technische universität dortmund

OBJECTIVE ASSESSMENT OF THE PERCEPTUAL QUALITY OF HMI-COMPONENTS WITH A PARTICULAR FOCUS ON THE HEAD-UP DISPLAY

Bei der Fakultät für Elektrotechnik und Informationstechnik
der Technischen Universität Dortmund
vorgelegte

Dissertation

zur Erlangung des akademischen Grades
Doktor der Ingenieurwissenschaften (Dr.-Ing.)

Von

M. Eng. Sonja Maria Köppl

Dortmund, 2017

Hauptreferent:	Prof. Dr. rer. nat. Christian Wöhler
Korreferent:	Prof. Dr.-Ing. Rüdiger Kays

Veröffentlichung als Dissertation in der Fakultät für Elektro- und Informationstechnik.

Dissertationsort:	Technische Universität Dortmund
Promotionsprüfung:	26.10.2017 um 16:00 Uhr
Hauptreferent:	Prof. Dr. rer. nat. Christian Wöhler
Korreferent:	Prof. Dr.-Ing. Rüdiger Kays

Chapter

I Danksagung

Nach Jahren intensiver Arbeit liegt sie nun vor Ihnen: meine Dissertation. Damit ist es an der Zeit, mich bei denjenigen zu bedanken, die mich in dieser herausfordernden Phase meiner akademischen Laufbahn begleitet haben.

Entstanden ist meine Dissertation während meiner Tätigkeit bei der Daimler AG in Ulm in Zusammenarbeit mit der Fakultät Elektrotechnik und Informationstechnik der Technischen Universität Dortmund.

Zunächst möchte ich mich ganz besonders bei Herrn Prof. Wöhler bedanken, der mir diese Arbeit ermöglicht hat. Dankbar bin ich auch über die Bereiterklärung von Herrn Prof. Kays, diese Arbeit als Zweitprüfer zu betreuen.

Ein weiterer Dank gilt Herrn Dr. Jostschulte von der Daimler AG, der mich mit vielen Diskussionen zu spannenden Forschungsfragen inspiriert hat. Ebenso möchte ich meinem Kollegen Herrn Dr. Presting von der Daimler AG dafür danken, dass er in zahlreichen Stunden meine Arbeit probegelesen hat. Ebenfalls bedanke ich mich bei Herrn Hellmann für die angenehme Zusammenarbeit.

Und am Ende danke ich meinen Eltern. Sie haben mir stets Mut zugesprochen, mir den Rücken frei gehalten und mich in meiner Arbeit bestärkt.
Eine größere und bessere Unterstützung hätte ich mir nicht wünschen können.

Chapter

II Abstract

This thesis focuses on applying classification methods to assess the perceived quality of virtual head-up display (HUD) images in luxury class vehicles. The quality of the virtual HUD image is affected by assembly tolerances. The assessment of the perceived image quality is done step by step. First, the perception of distortions and double images is assessed separately. The reason is that the different aberration types can be corrected by different measures. Distortions describe the difference between the real and the desired image geometry. Disturbing double images are created because the light is reflected at the inner and outer side of the windscreen. Only if the separate analysis of distortions and double images is successful, the combination of the 2 aberration errors is assessed.
The standard procedure to find the customer suitability of HUD images is the limit analysis. The goal is the specification of ergonomic limits. Compliance with these limits, an impairment-free reading is guaranteed. The standard method has a great weakness. Even simple combinations of various aberration types cannot be assessed. In this thesis, a new procedure is described to implement an assessment algorithm for HUD images. Here, classification methods are applied, which no longer show this weak point.

The customer suitability is recorded by the double-stimulus impairment scale method. This method is initially used to assess the subjective perception of television images. Here, this method is adapted to quantify the perceived quality of HUD images. 12 test persons label representative images according to the 5-grade impairment scale.
Occurring aberrations in HUD images are captured by 21 objective features. The main task of the assessment algorithm is to approximate the underlying relationship between the objective features and the subjective impressions.
Representative images are generated in an existing laboratory setup and identified by clustering methods. Thus, the distortion dataset has 1006 and the double image dataset has 345 training images. Additionally, each dataset comprises 360 test images. The dataset for the combined aberration types consists of 303 training images and 305 test images. All representative images are labelled by 12 test persons.

Classification tasks are divided into the training and the testing phase. During the training phase, the relationship between the subjective labels and the objective feature vec-

tors is detected. In the testing phase, the trained algorithm is applied to label an unlabelled test set. To find how well the classifier performs on unseen data. Here, the kNN (k-nearest neighbour classifier), the LVQ (learning vector quantisation) and the PC (polynomial classifier) are implemented as assessment algorithms.

The experimental evaluation shows that classification methods are well suited to assess the perceived quality of HUD images. Depending on the classifier type higher accuracy values are obtained on the test images as for the limit value consideration. Thus, the classification methods are able to assess combinations of various aberration types. For vehicles in the luxury segment, it is favourable to implement classifiers, which reduce the FPR (false positive rate). For the distortion dataset, the 2-class PC appears to be best suited. Here, the classifier obtains the FPR of 1.38% and the TPR (true positive rate) of 46.98%. The low TPR could be accepted since the rework costs are low. For the other 2 datasets, classifiers are intended that reach both a high TPR and a low FPR. The need for the high TPR is based on the time and cost consuming rework process. This conflict between a low FPR and a high TPR is best resolved for the double image dataset by the 2-class PC. The 2-class PC reaches the FPR of 1.36%, and the TPR of 97.18%. The best compromise for the distortion and double image dataset is reached by the LVQ. Here, the FPR of 13.04% and the TPR of 89.82% are obtained.

The proposed classifiers need many labelled training samples to yield a comprehensive and generalising recognition behaviour. However, the labelling of huge training datasets is costly and time-consuming. Thus, the semi-supervised learning (SSL) approach is applied to selected classifiers. SSL is an iterative procedure during which the learning process uses its own predictions to teach itself. The aim is to investigate if the manual labelling effort can be reduced by combining labelled and unlabelled training images. The experimental evaluation shows that unlabelled data in conjunction with a small amount of labelled data can improve the classification accuracy. The reason might be that the SSL process tries to avoid the use of poorly labelled training samples. SSL selects those samples, which improve the recognition behaviour in a helpful manner. Thus, the manual label effort can be considerably reduced without a loss of classification accuracy.
Finally, it is analysed if the labelling effort can also be reduced by the active learning (AL) procedure. The AL algorithm is able to select the most informative training images. These images are labelled by the test persons and transferred into the training dataset. The experimental evaluation shows that the AL process can improve the classification accuracy by using the most informative training samples. During the AL process, the manual labelling effort can be reduced to some informative training images. Thus, the training sets include uninformative images, which negatively affect the classification accuracy. The classifiers seem to be overcharged with the uninformative images.

Chapter

III

Table of content

Chapter

1 Introduction

This thesis is written to get the degree of Doctor Engineer, awarded by the University of Dortmund. This work is supervised by the Technical University of Dortmund from Prof. Dr. Christian Wöhler, Image Analysis Group at the faculty of Electrical Engineering and Information Technology and Prof. Dr.-Ing. Rüdiger Kays, field of communication technology at the faculty of electrical and computer engineering. The work has been performed at the research centre in Ulm of the Daimler AG.

In vehicles, human-machine-interface (HMI) components are distinguished between controlling and displaying elements. Both must meet the highest quality and safety requirements. So, the success of HMI components depends not only on price, reliability, and service life but also on handling and ease of use. The main need is that the operation of HMIs does not interfere with the driving task. For that matter, a fixed schema for the control and display areas has been established [MILICIC 10: p. 35-37]. HMI components that support the primary driving task are arranged in the direct field of view. Display elements are the instrument cluster and the head-up display. Steering wheel, pedals, and shift lever belong to the operating elements. The display regions below the instrument cluster and the controls around the steering wheel support the driver's secondary tasks. The secondary tasks include activities that are dependent on driving, like fading in and out, blinking, honking, and so on. All functions of the driver information systems are summarised in the centre console. They support the tertiary tasks and have nothing to do with the driving task itself. They are used to improving comfort and communication [MILICIC 10: p. 35-37].

HMI components must meet the highest quality standards. Thus, this chapter is used to clarify how customers make a quality evaluation. Afterwards, follows the problem definition and the structure of the presented thesis.

1.1 Quality perception

Due to various circumstances, customers have wishes and claims that are articulated as quality demands to the manufacturer. This wishes and claims are provided in form

of requirements and expectations. On the other side, the supplier provides a product with a certain state. The customer compares the state of the product with his demands and expectations. This comparison results in quality features, which are noticed through our 5 senses: seeing, hearing, feeling, smelling, and tasting. Thus, quality features are characteristics of a product and refer to a claim. A customer's satisfaction adjusts itself only if the expectations are fulfilled, as shown in Figure 1. So, the degree of the satisfaction depends on the extent to which the demands of the customers have been met [PFEIFER & SCHMITT 14: p. 372-385].

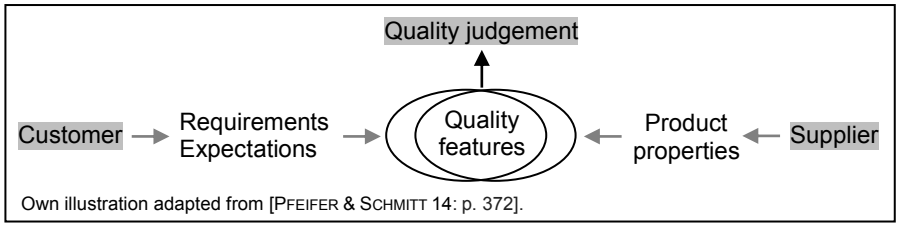


Figure 1: the quality judgement is delivered based on quality features

In relation to the head-up display, the expectations and requirements of the customers about the virtual image have to be identified. Comparing these expectations with the given image properties (distortions, double images, astigmatism…), it could be determined if the driver is satisfied. To fulfil the expectations, only vehicles should be sold that show a faultless image quality.

1.2 Definition of the project

The head-up display (HUD) system is an HMI component that supports the primary driving task in vehicles. It projects a virtual image above the engine hood. Presented are information that support the driver while driving; e.g. current speed, navigation hints, traffic signs, and so on. In order not to be distracted from driving, a good quality of the image must be ensured. The virtual image must be free of aberrations and both contrast and sharpness should be perfect. However, manufacturing defects and assembly tolerances downgrade the image quality. Depending on the size of the impact, observable aberrations are generated that disturb the overall impression. This results in the demand for an algorithm that is able to assess the perceived quality of head-up display images.

The aim of this thesis is to develop a new method, by which an assessment system for HUD images can be implemented. In this work, an algorithm is developed that is able to assess the perceived quality from head-up display images according to distortions and double images. For this purpose, the customer's requirements and expectations must be determined first. This is a complex task and the implementation entails considerable expense. The quality judgements are obtained through customer surveys. Here the test persons are interviewed about their impressions of representative HUD images. After that, the subjective ratings are linked to measurable variables. The aim is

to find a mapping between the subjective ratings and the objective features. In this work, the entire procedure is described systematically. The data acquisition, the subjective rating, the calculation of objective parameters and the development of an assessment algorithm are shown.

This work demonstrates how representative images can be selected from a plurality of images showing different varieties of aberrations. These selected images are then used for test person's questionings. It is shown that clustering methods (k-means, ward, and mean-shift) are able to split the amount of data into some subgroups. The centres of the clusters are the representative images, which are labelled by the test persons. This ensures that the determination of the subjective evaluation is made considerably easier. Likewise, the effort of the customer surveys is reduced because only 1 image of each cluster must be rated by the participants. Afterwards, the statements of the test persons result in the assessment algorithm. In contrast to the standard limit value consideration, methods of machine learning are implemented as assessment algorithm. The simple limit value analysis is the standard method to find the customer suitability of the HUD images [EICHHORN & ZINK 12], [SCHNEID 09: p. 23]. Unfortunately, this simple method has a great weakness. Even simple combinations of various aberration types cannot be considered. Thus, methods of machine learning are now applied, which no longer show this weak point. It is shown that these methods can result in more accurate results than the limit consideration method. As assessment algorithm, the polynomial classifier, the nearest neighbour classifier and the learning vector quantisation are applied. The results are then compared with those of the limit consideration.

In this thesis, it is shown for the first time that the perceived quality of HUD images can be mapped by methods of machine learning. With the help of clustering methods, the images are split into groups where the images in each group show no subjective difference. Thereafter, classification methods are used to evaluate the perceived quality of HUD images. This technique is quite new and has not been used in practice until now.

1.3 Structure of the thesis

The conceptual proceeding of the presented thesis is shown in Figure 2. It can be seen, that the basic part consists of 3 chapters. First, chapter 2 introduces the fundamentals of visual perception. It follows chapter 3, which analyses the head-up display system in more detail. The end of the basic part forms chapter 4 that describes the used test environment.
The main part includes 3 chapters and shows the development of an assessment algorithm for HUD images. In chapter 5, the subjective perception of varying levels of different aberration types is studied. Thus, objective features are defined to describe occurring aberrations. Chapter 6 introduces the used processes and method, which are needed to implement an assessment algorithm. Based on the findings of chapter 5, an assessment system is implemented in chapter 7, which relies on machine learning

methods. The thesis focuses on the assessment of the main aberration types like distortions, double images and the combination of both. First, the subjective quality assessment is determined for representative HUD images. This is done through extensive surveys. The representative images are selected by clustering methods like k-means, ward, and mean-shift. Once the subjective labels are known, an assessment system can be trained. The polynomial classifier, the nearest neighbour classifier, and the learning vector quantisation are implemented. To find the best-suited assessment algorithm, the results are compared with the findings of the conventional limit consideration. Finally, it is shown that semi-supervised and active learning rules can influence the classification results.
The presented thesis is completed with chapter 8, which gives the summary and the conclusion.

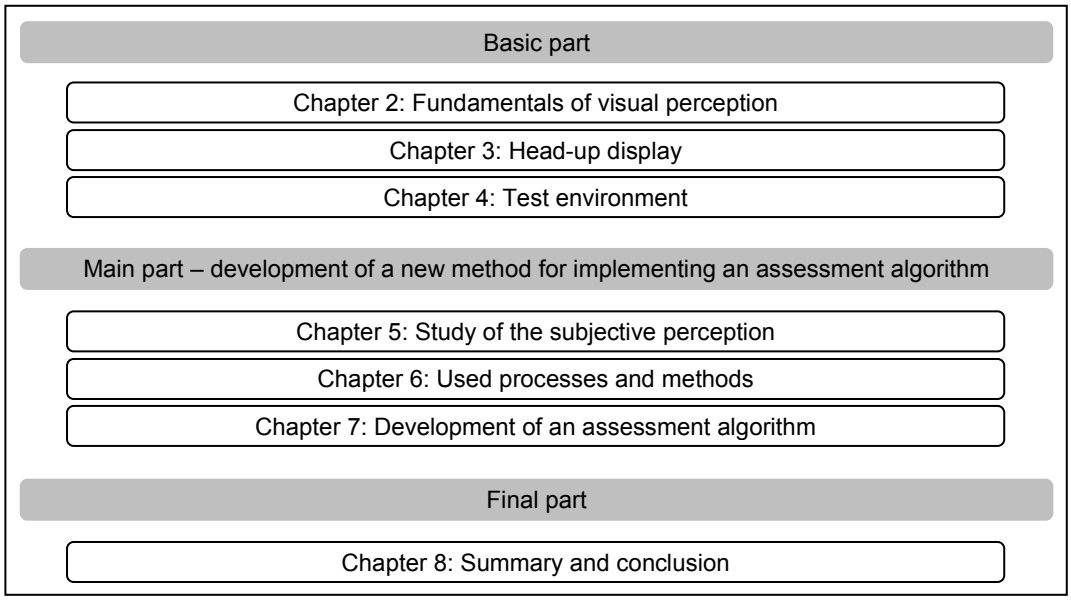


Figure 2: structure of the presented thesis

Chapter

2 Fundamentals of visual perception

Since the head-up display system is a purely visual system, the basics of visual perception are introduced. This chapter is used to analyse the nature of light, the human eye, and the resolution of the eye. Afterwards, it is investigated how it comes to accommodation, and adaption of the eye.

2.1 Nature of light

We perceive about 80% of the information from our environment by sight. All other senses, such as tasting, hearing, feeling, and smelling give a low percentage of information [DAHM 06: p. 41]. The information carrier for our eyes is the light, which is the visible part of the electromagnetic radiation. The wavelength of the visible light is in the range from 380 nm (violet) to 780 nm (red)[1] [HARTEN 74: p. 281].

Objects of our environment can be divided into 2 groups: bodies that emit light, and bodies that do not emit light. The non-self-luminous bodies can only be seen, if light emitted from a light source is reflected by the body [WINTER 42: p. 72f]. In summary, we can only see objects if the emitted or reflected light enters our eyes [HARTEN 74: p. 248-311].

The virtual image of the HUD is generated by a full-colour TFT display [BLUME et al. 13: p. 5]. Thereby, only the LED backlight is self-luminous. All other components emit no light. The generated image can only be seen because the emitted light rays are directed through a complex optical system into the driver's eyes. For details, see chapter *3.3*.

2.2 Structure and function of the human eye

The eye is a sensory organ for the perception of light stimuli. It is part of the visual system and enables us to see. A horizontal cross-section of the human eyeball is shown in Figure 3. The front of the eye is covered by a transparent surface called the cornea. The remaining outer cover, the sclera, is composed of a fibrous coat that surrounds the

[1] Wavelength of the visible radiation [HARTEN 74: p. 281]:
violet: 380 nm – 430 nm, blue: 430 nm – 490 nm, green: 490 nm – 570 nm,
yellow: 570 nm – 600 nm, orange: 600 nm – 640 nm, red: 640 nm – 780 nm

choroid, a layer containing blood capillaries. The light rays striking the eyes pass through the cornea, anterior chamber, pupil, lens, and vitreous to the retina. There, an inverted, real image is formed, which is passed through the visual nerve to the brain [GREHN 08: p. 3-16].
The amount of light that enters the eye is controlled by the iris, which acts as an aperture. This light is bundled on the retina surface by the lens that changes its shape for focusing on near and distant objects. This change is realised by the ciliary muscle, which is fixed by zonules to the lens [GREHN 08: p. 3-16].

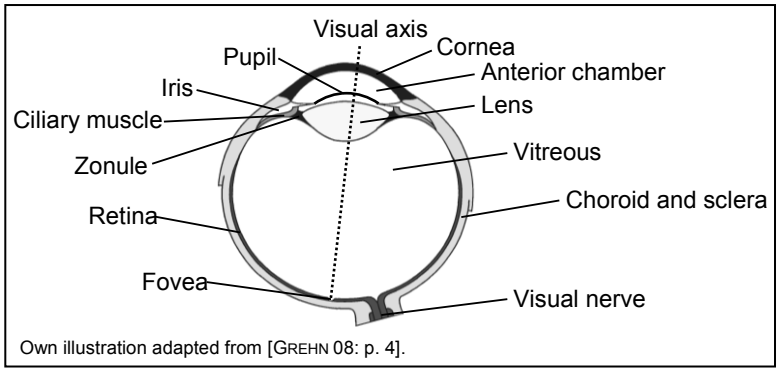


Figure 3: horizontal cross-section of the human eyeball

The retina consists of approximately 126 million photoreceptors that can be divided into 2 types, the rods, and the cones. The rods are long slender receptors, and the cones are generally shorter and thicker in structure. About 120 million rods represent the majority of the photoreceptors. They are responsible for the luminance sensation and react most sensitive to blue-green light of about 500 nm wavelength. In contrast, the 6 million cones are sensitive to light and are responsible for colour vision. 3 types of cone cells exist, each with a different spectral sensitivity (red, green, blue). Put together, they give an impression of colour [MEYER 08].
The rods and cones of the retina are distributed based on a horizontal line, the visual axis. However, we can only see well with a small area of the retina, called fovea, which covers only 2° of the visual angle. This area contains many cones and is the region of sharpest photopic vision. During the viewing process, the eye performs minimal movements. The brain creates from the different sharp seen information 1 image [DAHM 06: p. 44f].

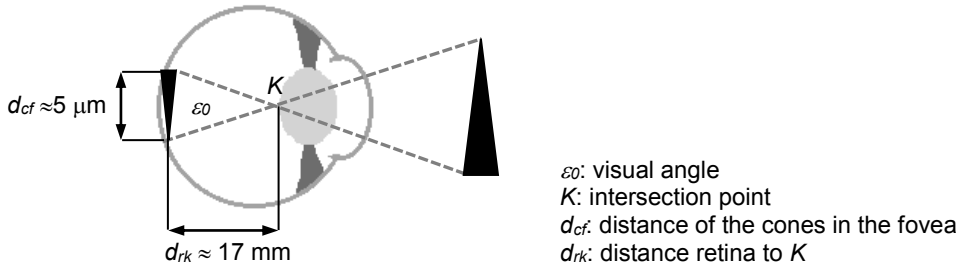
2.3 Resolution of the eye

The resolution of the eye is limited by the finite size of the cones and the distance from the cones to the retina. With the help of the reduced eye model, it is possible to find the minimal resolution angle. In this simplified model, the refractive elements of the eye, the cornea and the lens, are represented by a single surface. The intersection point[2] is

[2] The intersection point is the point on the optical axis, where light rays falling in at a certain angle, emerge on both sides at the same angle [BILLE & SCHLEGEL 05: p. 3].

in the centre of curvature. The distance between the retina and the intersection point is specified as 17 mm [BILLE & SCHLEGEL 05: p. 3].

In the fovea, the density of the cones reaches values between 50000 per mm^2 and 300000 per mm^2, depending on the eye. That corresponds to an average cone distance between 2 μm and 5 μm [CURCIO & ALLEN 90]. Thus, the minimal resolution angle of the eye could be determined, as shown in Figure 4.



Own illustration adapted from [BILLE & SCHLEGEL 05: p. 11].

Figure 4: minimal resolution angle based on the reduced eye model

Based on these geometrical conditions, a mathematical rule for the determination of the visual angle can be determined, as shown in Equation 1.

$$\tan\left(\frac{\varepsilon_0}{2}\right) = \frac{d_{cf}}{2} \cdot \frac{1}{d_{rk}}, \quad \Longrightarrow \quad \varepsilon_0 = 2 \cdot \arctan\left(\frac{d_{cf}}{2} \cdot \frac{1}{d_{rk}}\right)$$

ε_0: visual angle, d_{cf}: distance of the cones in the fovea, d_{rk}: distance retina to intersection point K

Equation 1: calculation of the minimal resolution angle of the eye

By rearranging the equation, and inserting the given values from the reduced eye model, the minimal resolution angle ε_0 is approximately determined as 0.017°. This roughly corresponds to 1' (1 minute of arc) [GREHN 08: p. 35].
In this thesis, the minimal resolution angle is required to determine the pixel size in the virtual HUD image; see chapter *3.4*.

2.4 Adaptation of the eye

Adaptation refers to the ability of the eye to adapt to different levels of brightness. In this process, the width of the iris is changed and the pupil size is adjusted, shown in Figure 5. The course and the time of the adaptation depend on the luminance values at the beginning and at the end of brightness change. During the light to dark adaptation, the vision process changes from cone vision to rod vision. This reduces both the sharpness and the colour perception. Additionally, the pupil is dilated so that the eye can absorb more light. Since the light sensitivity of the rods depends on the rhodopsin[3] concentration, a larger quantity of this substance has to be produced. This makes the

[3] Rhodopsin is a visual pigment, which is required to control the activity of the photoreceptors [GREHN 08].

eye more sensitive to light. The adaptation process from light to dark can take up to 35 minutes.

a) in the darkness, the pupil is dilated and much light can fall into the eye
b) good light conditions lead to a narrowing of the pupil

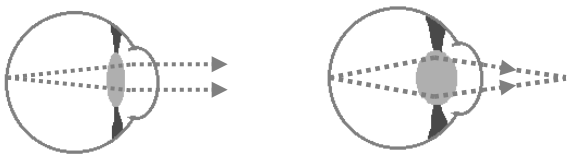
Own illustration adapted from [DAHM 06: p. 42].

Figure 5: adaptation of the iris to the ambient light

Against it, the eye can adjust within 2-5 minutes from dark to bright. The circular muscle of the iris contracts, the pupil becomes smaller so that less light hits the retina. During high light exposure, the rhodopsin degenerates and the light sensitivity of the eye decreases [GREHN 08: p. 49f].
A great advantage of head-up display systems in vehicles is that the adaptation time of the eye is reduced. For more details, see chapter *3.5.1*.

2.5 Accommodation of the eye

To focus objects at different distances, the eye changes its refractive power by altering the curvature of the flexible lens. The lowest optical fraction power of the lens is during minimal adjustment. The ciliary muscle, which affects the curvature of the lens, is completely relaxed and the lens is flat. If this muscle changes its tension, the lens changes to a curved shape. Thus, the focus point is moved closer (maximum accommodation), as shown in Figure 6. With the age, the lens loses its elasticity due to loss of water and the near point continues to move away from the eye [GREHN 08: p. 378f].



a) minimum adjustment: relaxed ciliary muscle, stressed zonule, flat lens
b) focus (accommodation): stressed ciliary muscle, relaxed zonule, curved lens

Own illustration adapted from [GREHN 08: p. 378].

Figure 6: accommodation of the eye

During the use of a head-up display system, the accommodation time for the eyes is decreased, as shown in chapter *3.5.1*.

Chapter

3 Head-up display

Over the years, the display technology in vehicles has changed significantly. Formerly, the original purpose of displays was the visualization of measured values from the vehicle. The first sign was a coolant thermometer, which was screwed on the radiator. The tank level indicator, the speedometer, the total distance recorder, and much more came shortly after. The clock was the first information source, which was not directly related to the vehicle state [MEROTH & TOLG 07: p. 79f].

Today, the displays are no longer limited to the instrument cluster. With the head-up display system (HUD), it is possible to project data in the driver's field of view. Automotive suppliers and manufacturers celebrate the head-up display as a great novelty. However, the system is already a few decades old. The roots of the HUD are in the military domain. Fighter jets have used the new display technology for years, to support the pilot in precarious situations. Then the display found the way into the civil aviation. Finally, in the 80s, General Motors offered, as first automobile manufacturer, the HUD system for vehicles. In addition, General Motors launched 2001 the first coloured variant in some cars. 2 years later, the first European car manufacturer BMW followed [KAUFMANN 04].

The idea of the head-up display is to have all relevant information during the car trip directly in the field of view. Therefore, the HUD (composed of a picture generation unit and complex mirror optics) is installed in the dashboard between the instrument cluster and the windscreen. The generated light rays are reflected by the mirrors and the windscreen. Due to the optical perception, the virtual image appears to hover above the hood [GÖTZE & BENGLER 15], [KAUFMANN 04], as shown in Figure 7.

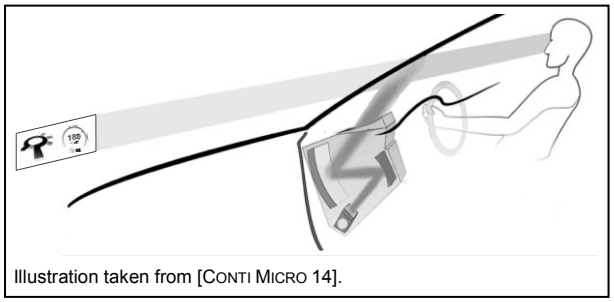
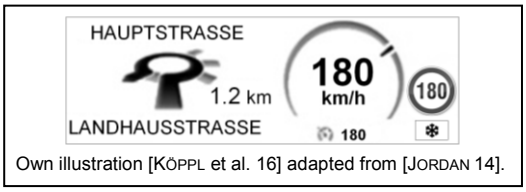


Illustration taken from [CONTI MICRO 14].

Figure 7: the virtual image appears to hover over the hood

The HUD systems can display different content. Basic information (current speed, turn indicator), speed limits, navigation instructions and details from driver assistance functions (road signs, distance warning) belong to it [JORDAN 14], as shown in Figure 8.



Own illustration [KÖPPL et al. 16] adapted from [JORDAN 14].

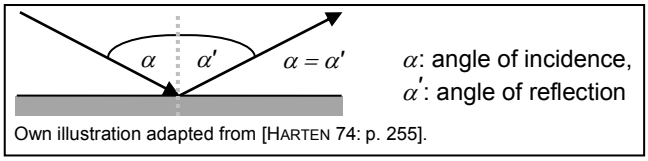
Figure 8: the virtual image shows driver relevant information

This chapter is used to get an idea of the head-up display system. First, the physical principles and the functionality of the HUD are discussed. Subsequently, the HUD-pixel size is determined and the interaction between man and machine via HUD is identified. Finally, it is investigated which kind of aberrations can occur and how they can be corrected.

3.1 Physical principles

The basic principle of the head-up display is the superposition of data with the real environment. Thereby the information is presented as a virtual image. To understand the origin of such a virtual image, the physical background is described briefly. In addition, the used vehicle coordinate system is shown.

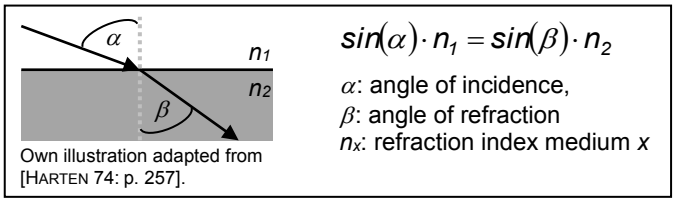
- *Law of reflection:* during reflection, the light encounters a change in direction. Thus, the angle of incidence and the angle of reflection have the same size [HARTEN 74: p. 255ff], as shown in Figure 9.



Own illustration adapted from [HARTEN 74: p. 255].

Figure 9: law of reflection

- *Law of refraction:* light propagates straight-lined. It encounters a change in direction during the transition of interfaces between media with different refraction indices [HARTEN 74: p. 257ff], as shown in Figure 10.



Own illustration adapted from [HARTEN 74: p. 257].

Figure 10: law of fraction

- *The interaction between light and material:* light rays that hit an object are partly reflected, partly transmitted, and partly absorbed, as shown in Figure 11. The sum of reflected, transmitted, and absorbed light is equal to the incoming light [Felder 11: p. 89].

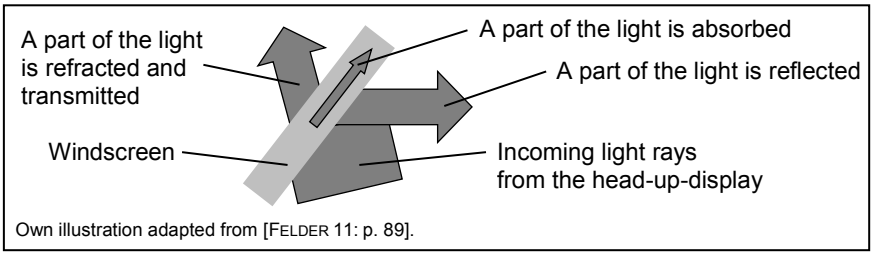


Figure 11: reflection, transmission, and absorption of light rays

- *Virtual image by plane mirrors:* plane mirrors generate virtual objects with the same size but mirror-inverted. If the reflected rays from the mirror fall in our eyes, we consider a straight course of the rays and the object emerges as a virtual image behind the mirror. The distance of the object in front of the mirror is equal to the perceived distance of the mirrored object behind the mirror [HARTEN 74: p. 255ff], as shown in Figure 12.

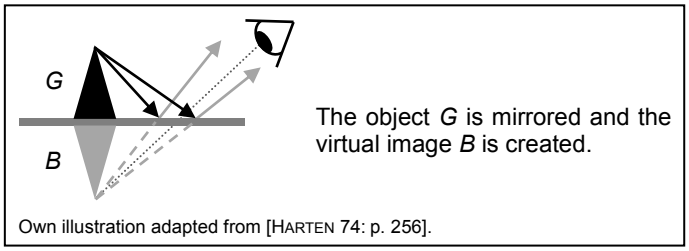


Figure 12: formation of a virtual image on a plane mirror

- *Virtual image by focusing lenses (magnifier glass):* if the object is arranged within the focal length, an enlarged virtual image is created. The image is upright, true-sided, and in the object space. Behind the lens, the rays do not intersect [HERING et al. 09: p. 288f], as shown in Figure 13. The virtual image is perceived at a greater distance than the object.

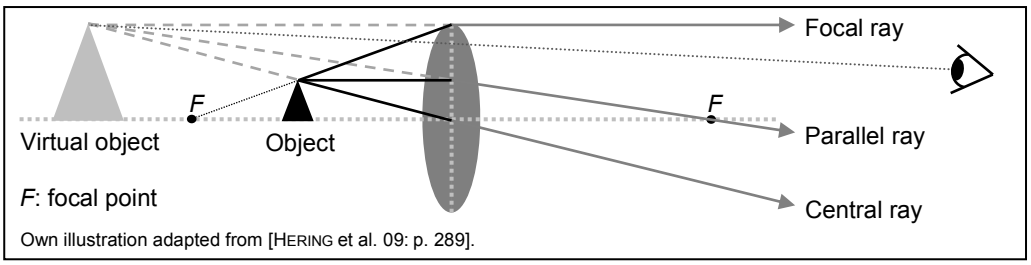


Figure 13: principle of the enlargement with a magnifier glass

- *The principle of the enlargement with a concave mirror:* concave mirrors combine the properties of mirrors and lenses. If the object is located within the focal

length, the virtual image is always upright, inverted, and enlarged. For the viewer, the virtual image is behind the mirror [HERING et al. 09: p. 279], as shown in Figure 14.

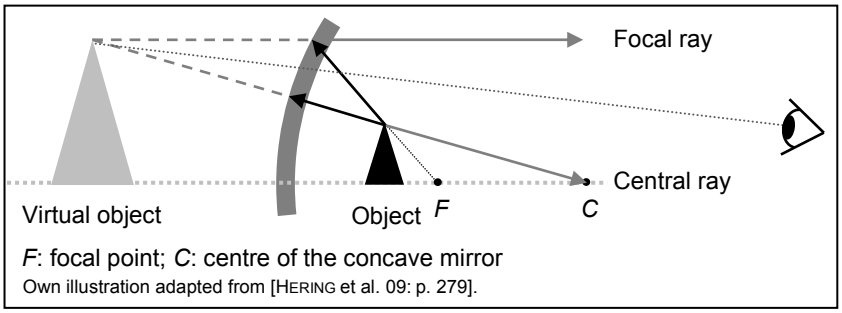


Figure 14: principle of the enlargement with a concave mirror

- *Used coordinate system*: the used orthogonal vehicle coordinate system is shown in Figure 15. The X-axis points in the opposite direction of the driving direction, the Z-axis upwards, and the Y-axis towards the passenger side [SEIFERT 05: p. 29].

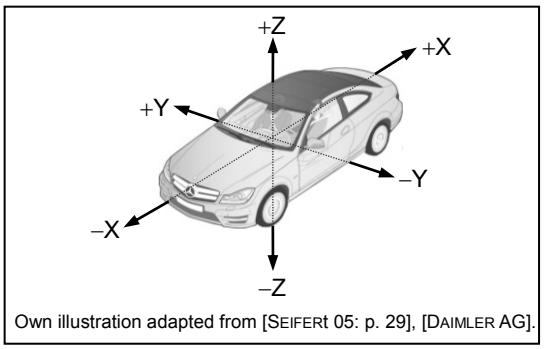


Figure 15: used vehicle coordinate system

3.2 Appearance of optical aberrations

This chapter is used to introduce the formation of optical aberrations. Optical aberrations are regular abnormalities, which are created by real optical elements compared to ideal elements [BILLE & SCHLEGEL 05: p. 9]. 2 groups of optical aberrations exist, chromatic and monochromatic aberrations.

3.2.1 Chromatic aberrations

Chromatic aberrations are caused by dispersion. It is expressed in the fact that refractive indexes depend on the wavelength of the incident light. Thus, chromatic aberrations split the light into its constituent parts and could be lateral colour or longitudinal colour, as shown in Figure 16 [THÖNIß 04: p. 6-10].

Longitudinal chromatic aberrations are caused by the fact that different colours show varying focal points. It describes the difference between the intersection points of the smallest and largest wavelengths with the focal plane [THÖNIß 04: p. 6-10].

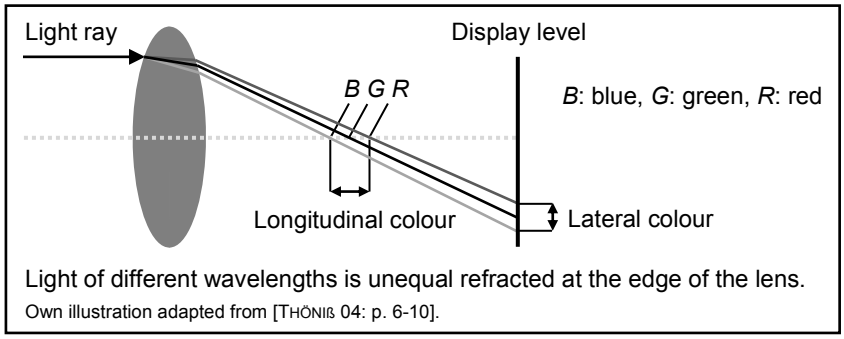


Figure 16: formation of longitudinal and lateral chromatic aberrations

Lateral chromatic aberrations occur on the projection level in the vertical direction and lead to colour changes in the image. They arise due to the longitudinal chromatic aberrations and the colour depending focal points [THÖNIß 04: p. 6-10].

3.2.2 Monochromatic aberrations

Monochromatic aberrations are independent of the wavelength. There exist monochromatic aberrations that display the image out of focus, e.g. spherical aberration, coma, astigmatism, and those that deform the image, e.g. curvature of field and distortion.

- *Spherical aberration*: after passing through a lens, incident parallel rays with different distances to the optical axis have varying focal distances, as shown in Figure 17. The complete image has a low contrast and details are less visible than expected [THÖNIß 04: p. 10f].

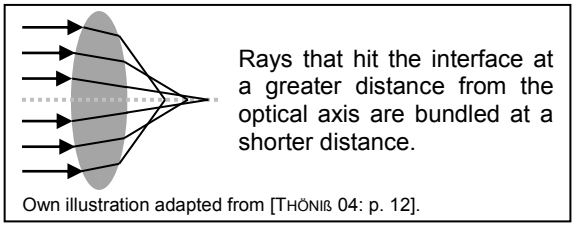
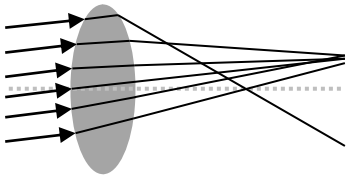


Figure 17: spherical aberration

- *Coma (asymmetrical error)*: coma occurs if oblique light rays strike the optical element, as shown in Figure 18. The emergence is based on the formation of the spherical aberration. However, the impact is stronger due to the oblique incidence angle. The image of sharply defined objects is blurred [THÖNIß 04: p. 13f].

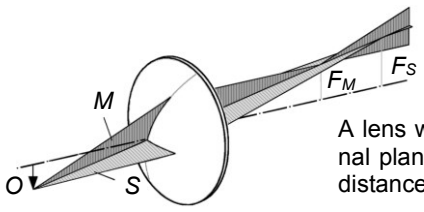


During passing an optical system, light rays lose a part of its symmetry characteristic.

Own illustration adapted from [THÖNIß 04: p. 13].

Figure 18: coma (asymmetrical error)

- *Astigmatism (focus error)*: astigmatism occurs if optical elements have different curvatures in horizontal and vertical direction. Rays in the meridional plane intersect at another point than the rays in the sagittal plane. Thus, different focal distances for horizontal and vertical lines exist. The distance between the 2 focal lines is called astigmatism and is perceived as blurring [THÖNIß 04: p. 14-16], as shown in Figure 19.



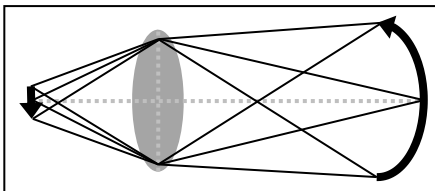
O: object,
M: meridional plane, *S*: sagittal plane,
F_M: focal line meridional, F_S: focal line sagittal

A lens with different curvatures has 2 focal lines (meridional plane and sagittal plane) instead of a focal point. The distance between these lines is called astigmatism.

Own illustration adapted from [THÖNIß 04: p. 15].

Figure 19: astigmatism (focus error)

- *The curvature of the field*: the image is not mapped on a flat layer but on a curved surface. Therefore, not all image points can be focused simultaneously. If the centre of the image is focused, the edges are blurred and vice versa. This error arises due to the fact, that abaxial rays are displayed closer to the optical axis as centred rays [THÖNIß 04: p. 16f], shown in Figure 20.



The image is not displayed on a plane surface, but on a curved region.

Own illustration adapted from [THÖNIß 04: p. 16].

Figure 20: curvature of field

- *Optical distortion:* a distorted image shows a dissimilar geometry than the object. This is based on the spherical aberration. The reproduction scale changes rotation-symmetrically with increasing the distance from the optical axis. The occurring dissimilarity can either be pincushion distortion or barrel distortion [THÖNIß 04: p. 17-19], as shown in Figure 21.

If the scale decreases towards the image border, a pincushion distortion appears. A barrel shape is generated, if the scale increases towards the image border.

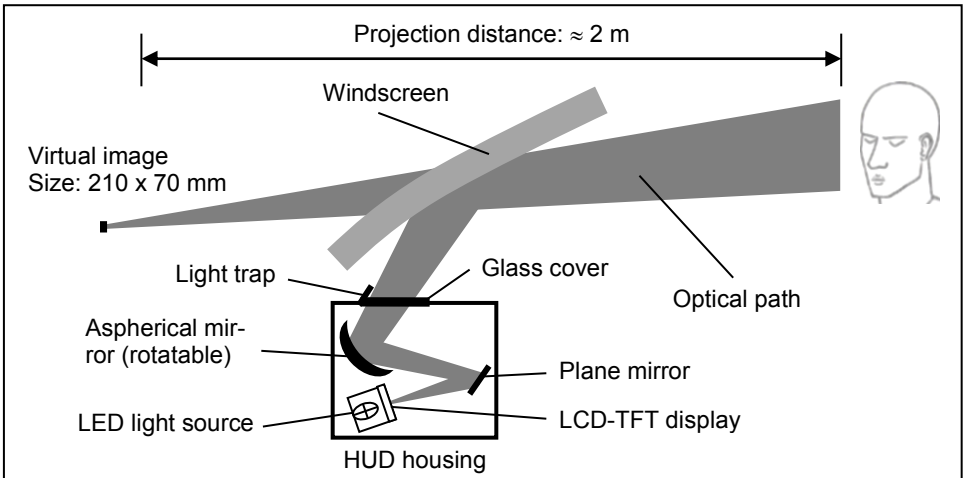
Own illustration adapted from [THÖNIß 04: p. 18].

Figure 21: different types of optical distortions

3.3 Functionality of the head-up display

The HUD is generally composed of a picture generation unit, mirror optics, and a special windscreen, as shown in Figure 22. The picture generation unit consists of an LCD-TFT display with LED backlights [CONTI 14]. A special light sensor, on the top of the roof, automatically adapts the display brightness to the surrounding light conditions. Thus, the brightness is adjusted dynamically to the particular situation and the readability is assured for both day and night [JORDAN 13].

The optic module directs the light rays from the picture generation unit to the windscreen. It consists of a fixed plane-folding mirror and a rotatable aspherical-concave mirror. The fixed plane mirror enables, based on the distracting properties, a convolution of the optical path. The aspherical mirror combines the displaying and the distracting proprieties. It enlarges the virtual image and enables through its rotation a height adjustment [SCHNEID 09: p. 6ff]. The projection distance of roughly 2 m results from the artificial extension of the ray path by the optic module [SCHNEID 09: p. 6ff]. Likewise, the size of the virtual image (210 x 70 mm [JORDAN 13]) is mainly determined by the enlarging properties of the concave mirror.



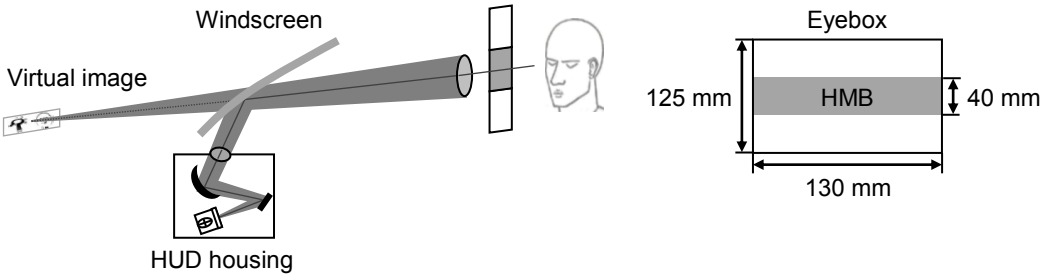
Own illustration adapted from [CONTI 14]; detailed information see [JORDAN 13], [CONTI 14].

Figure 22: functional description of the head-up display

The cover of the HUD housing consists of a glass plate. Thus, the optical elements are protected from dust and scratches. In addition, a light trap is mounted in front of the glass cover. It prevents that incident light is reflected towards the driver. Thus, stray light effects are prevented and a faultless projection, from the display information to the windscreen, is guaranteed [CONTI 14].

The windscreen is used as projection area for the virtual image and serves as the final semi-transparent mirror. However, the resulting image is not perceived as a reflection in the windscreen. Based on the main light translucent quality of the windscreen, the virtual image appears to hover above the forefront of the hood [SCHNEID 09: p. 6ff].

To see the virtual image, the eyes of the driver must be located within the head motion box (HMB). This is the range within the Y-Z plane, where the driver can perceive the virtual image [SCHNEID 09: p. 6], [CONTI 14], as shown in Figure 23. The used vehicle coordinate system is shown in Figure 15.



Head motion box: range where the eyes of the driver could perceive the virtual image.
Eyebox: range for the height adjustment of the head motion box.
Own illustration adapted from [CONTI 14]; detailed information see [SCHNEID 09: p. 25].

Figure 23: eyebox and head motion box

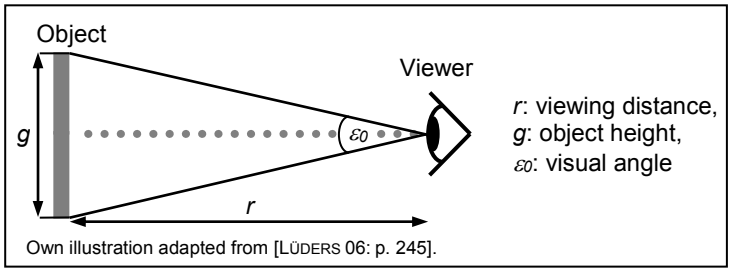
The dimension of the optical elements and the extent of the outlet port define the maximum size of the visible light rays and so the size of the HMB. With the help of the rotatable concave mirror, is it possible to do a height change of the virtual image and of the head motion box. Thereby, the position of the virtual image can be adjusted to the size of the driver. The range of the possible height adjustment is named eyebox. The eyebox represents the vertical range in Z-direction where the driver could adjust the head motion box [SCHNEID 09: p. 6ff], [CONTI 14]. A placement of the virtual image and the head motion box in the horizontal direction is unfortunately not possible. A typical size for head motion box is 130 x 40 mm, with a possible height adjustment of ± 42.5 mm [SCHNEID 09: p. 25].

To ensure a good readability of the virtual image, a maximal luminance of 10000 Cd/m^2 is realised [JORDAN 13]. Due to luminance losses in the optical path, the luminance of the LCD-TFT display is adjusted accordingly. For the optical elements of the HUD, a reflection coefficient of 80% can be estimated for both polarised and unpolarised light. The windscreen exhibits reflection coefficients of 25% and 13% for polarised and unpolarised light [SCHNEID 09: p. 7f]. Out of it, a TFT display luminance of 50000 Cd/m^2 for polarised light and 96154 Cd/m^2 for unpolarised light is determined[4].

[4] For comparison, some typical light density values, unit Cd/m^2:
mean clear sky: 8000 — night sky with full moon: 0.1 — disk of the sun at noon: $1.6 \cdot 10^9$
xenon lamp: $5 \cdot 10^9$ — halogen lamp wire: $30 \cdot 10^6$ — mat 60-W bulb: $120 \cdot 10^3$

3.4 Determination of the HUD-pixel size

Our eye can still distinguish 2 object points with an angular distance of about 1' (1 minute of arc). This is the smallest resolvable visual angle, as shown in chapter *2.3*. The general illustration of the visual angle is shown in Figure 24. Thus, it is possible to determine the minimal detectable size in the virtual HUD image.



Own illustration adapted from [LÜDERS 06: p. 245].

Figure 24: demonstration of the visual angle

Based on the virtual image, *r* represents the viewing distance of 2 m. The minimal visual angle ε_0 indicates the resolution of the human eye with 1'. The object size *g* is the height of the smallest perceptible part in the virtual image.
By identifying a calculation rule for the object size and inserting values for viewing distance and visual angle, the smallest detectable size in the virtual image is 0.58 mm, according to Equation 2.

$$g = tan\left(\frac{\varepsilon_0}{2}\right) \cdot r \cdot 2\,, \quad \Longrightarrow \quad tan\left(\frac{1'}{2}\right) \cdot 2\,m \cdot 2 \approx 0.58\,mm$$

g: object size, *r*: viewing distance, ε_0: visual angle

Equation 2: minimal detectable size in the virtual image

For the following calculations, a HUD-pixel in the virtual image has a height and width of 0.58 mm x 0.58 mm. If the size of the virtual image (210 mm x 70 mm) is divided by the HUD-pixel size, the resolution of the virtual image is determined. The virtual image has a resolution of 362 HUD-pixel x 120 HUD-pixel.

3.5 Interaction between man and machine via HUD

The HUD can be activated or deactivated via a button in the driver assistance bar. By settings in the instrument cluster menu, the driver can adjust the height of the virtual image. In addition, the display content and the brightness could be adapted individually. In vehicles with seat memory function, the chosen HUD settings are stored. The values are readjusted as soon as the driver has called the stored seating position [KRUSE 14]. If the driver has adapted the virtual image to its needs, the use of the HUD has some advantages and unfortunately some disadvantages.

3.5.1 Advantages using the head-up display

A driver takes about 1 second to read the speed from the tachometer. The duration of the deflection is influenced by 2 factors. On the one hand, the view is actually averted from the street to look at the dashboard. Contra wise, the accommodation, the process to focus the eyes, take its time. Researchers have calculated that the state of distraction could be reduced to 0.5 seconds by using the head-up display [KAUFMANN 04]. The reason is that the virtual image is displayed directly in the field of view. The driver does not avert the eyes from the road and the amount of accommodation is lower. This is because the displayed image is about 2 m away from the driver and not like the distance to the speedometer with roughly 25 cm [KAUFMANN 04].
In addition, an external light sensor adapts the luminosity of the virtual image to the ambient light. Thus, when focussing on the virtual image, the same light conditions are given and the adaptation time is negligible. This is the time, which would be needed to adjust the eyes from the sunlit street to the dark dashboard [MILICIC 10: p. 44]. Due to these advantages, the driver has its gaze always on the road and can react faster to unexpected incidents. This is how the head-up display makes the driving not only more user-friendly and comfortable but safer at the same time [KAUFMANN 04].

3.5.2 Disadvantages using the head-up display

Despite the described advantages (i.e., reduced re-accommodation, fewer eye/head movements, increased eye-on-the-road-time) it is explicitly warned against cognitive capture or attentional tunnelling. This refers to the non-perception of objects, due to the limited processing capacity of the human brain. Mainly older drivers are affected. Thereby, the attention of the driver is drawn apart from the driving situation and is directed to the virtual head-up display image. This is done subconsciously without the active intervention of the driver. The result could be that the driver reacts too late to critical situations [GISH & STAPLIN 95: p. 6].

3.6 Optical aberrations in HUD systems

Many optical elements are involved to generate a virtual image above the engine hood. Here, it is investigated if occurring optical aberrations affect the image quality. The formation of optical aberrations is explained in chapter *3.2*. Chromatic aberrations are strictly related to the refraction in a lens. Since the head-up display system does not include any lens, the longitudinal colour and the lateral colour can be ignored [DIAZ 05: p. 41]. The spherical aberration is caused by ball-shaped reflection surfaces. This kind of monochromatic aberration is suppressed by the use of properly designed and heavily controlled aspherical mirrors. Thus, coma and field curvature are also negligible in HUD systems [DIAZ 05: p. 42].
In contrast, the windscreen has different radii in the horizontal and vertical direction. This means that a distorted astigmatic image is produced. If astigmatism becomes too large, symptoms like headache, eyestrain, fatigue, and blurred vision are detected.

Thus, only astigmatism and optical distortions influence the image quality [DIAZ 05: p. 21, 42].

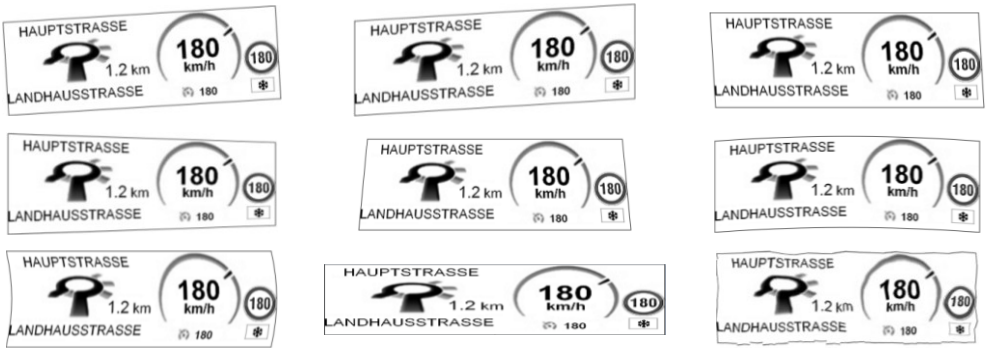
3.7 Origin and types of HUD typical irregularities

Besides to optical aberrations, many irregularities exist, which are only caused by the HUD composition of the vehicle. The quality of the virtual image is affected by assembly tolerances of the head-up display in the dashboard, the quality of the windscreen, and the poor manufacturing of the involved components. In the following sections, the individual irregularities are identified as distortions, double images, binocular misalignment, and dynamic variance.

3.7.1 Distortion formation in the vehicle

Distortions describe the difference between the real and the desired image geometry [MILICIC 10: p. 17]. Regardless of optical distortions, there are other sources of defects, which cause HUD typical distortions. The primary challenge is the curvature and the waviness of the windscreen. The plane image from the TFT display is projected on the curved surface and the resulting virtual image is deformed. Distortions also emerge based on assembly tolerances. The HUD housing and the windscreen can be installed only up to a certain accuracy. Such inaccuracies could affect the image quality, see chapter *4.3*.

Possible types of distortions are rotation, trapezoid, aspect deviation [EICHHORN & ZINK 12], misalignment, smile or local magnification [NEUMANN 12: p. 17, 33], [KÖPPL et al. 16], as shown in Figure 25.



Own illustration adapted from [HAN 05], [EICHHORN & ZINK 12], [NEUMANN 12: p. 17, 33].

Figure 25: example images with HUD typical geometric distortion types

Depending on the type of impact, one or more geometrical distortions can occur simultaneously. The perceived distortion of the virtual image results from the combination of all single types of distortions.

3.7.2 Double image effect - formation of the ghost image

The virtual image of the head-up display is formed by reflections at the windscreen. The windscreen, which is the last semi-transparent mirror, is made of 2 glass plates that are separated by a PVB[5] interlayer. As the light is also reflected at the inner and outer sides of the windscreen, a disturbing double image is created, shown in Figure 26.

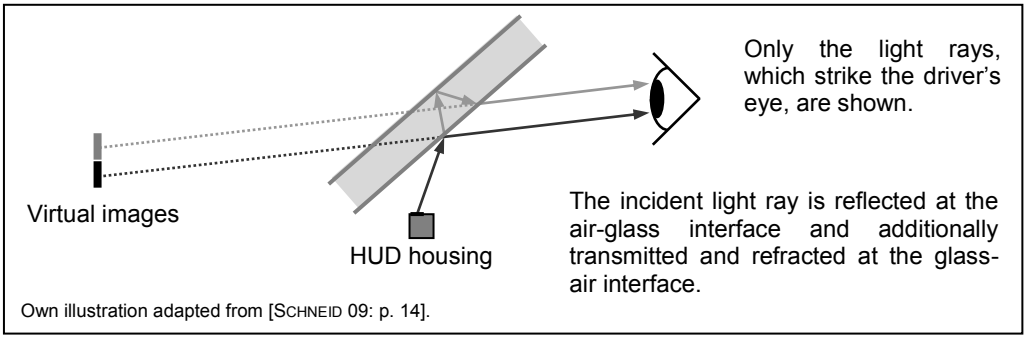
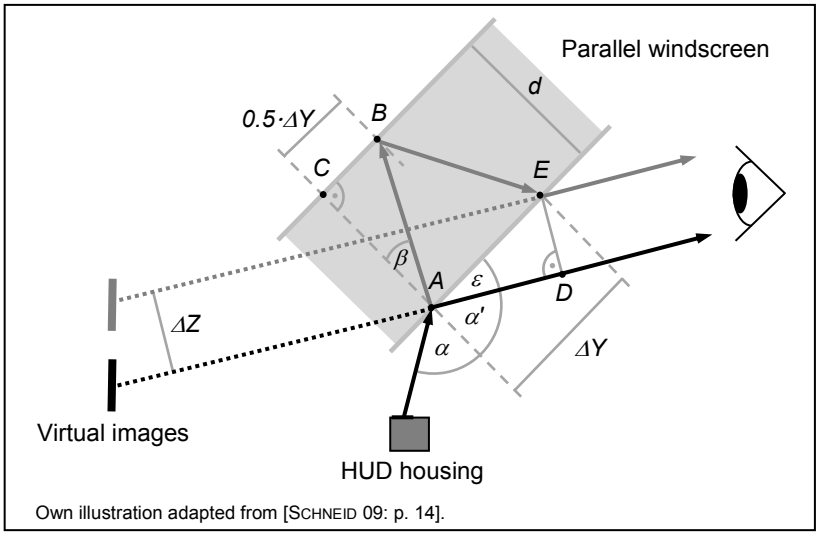


Figure 26: formation of a double image on a parallel windscreen

The incident light ray strikes the inner plate of the windscreen. At the air-glass interface, the light rays are partly reflected and partly transmitted. The driver can see the reflected part of the light as a primary virtual image. The transmitted percentage of the light is refracted, and the rays strike the outer plate of the double-glazing. At the glass-air interface, a part of the light is reflected back to the inner glass plate [SCHNEID 09: p. 13-16]. The remaining part of the light is transmitted and is observable from above, e.g. from a bridge looking down, over a small viewing angle [OTT & POGANY 09: p. 42f]. Back on the inner glass plate, a part of the light is refracted and transmitted at the glass-air interface. Thus, the driver can perceive an extra virtual image. The second image is in a slightly different position to the primary image and is called double image or ghost image [SCHNEID 09: p. 13-16]. In summary, the images are formed on the inner glass plate. The primary image is generated at the air-glass interface and the double image at the glass-air interface.

If a parallel windscreen is considered, the 2 images are displaced horizontally by an offset ΔZ. This offset is dependent on the thickness of the windscreen, the angle of incidence, and the refraction indices of air and glass. The exact calculation of the offset can be traced based on Figure 27.
Using the law of reflection, the angle of the reflected primary image α' can be determined as: $\alpha = \alpha'$. The angle of refraction β is determined for the air-glass interface, according to the law of refraction.

[5] PVB is the abbreviation for polyvinyl butyral. It is assumed that the PVB interlayer has almost the same refraction index as the 2 glass-plates.



Own illustration adapted from [SCHNEID 09: p. 14].

Figure 27: sketch for calculating the offset ΔZ for a parallel windscreen

The trigonometric functions of the right triangle *ABC* give $tan(\beta) = 0.5 \cdot \Delta Y \cdot d^{-1}$ and thus $\Delta Y = tan(\beta) \cdot 2 \cdot d$. For the right triangle *ADE* can be supposed that $\varepsilon = 90° - \alpha'$, $sin(\varepsilon) = \Delta Z \cdot \Delta Y^{-1}$ and so that $\Delta Z = \Delta Y \cdot sin(90° - \alpha')$. Thus, it is possible to find a general calculation rule for the distance between the virtual image and the double image, shown in Equation 3.

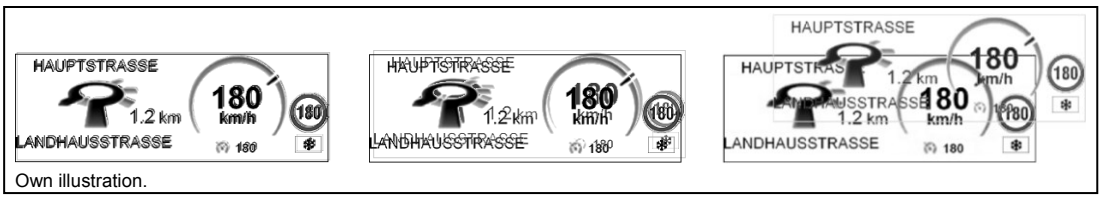
$$\Delta Z = tan(\beta) \cdot 2 \cdot d \cdot sin(90° - \alpha), \quad \text{with} \quad \beta = arcsin\left(\frac{sin(\alpha) \cdot n_{air}}{n_{glass}}\right)$$

ΔZ: distance between the images, α: angle of incidence, β: angle of refraction, n: refraction index, d: windscreen thickness

Equation 3: horizontal offset between the images for a parallel windscreen

By applying the following given values ($\alpha \approx 70°$, $d \approx 8$ mm, $n_{air} \approx 1$, $n_{glass} \approx 1.5$), the distance ΔZ amounts about 4.39 mm.

In the vehicle, the windscreen is not parallel. Depending on the complex non-symmetrical curvature, the secondary image may show a horizontal and vertical displacement, shown in Figure 28. In addition, the distance between the appearing double image and the primary image is not constant over the entire image.



Own illustration.

Figure 28: various distinct horizontal and vertical double images

3.7.3 Dynamic variance and binocular misalignment

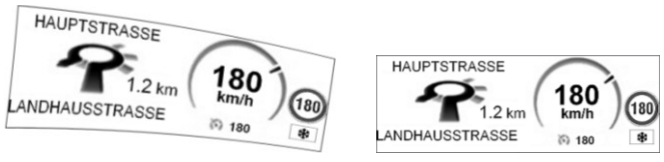
Due to the complex geometry, the windscreen exhibits a different curvature and waviness at each point. Thus, the virtual image shows varying aberrations (distortions and double images), viewed from different eyebox positions [MILICIC 10: p. 17]. If the driver moves his head within the HMB, aberrations could be perceived that change dynamically. The permanent variation of distortions and double images, even with slight head movements, is referred as dynamic variance [NEUMANN 12: p. 17]. It is also possible that the images for the right and the left eye are unequal. If the difference between the 2 images gets too large, the brain finally recognizes the 2 images no longer as belonging together. Binocular misalignment is the result. It occurs if a single object in the HUD image cannot be lined up on the retina with a suitable fixation, due to varying aberrations. The results are visual fatigue, binocular rivalry, and headache [GISH & STAPLIN 95: p. 15].

3.8 Prevention and correction of possible irregularities

The introduced irregularities can be eliminated with varying degrees of success. This chapter gives a short view of possible correction alternatives for distortions and double images.

3.8.1 Image warping – avoiding distortion

Distortions arise due to the optical system, the curvature of the windscreen, and assembly tolerances in the vehicle. It is tried to suppress occurring distortions by image warping. Warping assumes that an undistorted image is transformed to a distorted image by an optical system. Thus, a suitable pre-distorted image passed through the optical system is converted into a straight image [MILICIC 10: p. 17], shown in Figure 29.



A suitable pre-distorted image passed through the optical system is converted into a straight undistorted image.

Own illustration.

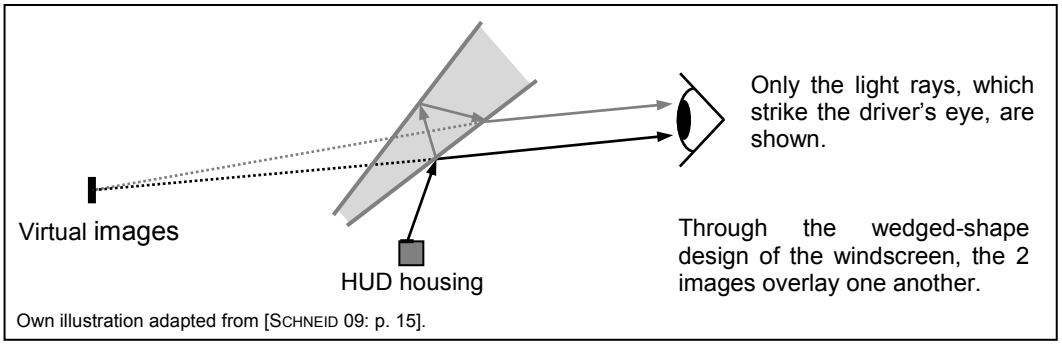
Figure 29: pre-distorted image on the TFT display and undistorted virtual image

The pre-distortion is described mathematically by 2D coordinate transformations [MILICIC 10: p. 17]. To determine the pre-distortion, a few sampling points are distributed on the display. It is calculated, how these points need to be shifted on the display to project a virtual image, which is free of distortions. Intermediate points are interpolated [NEUMANN 12: p. 17, 84].

Warping is only successful if the distortion is constant. Unfortunately, distortions change dynamically over the eyebox area. Thus, an average adaption over the complete range can be generated [MILICIC 10: p. 17]. Consequently, the dynamic variance and the binocular misalignment are also reduced.

3.8.2 Avoidance of double images

The windscreen generates 2 dislocated images. The primary image is generated at the air-glass interface and the second image at the glass-air interface. Since the second reflection cannot be avoided, the 2 images are superposed. Thus, special HUD windscreens are necessary, which are produced with a wedge-shaped PVB layer. This layer tapers in the direction of the engine bonnet [SCHNEID 09: p. 13-16], as shown in Figure 30.

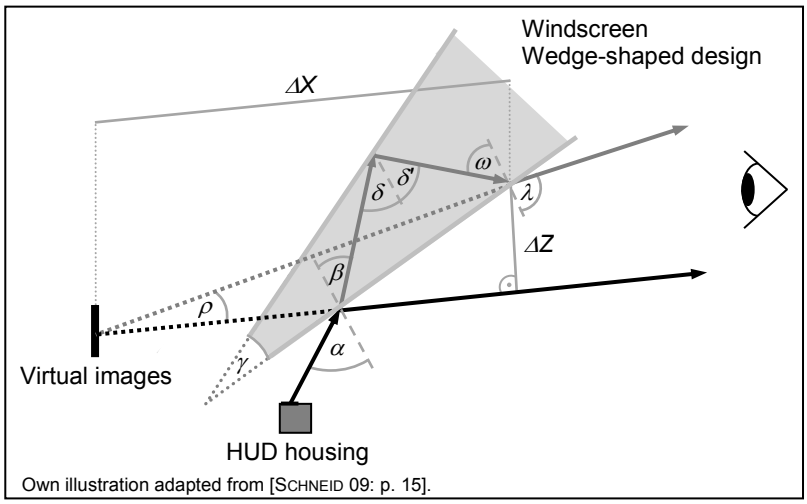


Own illustration adapted from [SCHNEID 09: p. 15].

Figure 30: windscreen with wedge-shape design to eliminate double images

The consequence of this wedge-shape is that the mirrored light rays are reflected at different angles. Thus, the 2 images overlay one another. Consequently, the driver sees the 2 images overlapped, giving the impression that only 1 exists. A mathematical sketch for the calculation of the wedged angle γ is illustrated in Figure 31.

For having 2 virtual images that overlay one another, the reflected light rays must be at a certain angle to each other. The angle ρ can be calculated as $\rho = arcsin(\Delta Z \cdot \Delta X^{-1})$. The law of reflection provides that $\delta = \delta'$. According to the wedge-shape of the windscreen, the angles are connected in the following ways: $\lambda = \alpha + \rho$, $\delta = \beta + \gamma$ and $\omega = \delta' + \gamma$.



Own illustration adapted from [SCHNEID 09: p. 15].

Figure 31: sketch for the mathematical calculation of the wedge angle

Out of this, the equations $\omega = \beta + 2 \cdot \gamma$, and $\gamma = 0.5 \cdot (\omega - \beta)$ can be derived. With the knowledge that ω could be determined by the law of refraction, a general calculation rule for the wedge angle γ is proposed, according to Equation 4.

$$\gamma = 0.5 \cdot (\omega - \beta), \quad \text{with} \quad \omega = \arcsin\left(\frac{\sin(\alpha + \rho) \cdot n_{air}}{n_{glass}}\right), \quad \text{with} \quad \rho = \arcsin\left(\frac{\Delta Z}{\Delta X}\right)$$

γ: wedge angle, α: angle of incidence, n: refraction index, β: angle of refraction, ρ: angle between the reflected light rays

Equation 4: wedge angle of the windscreen

By applying exemplarily given values ($\alpha \approx 70°$, $\beta \approx 38.79°$, $n_{air} \approx 1$, $n_{glass} \approx 1.5$, $\Delta X \approx 1.5\ m$, $\Delta Z \approx 4.39\ mm$), the angle γ amounts about 0.025° or 0.436 mrad.

Like the pre-distortion of the virtual image, the design of the wedge angle is only possible for constant double image distances. Unfortunately, this is not given by a curved windscreen. Thus, a perfect overlay of the 2 images can only be achieved for the design point. By the complex non-symmetrical form of the windscreen, the wedge angle does not ideally match the varying surface shape.
Besides, it must be ensured that the wedge-shape PVB layer is fitted properly between the glass plates. Otherwise, the double images become visible again.

3.9 Investigation area of the thesis

In this work, only the subjective perception of distortions and double images is investigated in detail. Disregarded are astigmatism, dynamic variance, and binocular misalignment.

Chapter

4 Test environment

This chapter is used to show the test environments that are deployed to generate, analyse, and assess different HUD images. Thus, the general procedure of the quality estimation of HUD images is introduced first. Subsequently, the presentation of the experimental setup and the generation of a database as basis for the assessment system follow. The section for the preparation of the images for the subjective assessment comes after. In the penultimate part of this chapter, the used test environment for the subjective assessment of the HUD images is shown and the selection of suitable test persons is presented. Finally, the standard procedure is described, which is usually used to assess the quality of HUD images.

4.1 General procedure to assess the HUD image quality

The basic procedure to capture the quality of HUD images is introduced in [EICHHORN & ZINK 12]. In summary, the quality estimation is carried out in 4 phases, as shown in Figure 32.

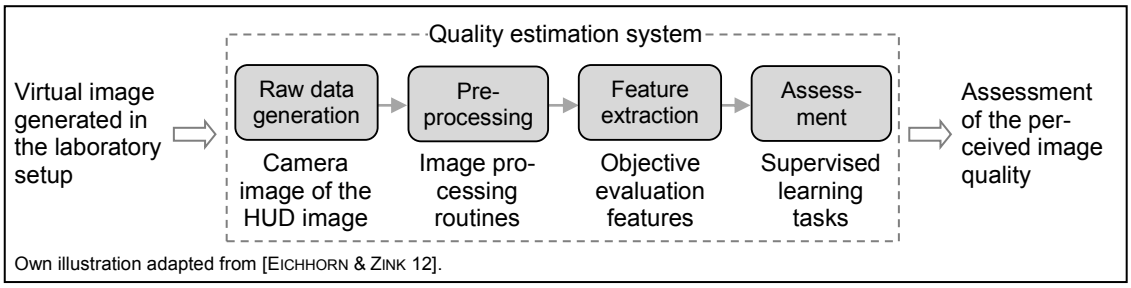


Figure 32: general procedure to predict the perceived image quality

- *Raw data generation:* the base of operation is a camera image of the virtual head-up display image.
- *Pre-processing:* details from the camera image are extracted by image processing routines. These details are then transferred into mathematical descriptions.
- *Feature extraction:* objective features values are calculated that describe the perceived quality of the virtual image.
- *Assessment:* the presented thesis focuses on this phase. The quality assessment of perceptible aberrations is achieved by methods of machine learning.

4.2 Experimental setup for the generation of virtual images

For generating HUD images, an existing laboratory setup is used. The Daimler AG owns a measuring construction that enables the generation of virtual HUD images. The structure of the setup is similar to the test stand described in [SCHNEID 09: p. 21f]. It consists mainly of a HUD, a windscreen and a driver's seat. The arrangement of the individual components corresponds to the geometry in actual vehicles. The movable mounting of the components allows the simulation of all translational and rotational assembly tolerances. By sitting in the driver's seat, the resulting virtual images can be considered. In addition, the windscreen can be changed by simple hand movements. Thus, not only assembly tolerances but also different double image occurrences could be investigated. Since the visibility of double images depends mainly on the quality of the windscreen, see chapter *3.8.2*.
It is possible to generate a camera image of the virtual HUD image from different positions. These camera images are the base for further investigations. The camera positions, which are used in this thesis, are located in the centre, the upper left corner, and the lower right corner of the eyebox [EICHHORN & ZINK 12], as shown in Figure 33. According to the used camera position, the HMB has to be adjusted in the upper, central, or lower position.

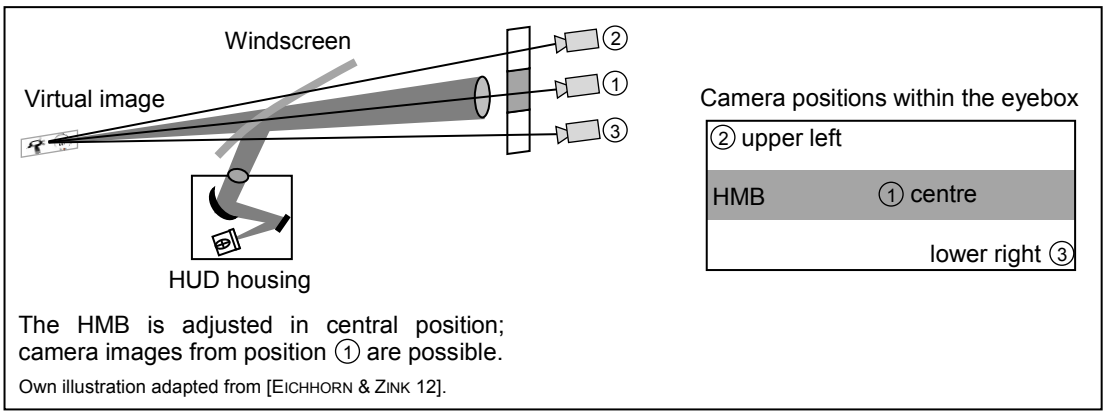


Figure 33: experimental setup for simulating aberrations

The camera positions and the assembly tolerances of the single components can be adjusted very accurately. Therefore, the measurements are reproducible. Camera positions and assembly tolerances can be adjusted via an existing self-written software from the Daimler AG. Thus, fully automated measurements are possible.

To analyse occurring distortions and double images, an already established test pattern is displayed onto the windscreen. Test pattern requires characteristic points or measuring marks that can be detected by accurate and robust image processing algorithms. In this work, the camera images of the displayed test pattern are analysed some existing image processing algorithms. These functions are developed by the Daimler AG and implemented in HALCON. The used test pattern consists of 9 x 21 white squares with an enlarged central point [KEM 06], [EICHHORN & ZINK 12], shown in

Figure 34. The central dot is known as the mass centre point or centre of gravity [EICHHORN & ZINK 12].

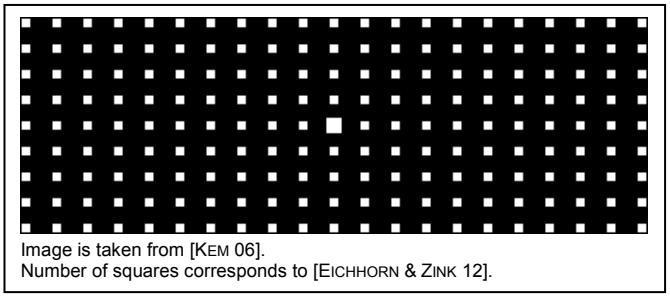


Figure 34: used test image to detect aberrations

For distortion detection, the image processing algorithms calculate the centre points of the 9 x 21 white squares. Each detected square is represented by its centre x-y coordinates. The positions of the centre points describe the distortion of the image [EICHHORN & ZINK 12]. Here, the 9 x 21 centre points are converted into 13 objective features that describe the distortion of the image, see chapter *5.1*. Additionally, the pixel offsets between the measuring marks and the corresponding double images are determined over the entire pattern [EICHHORN & ZINK 12]. The horizontal and vertical offsets between the double image and the corresponding white square are calculated. The pixel offsets are converted into 8 objective features that capture occurring double image distances of the HUD image, see chapter *5.1*.

4.3 Aberrations caused by assembly tolerances

The experimental setup described in the previous section is used to investigate which assembly tolerances lead to visible aberrations. Functional and optical errors of the HUD can be excluded, due to the 100% final inspection at the supplier [SCHUMM & WORZISCHEK 11: p. 36]. Thus, the quality and the assembly of the windscreen, as well as the assembly accuracy of the HUD in the dashboard could affect the image quality.

During a preliminary investigation, these inaccuracies are adjusted and the resulting virtual images are assessed subjectively. This preliminary study has already been carried out at the Daimler AG and is enhanced by this thesis to the subjective evaluation of the images. Several test series show that assembly tolerances of the windscreen do not lead to visible aberrations. The windscreen can be positioned wrongly or with restraints, but only negligible distortions and double images are generated. The generation of visible double images depends solely on the quality of the windscreen. To be more precise, the generation of double images depends on the quality of the wedge-shaped angle between the glass plates. This also explains why translational displacements of the HUD lead to visible double image distances. By shifting the HUD, the light rays hit no longer the design area of the windscreen. Outside this design range, the wedge angle does not exactly correspond to the angle of incidence and the double

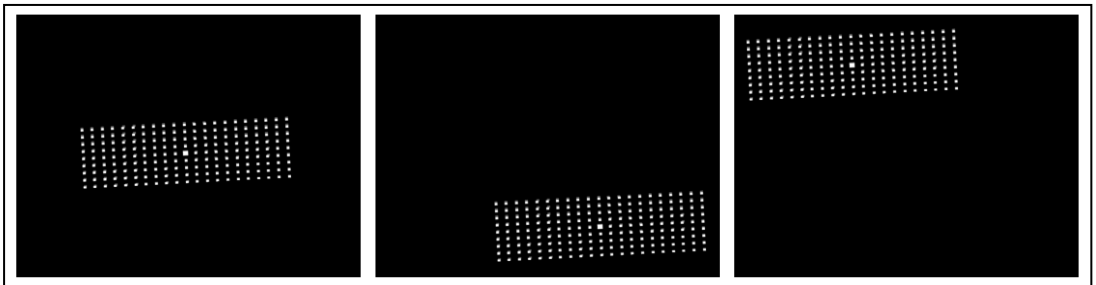
images are no longer suppressed. In contrast, a rotated HUD generates visible distortions and double images.

Next, a database of camera images is generated. These images show typical aberrations in HUD images. The aberrations are created by the simulation of assembly tolerances. Thus, only assembly tolerances of the HUD are considered. The test pattern, shown in Figure 34, is displayed onto the windscreen to analyse the images by existing image processing algorithms. The image processing routines are developed by the Daimler AG. 23625 different rotational and translational assembly tolerances of the HUD are simulated. For each resulting virtual image, a camera image is made. The measurement is executed with 3 eyebox positions and 2 pre-production windscreens. Since the resulting double images depend on the quality of the used windscreen. Thus, the adequate large number of 141750 images is accomplished (23625 different assembly tolerances · 3 different eyebox positions · 2 windscreens).
The camera images are analysed by existing image processing routines to detect occurring aberrations in size. The image processing is only possible if the enlarged central point of the 9 x 21 white squares could be detected. Unfortunately, this is only the case for 141263 images. The remaining 487 images are trimmed so far that more than the half image, with the central point, is missing. In addition, only images that are fully represented should be analysed. Trimmed Images are sorted out. 69368 images are trimmed and excluded from further analysis. Out of the 71895 remaining images, the feature extraction for distortion is possible for each image. For describing occurring distortions, all centre coordinates of the test pattern must be identified. In contrast, the extraction of the double image features is only possible for 64410 images. Only if the pixel offsets for all measuring marks can be determined, it is possible to describe occurring double images. Thus, the database consists of 71895 distorted images and 64410 images for the assessment of double images.

4.4 Preparation of the images for subjective assessment

The camera images of the database described in the previous section need to be prepared for the subjective assessment. The reason is that all interferences that may influence the evaluation should be eliminated. Interferences are the poor image contrast and the fact that the test pattern is displayed at different places in the camera image. In addition to this, a process is demanded that allows the separate assessment of distortion and double images. The separate assessment of distortions and double images is required for the implementation of the evaluation algorithm, see chapter 7.
Depending on the camera position, the virtual image is displayed at different places in the camera image, as shown in Figure 35. The reason is that the camera cannot adjust its inclination angle to the changed position. Preliminary studies have shown that the test persons are confused about the different places of the test pattern in the camera image. These displacements of the test pattern should not be included in the evaluation.

To meet this need, the centre of the test pattern (enlarged square) is shifted in the centre of the camera image.



Camera position ① centre | Camera position ② upper left | Camera position ③ lower right

The virtual image is taken from different camera positions. The camera does not adapt its angle of inclination to the changed position. Thus, the virtual image is displayed at another point on the camera image.

Own illustration.

Figure 35: resulting camera images from different camera positions

The other interference that must be eliminated is the poor contrast of the camera image. The contrast of the camera images is influenced by environmental factors such as room lighting or sunlight. The poor contrast of the camera image, shown at the top of Figure 36, leads to the fact that double images could not be uniquely identified. The grey double images are not clearly visible on the greyish black background. To get meaningful assessments all existing aberrations must be clearly visible.

Thus, Matlab routines are implemented, which prepare the images for the subjective assessment. Input parameters are the results of the image processing, see chapter *4.2*. The processing of the program generates 3 images, shown in Figure 36. All 3 images show the clear white test pattern on the pure black background. Here, the centre of the test pattern is identical to the centre of the entire image. The 1st image shows the resulting distortion without double images. The 2nd image shows the undistorted test pattern with the associated double images. Finally, the 3rd image shows, like the original camera image, the combination of distortions and double images. The difference to the original camera image is the very high contrast ratio. Thus, the double images are clearly visible. Since the background of the image is deep black, the grey double images are better visible than on a greyish black background.

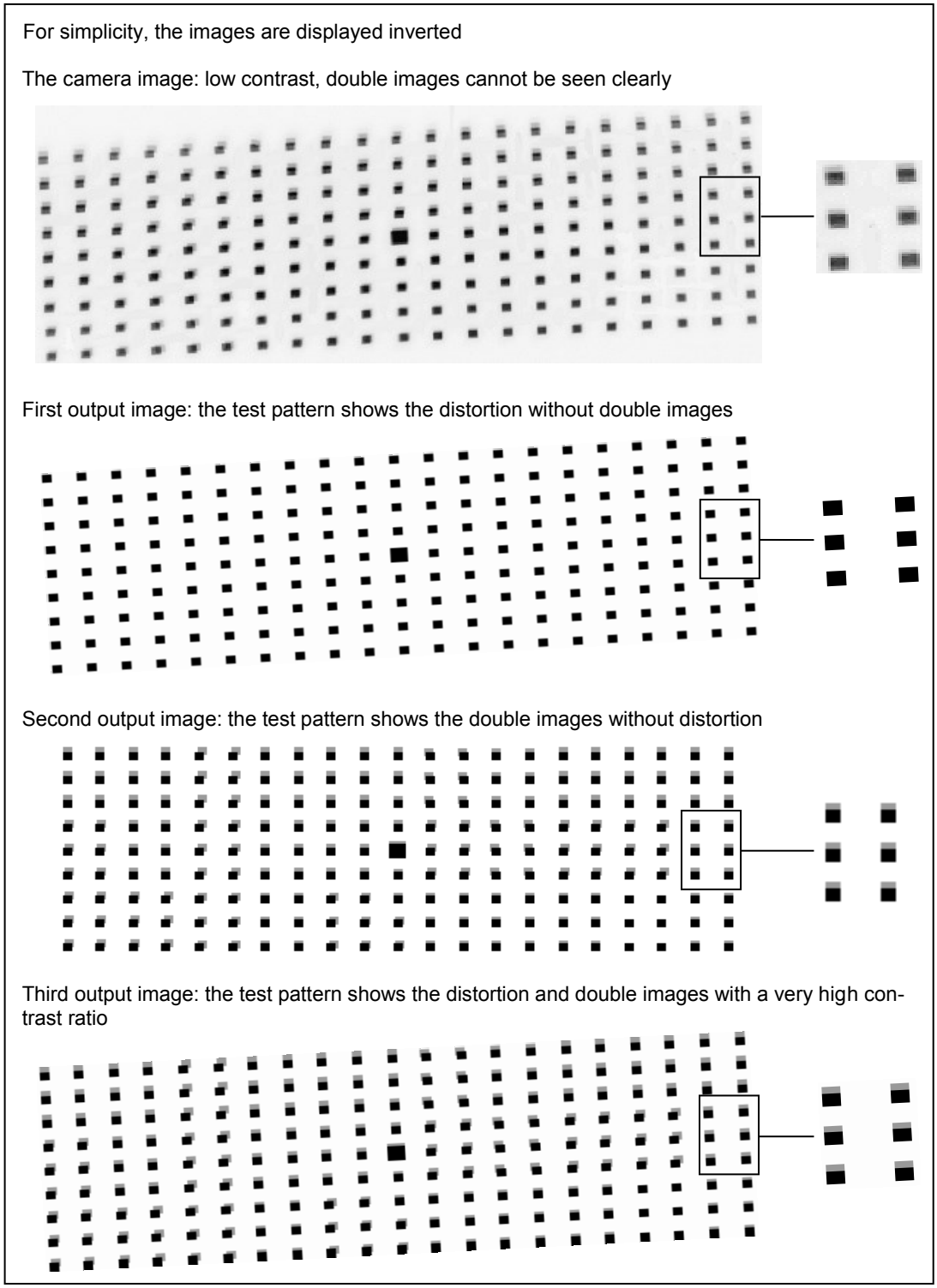


Figure 36: preparation of the image for subjective assessment

First, the distorted output image without double images is generated. Based on the 9 x 21 centre points the reference image is distorted. This is possible because the centre points contain the complete distortion information of the image. Based on already available functions of the Daimler AG, the distortion of the point cloud is transferred point by point to the reference image. Likewise, the distortion information can be transferred by this procedure to any arbitrary image, as shown in Figure 37.

Based on the 9 x 21 centre coordinates, which contain the complete information, each arbitrary image can be distorted approximately.

Own illustration.

Figure 37: the test image and the HUD image show the same distortion

The 2nd output image is generated from the double image distances determined by image processing routines. Based on these distances, a grey square is placed behind each existing white square of the reference image. This Matlab routine is designed especially for this thesis. Thus, the 2nd output image shows only the occurring double image distances on the undistorted test pattern.
The 3rd output image is created based on the 2nd output image and the existing functions for generating a distorted image. The distortion of the 9 x 21 centre points is transferred to the 2nd output image. Like the original camera image, the last output image shows the combination of distortions and double images only with a high contrast ratio.

The developed Matlab routines are applied to all images of the database. Altogether, 71895 distorted images without double images, 64410 undistorted test patterns with the associated double images, and 64410 images showing combined distortions and double images are available. These images form the foundation for the subjective assessment.

4.5 Test environment for the assessment of the image quality

Now, a test environment is needed to assess the processed images of the database. Here, it is important to ensure the same conditions for all test persons. For this reason, it is strongly advised not to assess the images in the experimental setup, described in chapter *4.2*. The resulting aberrations in the virtual image are dependent on the seating position of the test persons. For each seating position, the driver looks through a different area of the windscreen. The resulting aberrations depend on the viewing direction through the screen. The problem is that it cannot be guaranteed that each test person looks through the same area in the windscreen. Thus, the experimental setup is only partially suitable for assessing the image quality.
Thus, a special test environment is used. The construction of the test environment is based on the recommendation ITU-R BT.500-13. The BT.500-13 gives methodologies for assessing the quality of television pictures. Test methods, grading scales, and viewing conditions are also included [ITU-R BT 12]. During this thesis, the recommendation is applied to assess the quality of HUD images.

The test environment simply consists of a monitor and a chair, as shown in Figure 38. It is designed in such a way that the conditions are approximately the same than in real vehicles. The test person takes the chair, which is 2 m away from the monitor. Different images (size: 210 x 70 mm) are displayed on the monitor by the test leader. To ensure that the light conditions remain the same, the test is carried out in a darkened room. A black display screen is shown between 2 images. Thus, successive images are not directly comparable. This procedure corresponds to a proven standard of the Daimler AG [HELLMANN 12].

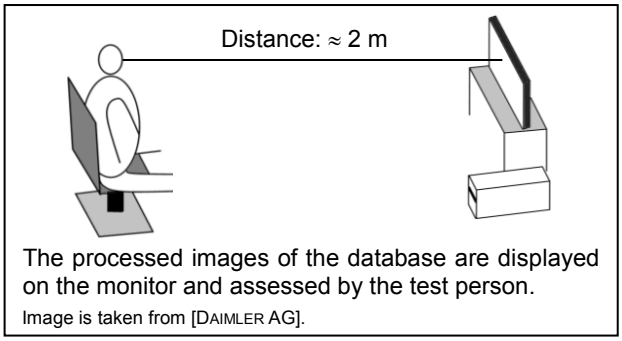


Figure 38: test environment for the subjective assessment of the images

The subjective assessment is recorded by applying the double-stimulus impairment scale method recommended by the ITU-R BT.500-13. This method is initially used to measure the robustness of television systems. For this thesis, the method is adapted to quantify the effect of failure characteristics in HUD images. First, the test persons see the unimpaired reference image and afterwards the faulty images. The participants are asked to vote on the defective image, keeping in mind the first. During the questioning, the test persons are presented with a series of images. These images are in a random order and with random impairments covering all required combinations. In addition, the unimpaired image is included in the set of images to be assessed [ITU-R BT 12: p. 10]. Assessed is the overall impression given by the image according to the 5-grade impairment scale [ITU-R BT 12: p. 11f], see Table 1.

Assessment	Quality	Impairment [ITU-R BT 12: p. 11]
1	bad	very annoying
2	poor	Annoying
3	fair	slightly annoying
4	good	perceptible, but not annoying
5	excellent	Imperceptible

Table 1: 5 divided rating scale recommended by ITU-R BT.500-13

Thereby, 5 points are equal to the imperceptible level of impairment and 1 point stands for the very annoying level [ITU-R BT 12: p. 10-13]. In addition, it is defined that the ratings 3, 4 and 5 correspond to an acceptable (AC) image quality. In contrast, the ratings 1 and 2 correspond to an unacceptable (UAC) image quality.

4.6 Selection of suitable test persons

The test environment, described in the previous section, is now used by different test persons to assess the quality of different aberrations. To get meaningful assessments, the test persons have to meet the following requirements:

- Test persons should not be professionally engaged in the development of HUDs.
- All age groups should be represented.
- The test person must have a driving licence and drive regularly longer distances.
- The test person should be technology interested.

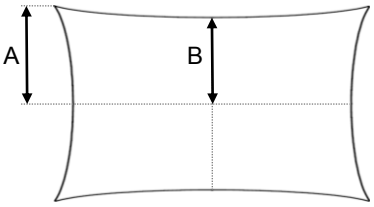
For this thesis, 12 test persons are interviewed. For time and cost reasons, unknown test persons are not questioned. Here, family members and friends are hired. The average age of the test persons is 38.1 years. The youngest test person is 19 years old and the oldest is 62 years old. 7 female and 5 male test persons are interviewed. All test persons drive daily several kilometres and are enthusiastic about technology.

The ITU-R BT.500-13 states that at least 15 test persons should be asked. For studies with limited scope, fewer than 15 test persons are enough [ITU-R BT 12: p. 8]. Since 12 test persons are asked, the standard is almost met.
The guideline of the Daimler AG recommends the questioning of at least 30 test persons [Daimler AG, department for product acceptance]. For time and cost reasons, this number is not reached in this thesis. Likewise, it is desirable not to use friends and family members as test persons. It would be perfect to interview test persons who drive a Mercedes car and consider using a HUD system in the near feature.

4.7 Standard method to assess the HUD image quality

The usual industry standard method to determine the customer suitability of HUD images is the limit analysis. It desires to specify ergonomic limits for the objective features. For each feature, an upper limit and a lower limit is set. Compliance with these limits, an impairment-free reading of the virtual image is guaranteed [EICHHORN & ZINK 12], [SCHNEID 09: p. 23].

Occurring double images are determined by the respective pixel-offsets [SCHNEID 09: p. 21]. In contrast, occurring distortions are described by a common method named TV-Distortion. This method is defined in the SMIA specification §5.20 [SMIA 04: p 61]. It is normally used to describe optical aberrations. Thus, the difference in height of the image corners and the image centre is calculated [SCHNEID 09: p. 21f], as shown in Figure 39.



$$TV\_Dist = \frac{A - B}{B} \cdot 100\%$$

The reported value is the average distortion of the 4 image corners.

Image is taken from [SMIA 04:

Figure 39: SMIA TV-Distortion

In relation to the HUD, distortions are not limited to aberrations caused by the optical system. The distortions are rarely symmetrical, they irregularly occur on all 4 sides. Therefore, new features are implemented, which are based on the geometric description [KÖPPL et al. 16], see chapter *5.1.*

The limit consideration method has a great weakness. Even simple combinations of various aberrations cannot be evaluated, see chapter *5.2*. Thus, methods of machine learning (nearest neighbour classifier, polynomial classifier, and learning vector quantisation) are applied, which no longer show this weak point, see chapter *7.*

Chapter

5 Study of the subjective perception

This chapter is divided into 2 main parts. First, objective features are presented that numerically describe occurring distortions and double images. The second part gives an overview of the subjective perception of individual and combined aberration types. Here it is shown that the limit consideration, explained in chapter *4.7*, is not suitable for the quality assessment of the HUD images.

5.1 Determination of objective evaluation features

To detect occurring distortions and double images, 21 objective features (13 features to detect distortions and 8 features to detect double images) are introduced. The basis for the calculation is the result of image processing routines, introduced in chapter *4.2*. The calculation instructions for the objective features are related to already existing rules of the Daimler AG [HELLMANN 12]. For this thesis, the calculation rules are adapted to the subjective perception of the images in the database.

To make distortions measurable, 13 objective features are used that describe the geometrical form of HUD images. The calculation is based on the 9 x 21 centre coordinates of the test pattern. It is necessary to design lines according to a linear or parabolic fit through the centre points. The curves are designed based on the mathematical scheme of minimising least squares [EICHHORN & ZINK 12]. Using the properties of these curves, the objective features can be determined. The features capture distortion types like rotation, trapezoid, aspect deviation [EICHHORN & ZINK 12], misalignment, smile, and local magnification [NEUMANN 12: p. 17, 33], see Table 2. Each feature is supplied with a unique number (Table 2, column 3), which is referenced in the course of the document. The first 3 values describe the size difference. The following 2 features are used to determine local deviations. Finally, the contour of the image is captured with the remaining 8 values.

Image property	Objective feature	Feature no.
Size difference	Adjusted in width	1.1
	Adjusted in height	1.2
	Aspect deviation	1.3
Local magnification	Enlargement horizontal	2.1
	Enlargement vertical	2.2
Contour	Rotation	3.1
	Misalignment horizontal	3.2
	Misalignment vertical	3.3
	Trapezoid horizontal	3.4
	Trapezoid vertical	3.5
	Smile horizontal top	3.6
	Smile horizontal bottom	3.7
	Smile vertical	3.8

Table 2: objective features to capture distortions numerically

The description of double images requires 8 objective features. Image processing routines provide horizontal and vertical double image distances for the 9 x 21 measuring marks. These distances are the basis for the calculation of statistical measures [SCHNEID 09: p. 21]. The maximal, average, 95% quantile, and the 80% quantile are determined, shown in Table 3. The features for horizontal and vertical distances are also supplied with unique numbers.

Distance property	Objective feature	Feature no.
Maximal	Horizontal	1.1
	Vertical	2.1
Mean	Horizontal	1.2
	Vertical	2.2
95% quantile	Horizontal	1.3
	Vertical	2.3
80% quantile	Horizontal	1.4
	Vertical	2.4

Table 3: objective features to capture double images numerically

It is ensured that the values of the features are directly comparable. So, all values are converted into HUD-pixels (0.58 mm x 0.58 mm), see chapter *3.4.* Possible problems with further processing (clustering and classification) are thus avoided.

5.2 Investigate the subjective assessment of aberrations

The 21 objective features, described in the previous section, can be used to specify the quality of any virtual HUD image. The next step is to find a relation between these characteristic values and the subjective perception. It is investigated, from which occur-

rence different aberration types are recognised and perceived as disturbing [HELLMANN 12]. For this purpose, some test person questionings are necessary. Thus, various images with different degrees of aberrations are subjectively assessed. The shown images are all produced artificially. The images show the test pattern, according to Figure 34, on a black background. The questionings are realised in the test environment intended for it, see chapter *4.5*. 12 test persons are interviewed to assess the impression of the images on the 5-divided rating scale recommended by the ITU.

Individual distortion types are generated by affine transformations. Affine transformations are explained in detail in [WOLLBERG 90: p. 41-49]. Here, the 9 x 21 centre point coordinates of the test pattern are distorted by the transformation matrix. Afterwards, the distorted centre points are converted back into an image. Special Matlab routines that generate distorted images by affine transformations are already available. They are provided by the Daimler AG. For this thesis, the existing program code is adapted to the 13 evaluation features for distortion.
Double images are created by placing grey coloured squares behind the existing measuring marks. These Matlab routines are not available and are specially developed for this thesis.

5.2.1 Perception of different distortion types

The subjective perception of individual distortions types is investigated in the first survey. Combinations of different distortion types remain unconsidered. 41 images are shown to 12 test persons. Time requirement is about 25 minutes per person. As an example, a few images of the first survey are shown in Figure 40.

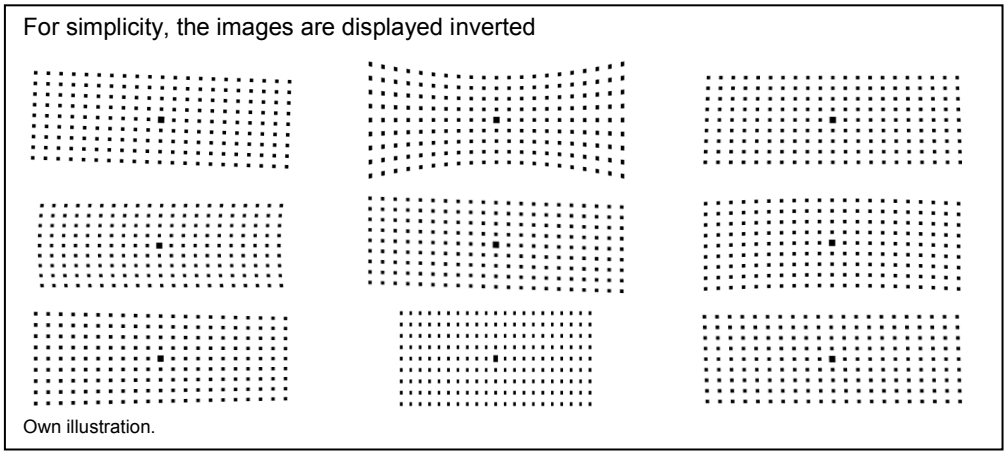


Figure 40: example images shown in the first survey

Among the distorted images, 1 undistorted image is shown, according to [ITU-R BT 12: p. 11]. Additionally, 1 distorted image is presented twice. After assessing the images, the test persons are asked 2 open questions: “Which type of distortion is the easiest to perceive?” and “Which type of distortion bothers you the most?”. The survey has a similar structure to an investigation already carried out at the Daimler AG

[HELLMANN 12]. The difference is that the shown images are specially adapted to the used features.

The evaluation of the survey is like the analysis of a recent survey carried out by [HELLMANN 12]. The obtained labels of the undistorted image and the labels of the twice-rated image are summarised in Table 4. Shown are the respective labels for each test person. The undistorted image is rated on average with 4.8 rating points. The score among the test persons varies between 4 and 5 rating points and the standard deviation is 0.5 points.

Image	Ratings of the test person											
	1	2	3	4	5	6	7	8	9	10	11	12
Undistorted	5	5	4	5	5	4	5	5	5	4	5	5
Distorted rated once	3	2	3	3	3	1	2	3	2	3	2	3
Distorted rated twice	3	3	3	4	3	2	3	3	2	3	3	3

Table 4: obtained labels of the undistorted and twice rated images

By comparing the labels of the twice-rated image, it is found that the score differs on average by 0.4 points. The first time the distorted image is rated on average with 2.5 points (standard deviation 0.7 points). After repeating the assessment of the same distorted image, an average score of 2.9 points (standard deviation 0.5 points) is obtained. From these findings, a limit can be defined for perceiving and accepting occurring individual aberration types, as shown in Equation 5. The calculation rules are introduced in [HELLMANN 12].

$$\textit{Perceptual limit} = \textit{maximal possible score} - \textit{scattering double rated images}$$
$$PLM = 4.8 - 0.4 \quad \Longrightarrow \quad PLM = 4.4\ \textit{[rating points]}$$

$$\textit{Acceptance limit} = \textit{score images just acceptable} + \textit{scattering double rated images}$$
$$ALM = 3.0 + 0.4 \quad \Longrightarrow \quad ALM = 3.4\ \textit{[rating points]}$$

Calculation rules correspond to the procedure at the Daimler AG [HELLMANN 12].

Equation 5: definition of perceptual limit and acceptance limit

An undistorted image is on average rated with 4.8 points. Thus, the perception limit is corrected to 4.4 points by the scattering of the double rated image. Here, a single aberration type is only detected, if the achieved assessment is less than 4.4 rating points. Likewise, the limit for acceptance is defined. The used rating scale implies that a labelling of 3 points is just acceptable. Due to the scattering of double rated images, the acceptance limit is increased to 3.4 points. If an assessment falls below this limit, the rated image is no longer customer suitable.

The remaining 39 images show 3 different characteristics for each distortion type. The resulting relationship between the feature values and the average subjective ratings is analysed by a best fitting straight line. The design of the straight line is based on the mathematical scheme of minimising least squares. Here, a prediction equation is developed for each of the 13 distortion types. The subjective labels are represented by the variable *y*, and the variable *x* contains the value of the feature. It must be ensured that the maximum value of the subjective rating does not exceed a value of 4.8 rating points. This score represents the highest possible rating of an undistorted image. Based on these straight-line equations, it is possible to determine values for the perception and acceptance. Thus, the intersection point between the line and the perception limit as well as the intersection with the acceptance limit is calculated [HELLMANN 12]. For example, Figure 41 shows the calculation for feature no. [2.1].

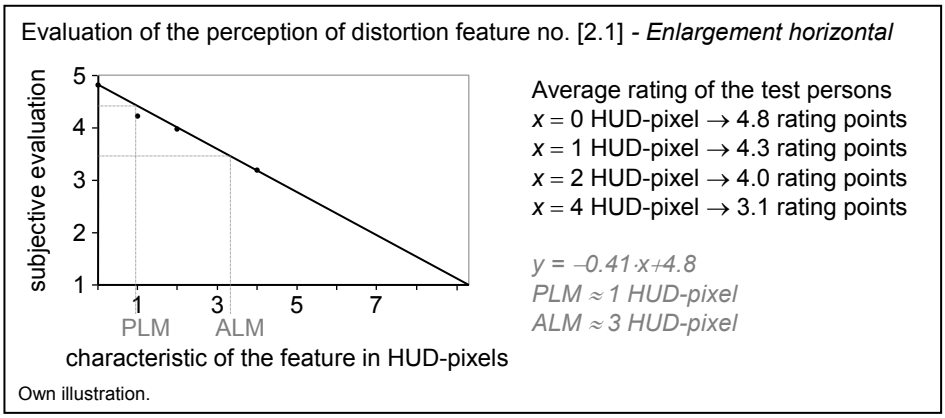


Figure 41: determination of the perceptual equation and perception limits

The resulting prediction equations and the associated perceptual and acceptance limits are summarised in Table 5. The resulting limit values are rounded to full HUD-pixels.

Feature no.	Prediction equation	Perceptual limit 4.4 rating points	Acceptance limit 3.4 rating points
Size difference [1.1]	$y = -0.01 \cdot x + 4.8$	± 40 HUD-pixel	± 140 HUD-pixel
Size difference [1.2]	$y = -0.03 \cdot x + 4.8$	± 13 HUD-pixel	± 47 HUD-pixel
Size difference [1.3]	$y = -0.05 \cdot x + 4.8$	± 8 HUD-pixel	± 28 HUD-pixel
Local deviations [2.1], [2.2]	$y = -0.41 \cdot x + 4.8$	± 1 HUD-pixel	± 3 HUD-pixel
Contour [3.1], [3.7]	$y = -0.26 \cdot x + 4.8$	± 2 HUD-pixel	± 5 HUD-pixel
Contour [3.2]	$y = -0.25 \cdot x + 4.8$	± 2 HUD-pixel	± 6 HUD-pixel
Contour [3.3], [3.8]	$y = -0.53 \cdot x + 4.8$	± 1 HUD-pixel	± 3 HUD-pixel
Contour [3.4], [3.5]	$y = -0.60 \cdot x + 4.8$	± 1 HUD-pixel	± 2 HUD-pixel
Contour [3.6]	$y = -0.80 \cdot x + 4.8$	± 1 HUD-pixel	± 2 HUD-pixel

Table 5: evaluation of the perception of single distortion types

The answers to the questions “Which type of distortion is the easiest to perceive?” and “Which type of distortion bothers you the most?” are shown in Table 6 [KÖPPL et al. 16].

To determine a subjective ranking, the features are assigned to the statements of the test persons, see Table 6 column 4 and 5. The analysis of the answers shows that the distortion, which is perceived first, is also the most annoying one. Skewing is mentioned first, followed by bending, discomposure, a tapered shape, and a stretched image.

Answers from participants	Easily noticed	Disturbing noticed	Evaluation feature	Feature no.
Skewing	67%	75%	Contour	[3.1], [3.2], [3.3]
Bending	50%	59%	Contour	[3.6], [3.7], [3.8]
Discomposure	42%	50%	Local deviations	[2.1], [2.2]
Tapered	25%	17%	Contour	[3.4], [3.5]
Stretching	9%	0%	Size difference	[1.1], [1.2], [1.3]

Table 6: subjective ranking of the objective features for distortion

During the second investigation, the test persons see 6 images that show combinations of 2 individual distortion types. The investigation is similar to a study carried out by [HELLMANN 12]. The required time is about 5 minutes per person. The average ratings of the images are summarised in Table 7. The ratings of the single distortion types are already determined in the previous experiment and shown in the left part of the table. The right part shows the assessments of the combination of the corresponding 2 distortion types.

Single distortion types				Combination Rating
Feature, occurrence	Rating	Feature, occurrence	Rating	
[3.1], 2 HUD-pixel	**4.2**	[3.3], 1 HUD-pixel	**4.5**	**3.2**
[3.1], 2 HUD-pixel	4.2	[3.5], 1 HUD-pixel	4.2	3.9
[3.1], 2 HUD-pixel	4.2	[3.7], 2 HUD-pixel	4.3	4.0
[3.3], 1 HUD-pixel	4.2	[3.5], 1 HUD-pixel	4.5	4.3
[3.3], 1 HUD-pixel	4.5	[3.7], 2 HUD-pixel	4.3	4.2
[3.5], 1 HUD-pixel	**4.2**	[3.7], 2 HUD-pixel	**4.3**	**3.3**

Table 7: evaluation of the perception of 2 combined distortion types

The average ratings of the single distortion types are all well above the acceptable limit of 3.4 rating points. The images are customer suitable. In contrast, not all average ratings of images showing a combination of these distortion types are above the crucial acceptance limit, see Table 7 row 1 and 6. Thus, it is possible that images showing combinations of single acceptable distortion types are assessed not to be customer suitable.

This simple evaluation already shows that the rating of images showing combined distortion types is not completely possible when using the limit consideration method. If the distortion type occurs alone, the image quality is assessed as acceptable. Another time, when the distortion type occurs in combination with other aberrations, the image

quality is assessed as unacceptable. This cannot be mapped to a simple limit consideration. Since it is assumed that, the compliance with meaningful limits guarantees the impairment-free reading of the virtual image.

5.2.2 Perception of different double image types

To analyse the perception of different double image types, another customer survey is executed. The survey is similar to an investigation carried out by [HELLMANN 12]. By the way of example, 3 images shown in the survey are illustrated in Figure 42.

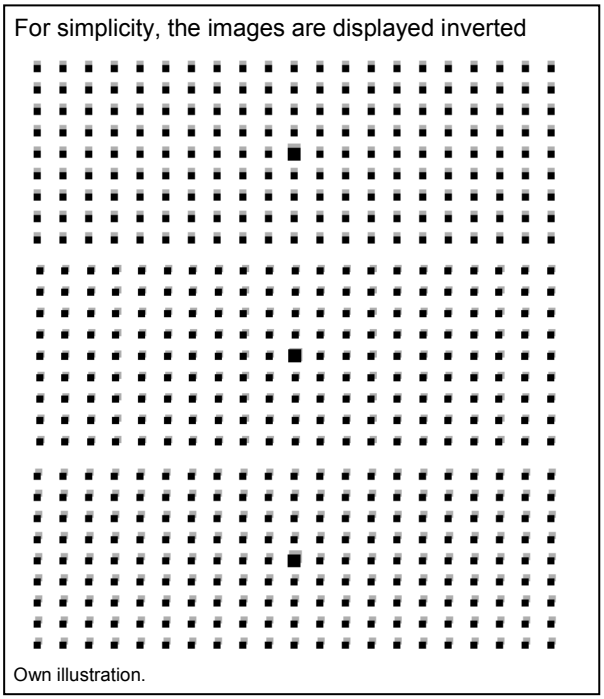


Figure 42: example images shown in the second survey

In the first step, 10 images are shown to 12 test persons. The time required is about 10 minutes per person. Only the mean (feature no. [1.2], [2.2]) and the maximal (feature no. [1.1], [2.1]) double image distances are analysed in more detail. Combinations of different double image distance types are not considered for the time being. The participants are asked to assess the quality of the shown images. In addition, they are interviewed if double images can be seen.

The evaluation of the survey is like the evaluation of a recent survey carried out by [HELLMANN 12]. The resulting assessments are summarised in Table 8. Here, the first 2 columns specify the presented double images. The other 2 columns show the average rating and the detection rate. It is noticeable that less than 45% of the test persons recognise horizontal or vertical double image distances of 1 HUD-pixel. However, no one who can perceive the double image will accept it. A double image at a distance of 1 HUD-pixel is perceived as blurring. If the distances increase, 2 separated images are identified. Thereby, the test persons mention that looking at blurry images causes

headaches and eyestrains. These findings are consistent with the outcomes in [SCHNEID 09: p. 23f] and [HELLMANN 12].
When comparing the labels, it is found that the perception of horizontal and vertical double images is almost identical. The labels vary by less than 0.4 rating points, which is the maximum scattering of the doubled rated image, see chapter *5.2.1*.

Double image Feature no.	Sizes	Average rating	Detection rate
Horizontal [1.1]	1 HUD-pixel	3.8	25%
Vertical [2.1]	1 HUD-pixel	4.1	17%
Horizontal [1.1]	2 HUD-pixel	3.6	33%
Vertical [2.1]	2 HUD-pixel	3.9	25%
Horizontal [1.2]	1 HUD-pixel	3.9	42%
Vertical [2.2]	1 HUD-pixel	4.0	33%
Horizontal [1.2]	2 HUD-pixel	2.7	83%
Vertical [2.2]	2 HUD-pixel	2.9	75%
Horizontal [1.2]	3 HUD-pixel	1.8	92%
Vertical [2.2]	3 HUD-pixel	1.8	92%

Table 8: survey results for the perception of individual double image distance types

Like the approach of [HELLMANN 12] and the analysis of single distortion types, the relationship between the double image feature values and the average subjective ratings is analysed according to the best fitting straight line. Here, 2 prediction equations, for horizontal and vertical double image distances, are defined. The variable y represents the subjective ratings. In addition, the variable x shows the different characteristic double image distances. Again, it must be ensured that the maximum value of the ratings does not exceed the highest possible label of 4.8 rating points. This score is determined in chapter *5.2.1* and represents the highest possible rating of a non-defective image.
Based on the straight-line equations, characteristic values for the perceptual limit and the acceptance limit can be calculated, see [HELLMANN 12]. The values result from the intersections of the line equation with the perception limit ($y = 4.4$ rating points) and the acceptance limit ($y = 3.4$ rating points); see Table 9. The derivation of these limits is presented in Equation 5.

Double image feature	Prediction equation	Perceptual limit	Acceptance limit
Horizontal	$y = -0.92 \cdot x + 4.8$	$\pm$ 1 HUD-pixel	$\pm$ 2 HUD-pixel
Vertical	$y = -0.85 \cdot x + 4.8$	$\pm$ 1 HUD-pixel	$\pm$ 2 HUD-pixel

Table 9: evaluation of the perception of single double image features

In the second part of the survey, the perception of 2 combined double image distance types is investigated. 6 images are labelled. The required time is about 5 minutes per person. The combinations of mean (feature no. [1.2], [2.2]) and maximal (feature no. [1.1], [2.1]) double image distances are rated. After labelling the images, the 12 test persons are asked 2 open questions: “Which kind of double image distances are per-

ceived fastest?" and "What are the most annoying double image distances?". This survey has a similar structure to an investigation already carried out at the Daimler AG [HELLMANN 12].

The evaluation of the 2 open questions shows that large double image distances that are evenly distributed over the entire image (large values of feature no. [1.2], [2.2]) are perceived first and are not acceptable. In contrast, single outliers (large values of feature no. [1.1], [2.1]) are acceptable if the average double image distances are small (low values of feature no. [1.2], [2.2]).
The detailed evaluation of the survey is given in Table 10. The first 4 columns show the ratings of the single double image types. The last column contains the average scores of the combination of the corresponding double image distance types. The ratings of the single double image distance types are taken from the previous survey. Looking at the average labels of the combinations, all shown images are rated with less than 3.4 points. Thus, the ratings are below the acceptable limit. It is possible that images, which show combinations of single acceptable double image types, are assessed as not acceptable; see Table 10, row 1, 3 and 5.

Single double image distance types				Combination Rating
Feature, occurrence	Rating	Feature, occurrence	Rating	
[1.1], 1 HUD-pixel	**3.8**	[1.2], 1 HUD-pixel	**3.9**	**2.9**
[1.1], 2 HUD-pixel	3.6	[1.2], 2 HUD-pixel	2.7	2.4
[2.1], 1 HUD-pixel	**4.1**	[2.2], 1 HUD-pixel	**4.0**	**3.2**
[2.1], 2 HUD-pixel	3.9	[2.2], 2 HUD-pixel	2.9	2.6
[1.2], 1 HUD-pixel	**3.9**	[2.2], 1 HUD-pixel	**4.0**	**3.1**
[1.2], 2 HUD-pixel	2.7	[2.2], 2 HUD-pixel	2.9	1.3

Table 10: evaluation of the perception of combined double image distance types

This study makes clear that the assessment of images showing combined double image types is not completely possible when using the limit consideration method. Once the double image value is considered as acceptable and another time as not acceptable. This cannot be represented by a simple limit consideration.

5.2.3 Perception of combinations of distortions and double images

Now, images containing both distortion and double images are shown to the test persons. This survey is similar an investigation already carried out at the Daimler AG [HELLMANN 12]. 2 different distortion types are combined sequentially with 3 different double images. Thus, the study consists of 6 images, which are again assessed by 12 test persons. A test person needs about 10 minutes for this. After the test persons have rated the images, the open question "What would you rather accept, a distorted image or an image with double vision?" is asked [HELLMANN 12].

The analysis of this question shows that 75% of all test persons perceive distortions first. Also, 67% of all participants would rather accept distortions than double images. The reason commonly given by the participants is that looking at blurry images causes headaches and eyestrain.
The results of the survey are listed in Table 11. The ratings of the single distortions and double images are taken from the previous experiments. These ratings are shown in the front columns. The last column contains the average scores for combined distortion and double image distances.

Single distortion types		Single double images		Combination Rating
Feature, occurrence	Rating	Feature, occurrence	Rating	
[3.3], 1 HUD-pixel	**4.5**	[1.2], 1 HUD-pixel	**3.9**	**3.3**
[3.1], 2 HUD-pixel	**4.2**	[1.2], 1 HUD-pixel	**3.9**	**3.1**
[3.2], 7 HUD-pixel	2.9	[1.2], 1 HUD-pixel	3.9	2.6
[3.3], 1 HUD-pixel	4.5	[2.2], 1 HUD-pixel	4.0	3.4
[3.1], 2 HUD-pixel	4.2	[2.2], 1 HUD-pixel	4.0	3.5
[3.2], 7 HUD-pixel	2.9	[2.2], 1 HUD-pixel	4.0	2.5

Table 11: evaluation of the perception of combined distortions and double images

The average labels of the resulting combinations are less than or equal to the acceptance limit of 3.4 rating points. It is possible that images showing combinations of acceptable distortion types and acceptable double image types are assessed as not acceptable, shown in Table 11, rows 1 and 2. Similar to the previous studies, this investigation demonstrates that images showing combinations of acceptable aberrations can be considered unacceptable. Likewise, a limit consideration to assess the image quality does not make sense.

5.3 Impact of the results on the assessment system

As the investigation in the previous sections shows, an assessment of the image quality by using the limit consideration method is only possible to a limited extent. This is because images showing a combination of different aberration types cannot be assessed by the limits. It is possible that combinations of single acceptable aberration types are considered unacceptable. Thus, the aberration types are once considered as acceptable and another time as unacceptable.

In the course of this thesis, other assessment methods will be used. Supervised learning methods for assessing combined aberration types are implemented.

Chapter

6 Used processes and methods

The main goal of this thesis is to develop a novel approach to assess the perceived quality of HUD images. Here, methods of machine learning are used. Thus, each head-up display image is represented as a data point in the *q*-dimensional feature space. 21 features are used. 13 features to describe distortions, and 8 features to capture double images, see chapter *5.1*. The feature values are written as *q*-dimensional feature vector v; $v = (v_1\ v_2\ \ldots\ v_q)^T$ [SCHÜRMANN 96: p. 14]
The first step in implementing an assessment algorithm is to identify the perceived quality of some representative images. The perceived quality is determined by empirical studies and mapped to labels *y*. The labels correspond to the 5-grade impairment scale of the ITU-R BT.500-13 directive, see chapter *4.5*. Representative images are selected by clustering methods (ward, k-means, and mean-shift). Thus, the database is structured and reduced to a few characteristic images. The database is described in chapter *4.3*.
Finally, the objective feature vectors *v* describing the HUD image quality are linked to the subjective labels *y* by classification algorithms (nearest neighbour classifier, polynomial classifier, and learning vector quantisation). Here, the following learning rules are applied: supervised learning, semi-supervised learning, and active learning.

This chapter introduces the applied processes and methods. It begins with the procedure of empirical studies and an overview of machine learning methods. The last part of this chapter describes the used clustering and classification methods in detail.

6.1 Conception of empirical studies

The word empirical stands for based on experience [ATTESLANDER 03: p. 3]. In contrast to theory, empirical statements are not sufficiently proven in practice. However, the transmission from empirical experience or knowledge to theoretical awareness is gradual [HÄDER 10: p. 22].
Here, the results of some empirical studies are the basis for implementing an assessment system for HUD images. The preparation and analysis of these studies are achieved by methods of machine learning.

The course of an empirical study can be roughly divided into 5 main phases, which will be explained in more detail below [DIEKMANN 95: p. 187].

- *Formulation and clarification of the research problem:* the aim of the research problem identification is the detailed answer to the question: “What should be investigated exactly?” [DIEKMANN 95: p. 187].
  In this thesis, the perceived quality of HUD images is investigated in detail.

- *Planning and preparation of the survey:* in this phase, the required survey tools are defined. Similarly, the research features are determined and assigned to measurable and assessable indicators [HÄDER 10: p. 76].
  Here, this phase involves selecting representative images that should be subjectively assessed. This is achieved by unsupervised learning methods like ward, k-means and mean-shift clustering.

- *Data collection:* the data acquisition is done by volunteer surveys [HÄDER 10: p. 76].
  The choice and the number of test persons depend on the complexity of the assessment task, the defined quality features, and the time and financial resources [GENUIT 10: p. 141f]. This phase ends with creating a machine-readable dataset [HÄDER 10: p. 76].
  In this thesis, 12 test persons, who must meet certain requirements, are interviewed. The questioning takes place in a special test environment. The equipment of the test environment is based on the recommendations of the ITU-R BT.500-13. Detailed information can be found in chapters *4.5* and *4.6.*

- *Data analysis:* the evaluation of the study includes the preparation of tables, overviews, and statistical calculations [HÄDER 10: p. 76].
  In this thesis, the relationship between the subjective labels and the objective features is analysed. For this purpose, classification algorithms such as the nearest neighbour classifier, the polynomial classifier, and the learning vector quantisation are used.

- *Reporting and documentation:* the completion of the study includes the documentation of the process and the obtained research results [HÄDER 10: p. 76].

The statistical literature recommends the regression analysis as a standard method for evaluating empirical studies [DIEKMANN 95: p. 660], [HÄDER 10: p. 435], [ATTESLANDER 03: p. 298]. Based on the perceived quality of HUD images, classification methods are used instead.

6.2 Methods of machine learning

The volunteer studies are prepared and evaluated by using methods of machine learning. Machine learning deals with the artificial generation of knowledge from experience. An artificial system learns from known examples. After completing the learning phase, the system is able to generalise to unknown examples [MARSLAND 11: p. 5f]. In this thesis, 4 types of learning are applied, namely, unsupervised learning, supervised learning, semi-supervised learning, and active learning.

6.2.1 UL: unsupervised learning

The unsupervised learning (UL) algorithm attempts to find similarities between data points [MARSLAND 11: p. 6f]. For this, no teacher provides supervision how the single data points should be handled. Thus, only unlabelled data are used. General objectives of unsupervised learning algorithms are clustering and reducing the amount of data [ZHU & GOLDBERG 09: p. 2f], as shown in Figure 43.

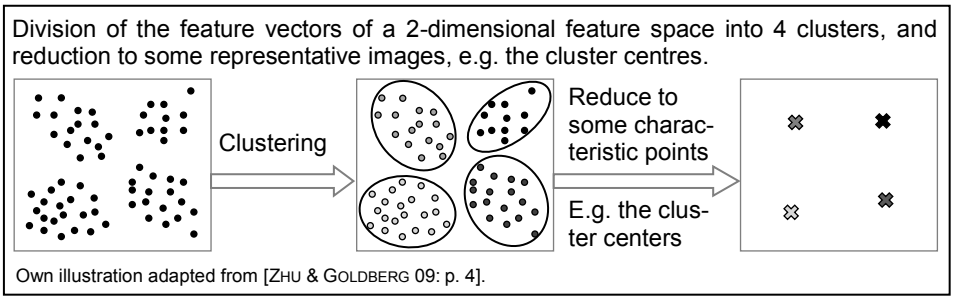


Figure 43: principle of a cluster analysis

Clustering methods are always relevant to subdivide a set of data points $U = \{v\}$ into homogeneous groups called clusters. The clusters are not available at the time of collection. The goal is to find groups of similar data points without explicitly saying that these data points belong to the same cluster. Instead, the algorithm has to discover the similarities by itself [MARSLAND 11: p. 195f]. The data points are clustered in such a way that the difference within a cluster is as small as possible and the difference between the clusters as large as can be done [BORTZ & SCHUSTER 10: p. 453]. After grouping, the clusters can be reduced to a few characteristic points. The cluster centres are best suited for this purpose [ZHU & GOLDBERG 09: p. 2f].

6.2.2 SL: supervised learning

A training set of data points with the correct labels is provided. From this, the algorithm derives generalisations in order to correspond correctly to possible inputs, as shown in Figure 44.

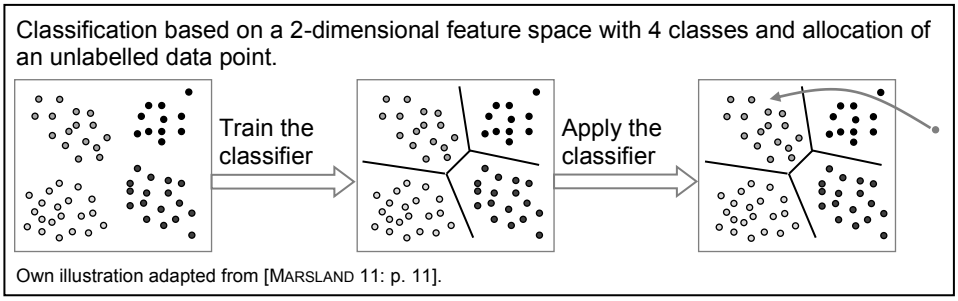


Figure 44: principle of a classification approach

Depending on the domain of labels, the supervised learning problem can be subdivided into classification and regression, as shown in Figure 45. Classification is the supervised learning problem with discrete labels that correspond to the classes. If a continuous domain of labels is used, the supervised learning problem is called regression [ZHU & GOLDBERG 09: p. 5].

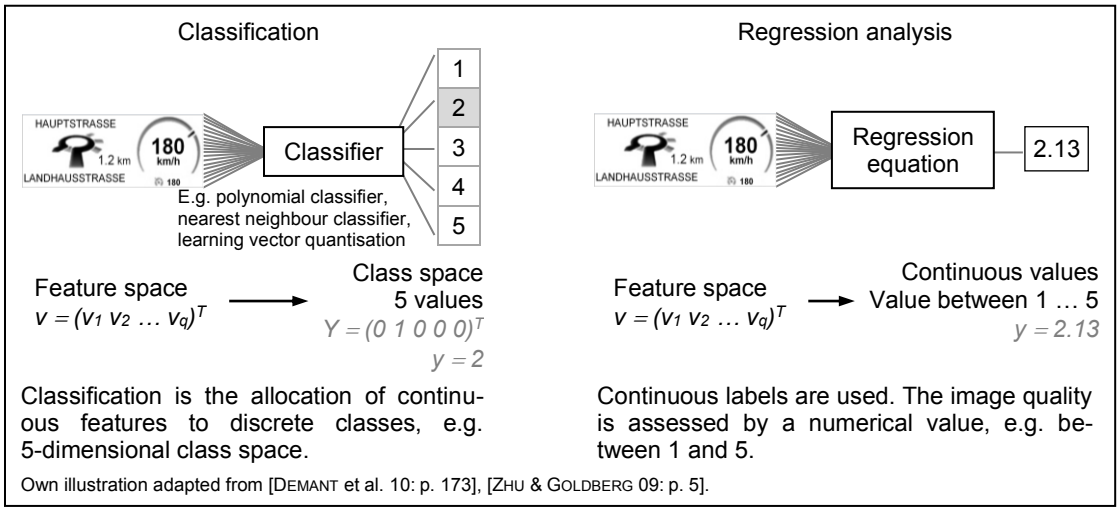


Figure 45: classification versus regression analysis

To obtain a good generalisation, supervised learning (SL) algorithms need a large number of labelled training samples preferably from a range of variation [MARSLAND 11: p. 6]. The supervised learning task is divided into the training and the testing phase. During the training phase, the relationship between the subjective labels and the objective feature vectors is detected. This requires a training set $S_{Train} = \{v; y\}$ consisting of objective feature vectors with associated labels. Upon completion of the learning phase, the supervised learning algorithm is able to recognize the learned relationship in unlabelled data. Thus, objective feature vectors receive the corresponding label. During the testing phase, the trained algorithm is applied to label an unlabelled test set[6] $S_{Test} = \{v\}$. To verify the quality of the trained algorithm, the determined labels y_{model} are compared with the actual given labels y_{given}. Therefore, the root mean square error (RMSE) is calculated. The RMSE is a measure of the difference between the values delivered by a

[6] The procedure of checking the quality of the labels, based on the training set is called reclassification. The procedure of checking it with an independent test set is called generalisation. Obviously, only the generalisation is relevant [SCHÜRMANN 96: p. 17f].

model and the values actually observed from the environment [GABLER 14], as shown in Equation 6. If all prognoses exactly apply, the RMSE is 0. The larger the RMSE value, the lower is the agreement of the labels with the subjective perception.

$$RMSE = \sqrt{\frac{1}{n} \cdot \sum_{i=1}^{n} \left(y_{model\ i} - y_{given i} \right)^2}$$

RMSE: root mean square error, *n*: number of samples in the test set, y_{model}: labels submitted by the evaluation system, y_{given}: labels corresponding to the subjective perception [GABLER 14]

Equation 6: root mean square error

The labels correspond to the 5-grade impairment scale of the ITU-R BT.500-13. The labels 1 and 2 represent an unacceptable image quality and the labels 3, 4 and 5 stand for an acceptable quality. The ability of the classification algorithm is quantified by counting the numbers of matches and matching failures see Table 12 [SZELISKI 10: p. 201].

Assessment of the image quality [SZELISKI 10: p. 201]		Assessments by the test persons	
		ACCEPTABLE	UNACCEPTABLE
Assessments by the algorithm	ACCEPTABLE	*TP*: true positive	*FP*: false positive
	UNACCEPTABLE	*FN*: false negative	*TN*: true negative

Table 12: confusion matrix to quantify the performance of the algorithm

Based on these matches and matching failures, the classification accuracy, the true positive rate (TPR) and the false negative rate (FPR) can be determined, as shown in Equation 7.

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \cdot 100 \ [\%]$$

$$TPR = \frac{TP}{TP + FN} \cdot 100 \ [\%], \ FPR = \frac{FP}{FP + TN} \cdot 100 \ [\%]$$

Accuracy: classification accuracy, TPR: true positive rate, FPR: false positive rate, *TP*: true positive, *FP*: false positive, *FN*: false negative, *TN*: true negative [SZELISKI 10: p. 202]

Equation 7: quality estimation of the classification results

If all prognoses are accurate, the classification accuracy is 1. If the accuracy decreases, the consistency between the quality estimation and the subjective perception decreases as well. The true positive rate should be close to 1, while the false positive rate should be close to 0.

6.2.3 SSL: semi-supervised learning

The manual labelling of many training samples is costly and time-consuming. In contrast, unlabelled data are easy to get and thus cheap. The goal of semi-supervised

learning (SSL) is to verify whether combining labelled and unlabelled data may improve the learning behaviour. Thus, semi-supervised learning is somewhere between unsupervised learning and supervised learning.
SSL is characterised by an iterative procedure in which the learning process uses its own predictions to teach itself. Given are labelled training data $L = \{v; y\}$ and the unlabelled data $U = \{v\}$. First, the algorithm is trained using L and then applied to predict the labels of U. A subset $S_{SSL} = \{v; y_{SSL}\}$ with the most confident predictions is removed from U $(U = U - S_{SSL})$ and added to L $(L = L + S_{SSL})$. The algorithm is retrained with the enlarged L and the procedure repeats. This is usually done until convergence, [HADY & FAROUK 11], [ZHU & GOLDBERG 09: p. 9-12].
There are 2 slightly different semi-supervised learning scenarios, the inductive and the transductive learning. The goal of inductive learning is to predict the labels of future test data. In contrast, the goal of transductive learning is to predict the labels of the unlabelled feature vectors in the training set [ZHU & GOLDBERG 09: p. 12].

6.2.4 AL: active learning

The key idea behind active learning (AL) is that the machine-learning algorithm selects the data from which it learns. Thus, a higher accuracy is achieved and the number of labelled training samples is reduced. Like semi-supervised learning, the active learning is based on the fact, that unlabelled data could be easily obtained, but labels are time-consuming to get [SETTLES 09]. The difference is that active learning attempts to cut training costs by selecting the informative training samples. In contrary, semi-supervised learning attempts to extrapolate previously acquired knowledge to unlabelled data, to increase the number of labelled training samples without causing extra costs [WUTTKE et al. 14], [YEH & GALLAGHER 08].
Active learning is an iterative procedure in which the learning process independently selects the data used for training. The algorithm asks an Oracle (e.g. human annotator, test persons) for the right labels. Thus, a set of labelled samples $L = \{v; y\}$ as well as a larger set of unlabelled samples $U = \{v\}$ is used as input. The algorithm is trained using L and then applied to predict the labels of U. A subset $S_{Active} = \{v\}$ for which the process will achieve a wrong classification by a high probability is removed from $(U = U - S_{Active})$ and the Oracle is asked for the right labels. From this subset $S_{Active} = \{v; y_{Oracle}\}$ the process learns the most. Thus, the subset is added to L $(L = L + S_{Active})$. The algorithm is retrained with the enlarged L and the procedure repeats. Therefore, the algorithm is trained without having more training samples than necessary. The learning process aims to keep the Oracle labelling effort to a minimum, it is only asked for advice if the "usefulness" of training is high [HAQUE et al. 13], [PERSELLO & BRUZZONE 12], [SETTLES 09].

6.3 Unsupervised learning: clustering analysis

Many clustering algorithms exist that differ by the applied similarity criterion [MARSLAND 11: p. 195f]. The following section shows common clustering methods for structuring the HUD images in the database. The database is described in chapter *4.3*. Thus, the images represented by the feature vectors *v* are divided into clusters. Subsequently, a representative image is selected from each cluster for personal interviews.

- *Ward:* initially, each data point forms a cluster of its own. From all possible pairs of clusters centres, the cluster centres $\overline{V_i}$ and $\overline{V_{i'}}$ whose combination results in the minimal increase of the sum of squared differences *QSe* of the corresponding elements are merged, as shown in Equation 8.

$$QSe = \frac{n_i \cdot n_{i'}}{n_i + n_{i'}} \cdot \sum_{j=1}^{q} \left(\overline{V_{i\,j}} - \overline{V_{i'\,j}} \right)^2$$ For merging cluster centres $\overline{V_i}$ and $\overline{V_{i'}}$

QSe: sum of squared differences, $\overline{V}$: cluster centre that corresponds to the mean vector of all elements in the cluster, *n*: number of images in a cluster, *q*: number of objective features to represent the image quality [BORTZ & SCHUSTER 10: p. 463]

Equation 8: sum of squared differences for merging to clusters

Merging of cluster pairs is stopped once a predefined threshold for the value of *QSe* is exceeded [BORTZ & SCHUSTER 10: p. 462-465].

- *K-means:* initially, a predefined number of cluster centres are chosen randomly. For all data points *v*, the (usually Euclidean, as shown in Equation 9) distance to each cluster centre is computed, and the cluster membership of each data point is set according to the cluster centre with the smallest distance.

$$d_{gi} = \left[\sum_{j=1}^{q} \left(v_{g\,j} - \overline{V_{i\,j}} \right)^2 \right]^{0.5}$$ Distance between feature vector v_g and cluster centre $\overline{V_i}$

d_{gi}: Euclidean distance, *v*: feature vector, $\overline{V}$: cluster centre that corresponds to the mean vector of all elements in the cluster, *q*: number of objective features to represent the image quality [BORTZ & SCHUSTER 10: p. 456]

Equation 9: Euclidean distance between feature vector and cluster centre

In the next step, the cluster centres are redefined by computing the mean vectors of the data points assigned to them. This procedure is repeated until the assignment of the data points to the cluster centres does not change anymore. The final clustering result depends on the initialisation [BORTZ & SCHUSTER 10: p. 465f].

- *Combine ward and k-means:* it is often favourable to utilise the result of the ward algorithm as an initialisation to the k-means algorithm. This circumvents the property of the ward algorithm that the assignment of a specific data point to a

cluster remains unchanged, and the result of the k-means algorithm does not depend on an arbitrary random initialisation anymore [BORTZ & SCHUSTER 10: p. 462].

- *Mean-shift:* according to [CHENG 95], for an initial position in feature space, a kernel-based approximation to the gradient of the distribution of data points is computed in an iterative manner and the cluster centre is moved in the direction of this approximated gradient. This procedure is repeated until convergence. Cluster centres thus correspond to local maxima of the distribution of the points in feature space. All data points that can be assigned to the same local maxima belong to the same cluster. For a distribution of data points with several local maxima, the algorithm needs to be started from a (possibly large) number of different initial points in order to determine all local maxima of the distribution, where the number of detected local maxima may also depend on the form of the utilised kernel function [CHENG 95].

Finding an appropriate number of clusters is difficult. The resulting number of clusters is always a compromise between manageability (a small number of clusters) and the demand for homogeneity (a large number of clusters) [SCHÄFER 09].

6.4 Utilised classification methods

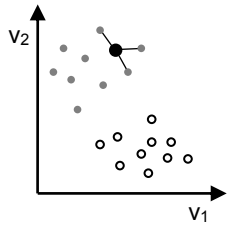
Classification methods are used to assess HUD images, represented by objective feature vectors *v*, with labels *y*. The labels reflect the perceived quality of the HUD images. In general, classification is the assignment of continuous feature vectors to discrete classes. The feature vectors are transferred via a mapping function from the feature space into the class space and are assigned to exactly 1 single class [SCHÜRMANN 96: p. 11f]. In this thesis, the labels are displayed in a 5-dimensional class space, which corresponds to the used rating scale. All individual classes, ω_1 to ω_5, are summarised in the class space Ω. Thus, Ω covers all possibilities that may occur. The mapping between the feature vector *v* and the corresponding class is achieved by the target vector *Y*. Here, all elements are 0, excluding this element that corresponds to the class index of the input pattern [SCHÜRMANN 96: p. 11f]. In this thesis, the nearest neighbour classifier, the polynomial classifier, and the learning vector quantisation are implemented as the mapping function.

6.4.1 Nearest neighbour classifier

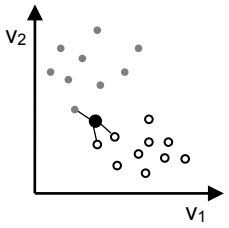
The most straightforward classifier in the machine learning techniques is the nearest neighbour classifier. Examples are classified based on the class of their nearest neighbours [CUNNINGHAM & DELANY 07]. This classifier uses a number of reference patterns to represent each class. For this, the classifier uses all training samples as a reference and assigns the unlabelled feature vector to that class to whose reference vector it is most similar. As similarity measure, the Euclidean distance is used. Each unlabelled feature vector is compared with all reference patterns and is assigned to the class of the reference vector, to which it is closest, respectively, for which the Euclidean distance is minimal [DEMANT et al. 10: p. 179-181], [DIVAKARAN et al. 15]. Since the training samples $S_{Train} = \{v; y\}$ are needed at runtime, the classifier requires no special training. The classification is based directly on the training samples [CUNNINGHAM & DELANY 07]. A great disadvantage of the kNN is that all the work is done at runtime. Therefore, the kNN can have poor runtime performance if a lot of training samples are used [CUNNINGHAM & DELANY 07]. In addition, this method suffers from the curse of dimensionality of the feature vector *v* [MARSLAND 11: p. 183].

It is often useful to consider more than 1 neighbour. Therefore, the technique is more commonly referred to as k-nearest neighbour (kNN) classification. Here *k*-nearest neighbours are used for determining the class, as shown in Figure 46.

Classification examples for k = 3 nearest neighbours as a 2-class problem in a 2-dimensional feature space [CUNNINGHAM & DELANY 07].



The decision is straightforward; all 3 nearest neighbours belong to the same class. So the input sample is also assigned to that class.



The decision for the input sample is a bit more complicated. It has 2 neighbours belonging to the same class and the 3rd neighbour belongs to another class. This can be resolved by simple majority voting.

Own illustration adapted from [CUNNINGHAM & DELANY 07].

Figure 46: basic idea of a 3-nearest neighbour classifier

The kNN classification has 2 stages. The first is the determination of the nearest neighbours and the second is the identification of the class memberships of these nearest neighbours [CUNNINGHAM & DELANY 07]. The choice of *k* is not trivial. Make it too small and the kNN is sensitive to noise. In contrast, a too large *k* can reduce the accuracy because points that are too far away are considered [MARSLAND 11: p. 183].

The allocation rule of the kNN can also be modified and reduced to a 2-class problem. Here only the customer suitability of the HUD images is assessed as acceptable (rating classes 3, 4 and 5) or unacceptable (rating classes 1 and 2). A test image is classified as acceptable only if at least *m* of the *k*-nearest training images represent a customer suitable image quality. On the other hand, the training image is rejected and classified as non-customer suitable [JIANG et al. 06]. By varying the number of relevant nearest training images *m*, the ROC (receiver operator characteristic) curve of the classifier can be determined. The ROC curve is interpolated by plotting the true positive rate (*TPR*) against the false positive rate (*FPR*) at different classifier parameters. The closer this curve is to the upper left corner, the larger the area under the curve (*AUC*), and the better the performance of the classifier [SZELISKI 10: p. 202].

6.4.2 Learning vector quantisation

The learning vector quantisation (LVQ) is discussed in detail in [KOHONEN 01: p. 245-263]. The LVQ is a special case of an artificial neural net and is a prototype based supervised classification algorithm. The LVQ uses prototype vectors to classify unlabelled feature vectors. After training, the prototype vectors approximate the underlying distribution of samples from the training set. According to the competitive winner-takes-it-all strategy, an unlabelled input sample is assigned to the same class to which the nearest prototype vector belongs, as shown in Figure 47. From this class assignment, only the corresponding label has to be determined $Y \rightarrow y$.

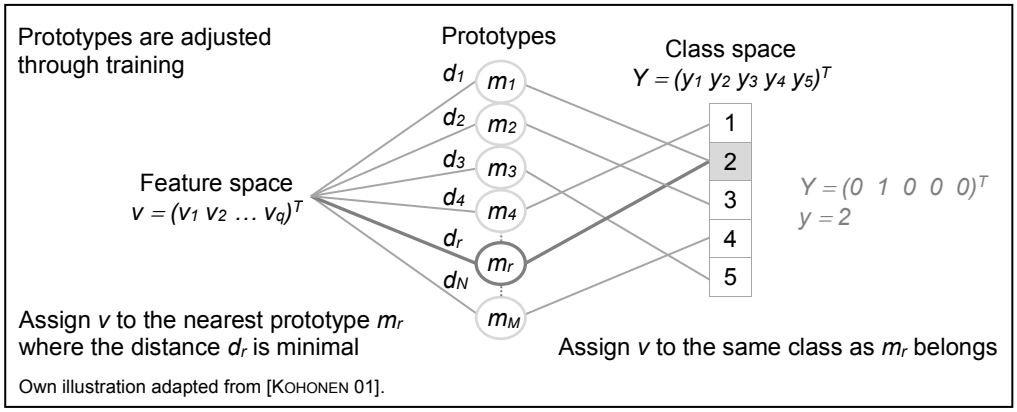


Figure 47: conceptual approach of the LVQ

At the beginning, the prototype vectors are initialised arbitrarily. The number of prototype vectors is set in proportion to the number of training samples in each class. Usually, the number of prototypes is much lower than the number of training samples. Training now means that the prototypes are reinforced. Prototype vectors that approximately minimise the misclassification errors can be found by several learning rules, such as LVQ1, LVQ2.1, and LVQ3. For this, the distance between a single training sample and all prototype vectors is calculated. The prototype vectors that are closest to the training pattern are always reinforced. Positive reinforcement takes place if the class allocation of the nearest prototype vector equals the class allocation of the training sample. Here, the prototype vector is moved closer to the applied training record. In contrast, negative reinforcement takes place by a different class assignment. The prototype vector is pushed away from the applied training sample [KOHONEN 01: p. 245-263]. Below, the used learning rules are presented briefly.

- *LVQ1:* the nearest prototype vector m_r to v is reinforced and moved closer or further away from the feature vector, as shown in Equation 10. All other prototypes remain untouched. The magnitude of the reinforcement is described by the learning rate α [KOHONEN 01: p. 246-249].

$$m_r(t+1) = m_r(t) + \alpha(t) \cdot [v(t) - m_r(t)] \quad \text{If } v \text{ and } m_r \text{ belong to the same class}$$
$$m_r(t+1) = m_r(t) - \alpha(t) \cdot [v(t) - m_r(t)] \quad \text{If } v \text{ and } m_r \text{ belong to different class}$$

m_r: nearest prototype vector, y: feature vector, α: learning rate [KOHONEN 01: p. 247]

Equation 10: LVQ1 learning rule

- *LVQ2.1:* the procedure is similar to the LVQ1. Yet, the 2 nearest prototype vectors m_r and m_i to v are selected. If one prototype belongs to the correct class and the second does not, the 2 prototypes are reinforced at the same time. In addition, the feature vector has to fall inside a 'window' defined by the centre plane of m_r and m_i, as shown in Equation 11 [KOHONEN 01: p. 252f].

$$min\left(\frac{d_i}{d_r}, \frac{d_r}{d_i}\right) > \frac{1-w}{1+w}$$ Precondition, *v* must be inside the 'window'

$$m_r(t+1) = m_r(t) + \alpha(t) \cdot [v(t) - m_r(t)]$$ *v* and m_r belong to the same class

$$m_i(t+1) = m_i(t) - \alpha(t) \cdot [v(t) - m_i(t)]$$ *v* and m_i belong to a different class

m_r, m_i: 2 nearest prototype vectors, *v*: feature vector, α: learning rate, d_i, d_r: Euclidean distances of *v* from m_i and m_r, *w*: width of the window [KOHONEN 01: p. 253]

Equation 11: LVQ2.1 learning rule

- *LVQ3:* the LVQ3 is an enhancement of the LVQ2.1. In addition to the existing conditions, the 2 nearest prototype vectors are also reinforced, if both belong to the same class as the feature vector. Thus, the algorithm appears to be self-stabilising and the optimal placement of m_i does not change in continual learning, as shown in Equation 12. The adjusted learning rate ε depends on the size of the 'window' and is smaller for narrower 'windows' [KOHONEN 01: p. 253f].

In addition to the requirements of LVQ2.1

$$m_r(t+1) = m_r(t) + \varepsilon \cdot \alpha(t) \cdot [v(t) - m_r(t)]$$ *v* and m_r belong to the same class

$$m_i(t+1) = m_i(t) + \varepsilon \cdot \alpha(t) \cdot [v(t) - m_i(t)]$$ *v* and m_i belong to the same class

m_i, m_r: 2 nearest prototypes belonging to the correct class, *v*: feature vector, $\varepsilon \cdot \alpha$: adjusted learning rate [KOHONEN 01: p. 253]

Equation 12: LVQ3 learning rule

The performance of the resulting LVQ algorithm depends mainly on the number of prototype vectors, the initialising of the prototype vectors, the LVQ learning rule, the learning rate, and the termination criterion.

The mapping rule of the LVQ can also be reduced to a 2-class problem where the image quality is classified only as AC (rating classes 3, 4 and 5) or UAC (rating classes 1 and 2). The prototype vectors are still determined by the abovementioned learning rules. Thereafter a test pattern is classified according to the *k*-nearest prototype vectors. A test image is only classified as acceptable if at least *m* of *k*-nearest prototype vectors represent a customer suitable image quality [JIANG et al. 06]. Otherwise, the quality of the test image is not acceptable. Again, the ROC curve of the classifier is determined by varying the number of relevant prototype vectors *m*. For each *m* value, the resulting TPR and FPR values of the test images are calculated and the ROC curve is interpolated.

6.4.3 Polynomial classifier

The polynomial classifier (PC) discussed in [SCHÜRMANN 96: p. 102-186] tries to approximate some decision functions directly. Thereby, the relationship between the subjective labels *y* and the objective feature vectors *v* is adapted to decision equations *d(v)*, which are determined by a full polynomial approach. If all polynomial

terms, up to a certain degree *G*, are used, the polynomial length *M* is given by Equation 13.

$$M = \binom{q+G}{G} = \frac{(q+G)!}{G! \cdot q!}$$

M: length of the polynomial, *G*: degree of the polynomial, *q*: number of objective features to represent the image quality [SCHÜRMANN 96: p. 102]

Equation 13: length of a polynomial with *q* objective features

This general polynomial *d(v)* consists of the constant term a_0 followed by linear terms, quadratic terms, cubic terms, and so on, up to an arbitrary degree *G*, as shown in Equation 14.

$$\begin{aligned} d(v) = {} & a_0 + a_1 \cdot v_1 + a_2 \cdot v_2 + \ldots + a_q \cdot v_q + \\ & a_{q+1} \cdot v_1^{\,2} + a_{q+2} \cdot v_1 \cdot v_2 + a_{q+3} \cdot v_1 \cdot v_3 + \ldots \\ & a \ldots \cdot v_1^{\,3} + a \ldots \cdot v_1^{\,2} \cdot v_2 + a \ldots \cdot v_1^{\,2} \cdot v_3 + \ldots + a_{M-1} \cdot v_q^{\,G} \end{aligned}$$

d(v): decision function for a full polynomial approach, *a*: coefficients, v: feature vector, *q*: number of objective features to represent the image quality, *M*: length of the polynomial, *G*: degree of the polynomial [SCHÜRMANN 96: p. 102]

Equation 14: basic structure of a general polynomial

This general polynomial can be written in a more compact form by introducing a vector-valued mapping $v \rightarrow x$ generating the *M*-dimensional vector *x*:

$$x(v) = (1 \quad v_1 \quad v_2 \ldots \quad v_q \quad v_1^{\,2} \quad v_1 v_2 \quad v_1 v_3 \ldots \quad v_1^{\,3} \quad v_1^{\,2} v_2 \quad v_1^{\,2} v_3 \ldots \quad v_q^{\,G})^T .$$

The function *x(v)* determines the type of the polynomial *d(v)* and is called polynomial structure. Therefore, the general polynomial can be written in a more compact form *d(x)*: $d(x) = a_0 + a_1 \cdot x_1 + a_2 \cdot x_2 + \ldots + a_q \cdot x_q + \ldots a_{q+1} \cdot x_{q+1} + \ldots a_{M-1} \cdot x_{M-1}$.

Following the same principle, the *M*-dimensional coefficient vector *a* is introduced [SCHÜRMANN 96: p. 102-186].

The polynomial classifier requires for each class, ω_1 to ω_5, its own decision function $d\omega_1(v)$ to $d\omega_5(v)$. Subsequently, each decision function is transformed into the vector-valued polynomial function $d_\omega(v) = a_\omega^{\,T} \cdot x(v)$. Combining the class-specific coefficient vectors a_ω into the coefficient matrix $A = (a_{\omega 1} \; a_{\omega 2} \; a_{\omega 3} \; a_{\omega 4} \; a_{\omega 5})$, the compact form of the vector-valued polynomial functions is determined, as shown in Equation 15 [SCHÜRMANN 96: p. 102-186].

$$d_{\omega 1 \ldots \omega 5}(v) = A^T \cdot x(v)$$

$d_{\omega 1 \ldots \omega 5}(v)$: decision functions, *A*: coefficient matrix, *x*: polynomial structure vector [SCHÜRMANN 96: p. 103]

Equation 15: compact form of the vector-valued polynomial functions

In the training phase, the coefficient matrix *A* is determined. The polynomial structure vector *x(v)* is predetermined and remains unchanged. Based on given training samples, the deviations between the calculated labels and the given labels that correspond to the subjective perception are minimised. Thus, the least-mean square approach can be determined, as shown in Equation 16.

$$S^2 = E\left\{ \left| A^T \cdot x(v) - Y \right|^2 \right\} = \min_A$$

S^2: residual variance, A: coefficient matrix, *x*: polynomial structure vector, *Y*: target vectors of the training samples corresponding to the subjective perception [SCHÜRMANN 96: p. 107]

Equation 16: least-mean square approach of the polynomial regression

Here, the labels of the training samples are presented as 5-dimensional target vectors *Y*. Through several mathematical transformations, it is possible to solve this minimisation problem and the coefficient matrix *A* can be determined, according to Equation 17.

$$A = E\left\{x \cdot x^T\right\}^{-1} \cdot E\left\{x \cdot Y^T\right\}$$

A: coefficient matrix, *x*: polynomial structure vector, *Y*: 5-dimensional target vectors of the training samples corresponding to the subjective perception, E: moment matrixes [SCHÜRMANN 96: p. 109]

Equation 17: determination of the coefficient matrix

Thereby, $E\{x \cdot x^T\}$ and $E\{x \cdot y^T\}$ represent moment matrices. The polynomial structure vector can only be determined, if $E\{x \cdot x^T\}$ is invertible [SCHÜRMANN 96: p. 102-186].

The training phase is completed by the determination of the coefficient matrix *A*. Afterwards the polynomial classifier is able to calculate a label for any feature vector, as shown in Equation 18.

$$v_{test} \rightarrow x_{test}, \quad Y_{test} = A^T \cdot x_{test}$$

A: coefficient matrix, *x*: polynomial structure vector, *y*: target vectors of the test samples determined by the polynomial classifier [SCHÜRMANN 96: p. 103]

Equation 18: application of the polynomial classifier on a test dataset

For this purpose, the feature vector v_{test} of the test sample is transformed into the polynomial structure vector $x_{test}(v)$. The target vector Y_{test} is then obtained from the multiplication of the coefficient matrix *A* determined by training and the structure vector $x_{test}(v)$ [SCHÜRMANN 96: p. 102-186].

The peculiarity of this classifier is that no single value is determined. The outcome of the polynomial classifier is a 5-dimensional assignment probability vector *d'*, shown in Figure 48.

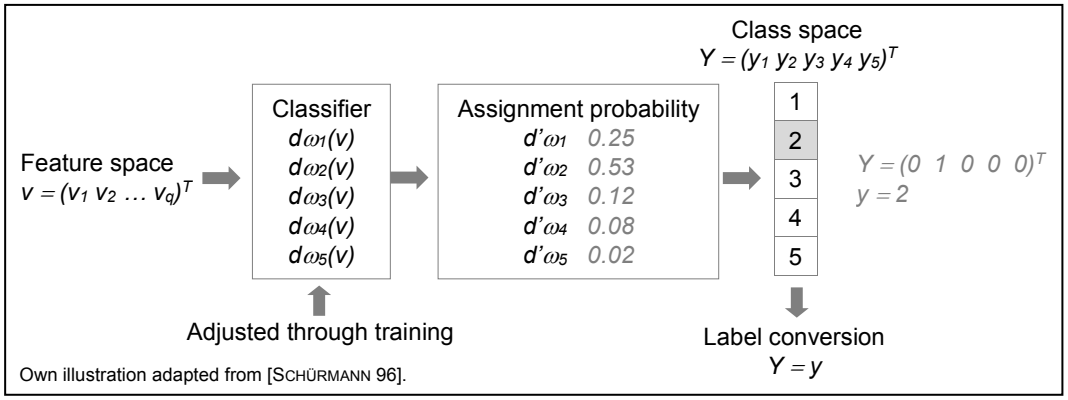


Figure 48: conceptual approach of a polynomial classification

The assignment probability vector *d'* contains for each of the 5 classes the probability that the feature vector belongs to that class. The sum of all probabilities is always 1. Then, the feature vector is assigned to the class with the greatest probability $d' \rightarrow Y$. The result is the 5-dimensional target vector, where all elements are 0, excluding the element that corresponds to the class with the largest assignment probability [SCHÜRMANN 96: p. 102-186]. Finally, the target vector needs to be converted into a continuous label $Y \rightarrow y$.

With increasing polynomial degree *G*, the complexity of the classification system increases as well. Here, according to [MARSLAND 11], it is important to ensure that the number of samples *N* in the training set is 10 times greater than the number of polynomial terms *P*, $P \geq 10 \cdot N$.

Again, the assignment task is reduced to a 2-class problem where only the customer suitability of the HUD images as acceptable or unacceptable is assessed. So far, the feature vector is assigned to the class with the greatest probability. This rule can be modified that the decision for an acceptable quality is only made if the probability of the feature vector to belonging to an acceptable quality is higher by a multiple ϕ than the probability of belonging to an unacceptable quality [KRISTIAN et al. 11: p. 203], as shown in Figure 49.

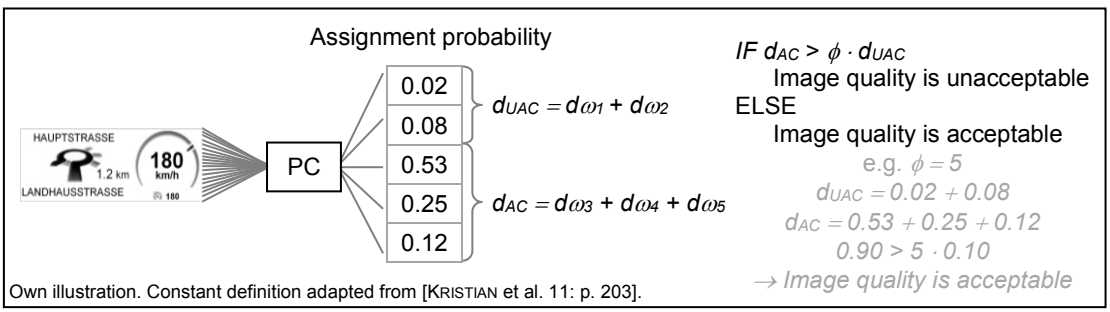


Figure 49: 2-class polynomial classification for assessing the image quality

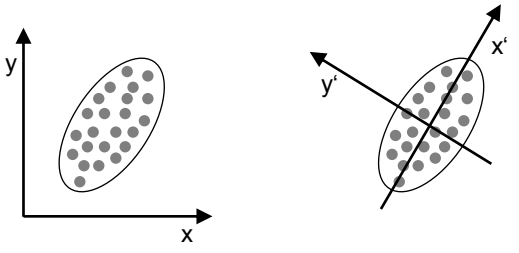
By varying the constant ϕ, the ROC curve of the classifier can be determined. The ROC curve is interpolated by plotting the true positive rate (*TPR*) against the false positive rate (*FPR*) at various constant values ϕ [KRISTIAN et al. 11: p. 219].

6.5 Dimensionality reduction

Dimensionality reduction is used to lower the computational cost of many algorithms. However, it can also remove noise, significantly improve the results of the learning algorithms, make the dataset easier to work with and make the results easier to understand [MARSLAND 11: p. 221]. In this thesis, 2 methods of dimensionality reduction are used, the principal component analysis and the feature selection approach.

6.5.1 Principal component analysis

The main objective of the principal component analysis (PCA) is to reduce the number of characteristic features. The PCA projects high dimensional data into a low dimensional space with minimal information loss, as shown in Figure 50.



The diagrams show 2 different sets of coordinate axes. The second consists of a rotation and translation of the first and was found using PCA.

Illustration taken from [MARSLAND 11: p. 227].

The potential for dimensionality reduction lies in the fact that the y' dimension does not demonstrate much variability. Thus, it might be possible to ignore it and use the x' axis values alone without compromising the results of the learning algorithm [MARSLAND 11: p. 227].

Figure 50: principle of the PCA

This is based on the assumption that a complete set of measured variables hides only a small number of fundamental components that are not correlated [BORTZ & SCHUSTER 10: p. 385ff]. The performance of the principal component analysis is described in [BACKHAUS et al. 00: p. 252-327]. The first step is the calculation of the correlation matrix *R* of the feature vectors. Based on the correlation matrix *R*, it can be seen whether relations between the objective feature exist or not [BACKHAUS et al. 00: p. 262], as shown in Equation 19.

$$R = \begin{pmatrix} 1 & \Lambda & r_{1q} \\ \mathrm{M} & 1 & \mathrm{M} \\ r_{q1} & \Lambda & 1 \end{pmatrix}, \quad r_{1q} = \frac{\sum_{i=1}^{N} \left(v_{1i} - \overline{V_1}\right) \cdot \left(v_{qi} - \overline{V_q}\right)}{\sqrt{\sum_{i=1}^{N} \left(v_{1i} - \overline{V_1}\right)^2 \cdot \sum_{i=1}^{N} \left(v_{qi} - \overline{V_q}\right)^2}}$$ correlation between v_1 and v_q

R: correlation matrix, v: feature vector, q: number of objective features to represent the image quality, N: number of images taken for investigation, r_{1q}: correlation coefficient between feature v_1 and v_q, $\overline{V}$: average value of the features for all images [BACKHAUS et al. 00: p. 263]

Equation 19: correlation coefficient and correlation matrix

The correlation coefficient r can take values between ± 1. A value of ± 1 indicates a completely positive or negative linear correlation. In contrast, 0 shows that the features are independent of each other [NEUE STATISTIK 03].

The principal component analysis is only useful if the features show substantial correlations. This is the case if the correlation matrix differs significantly from the unit matrix [BORTZ & SCHUSTER 10: p. 417]. The higher the variables correlate, the fewer components are needed to explain the total variance. The total variance is always equal to the number of used features [BORTZ & SCHUSTER 10: p. 392f].

In the next step, the eigenvalues of the correlation matrix are calculated. They indicate how much of the total variance is covered by the corresponding principal component. The sum of the eigenvalues corresponds to the total variance [BORTZ & SCHUSTER 10: p. 408f]. Based on the size of the eigenvalues, it is often decided how many components should be extracted. Thus, the eigenvalues indicate which number of components explains the total variance sufficiently well [BORTZ & SCHUSTER 10: p. 415f].
After determining of the number of principal components, the component loadings and component values are calculated. The component values describe the projection of each data point to the principal component. The component loadings reflect the correlation between the principal component values and the initial values of the features, as shown in Equation 20 [BORTZ & SCHUSTER 10: p. 392f].

$$A = \begin{pmatrix} a_{11} & \Lambda & a_{1f'} \\ \mathrm{M} & \mathrm{O} & \mathrm{M} \\ a_{q1} & \Lambda & a_{qf'} \end{pmatrix}, \; A = V \cdot diag\left\{\sqrt{\lambda}\right\}, \quad F = \begin{pmatrix} f_{11} & \Lambda & f_{1f'} \\ \mathrm{M} & \mathrm{O} & \mathrm{M} \\ f_{N1} & \Lambda & f_{Nf'} \end{pmatrix}, \; F = X \cdot A \cdot \left(A^T \cdot A\right)^{-1}$$

A: component loading matrix, F: component value matrix, V: transformation matrix that contains the eigenvectors of the correlation, X: raw data matrix that contains all feature vectors, λ: eigenvalues of the correlation matrix, N: number of images taken for investigation, f': number of significant components ($f' \leq N$) [BORTZ & SCHUSTER 10: p. 412].

Equation 20: component ladings and component values

The PCA results in principal components, which are independent of each other and explain successively maximum variance [BORTZ & SCHUSTER 10: p. 392].

Finally, the Varimax-rotation creates a simple structure of the principal components. This method is an orthogonal rotation technique that preserves the independence of the components. The aim is to maximise the variance of the squared component loadings per component. For a detailed description of the Varimax-rotation please refer to [BORTZ & SCHUSTER 10: p. 419f].

6.5.2 Feature selection

Feature selection refers to looking through the available features and determining whether they are actually useful [MARSLAND 11: p. 221]. Irrelevant and redundant features are removed. Only the features that will give good system accuracy are used.
Feature selection differs from dimensionality reduction with PCA. Both methods attempt to reduce the number of attributes in the dataset. The PCA creates new components, while feature selection methods include and exclude features in the data without changing them [BROWNLEE 14].

Chapter

7 Development of an assessment algorithm

This chapter examines how an assessment algorithm for HUD images can be implemented. The main task of an assessment algorithm is to display the relationship between any forms of aberrations and their resulting quality impressions. It is shown how such an assessment algorithm can be derived from the results of conducted subject studies. The subjective perception of distortions and double images is investigated. The structure of the implemented algorithm is shown in Figure 51.

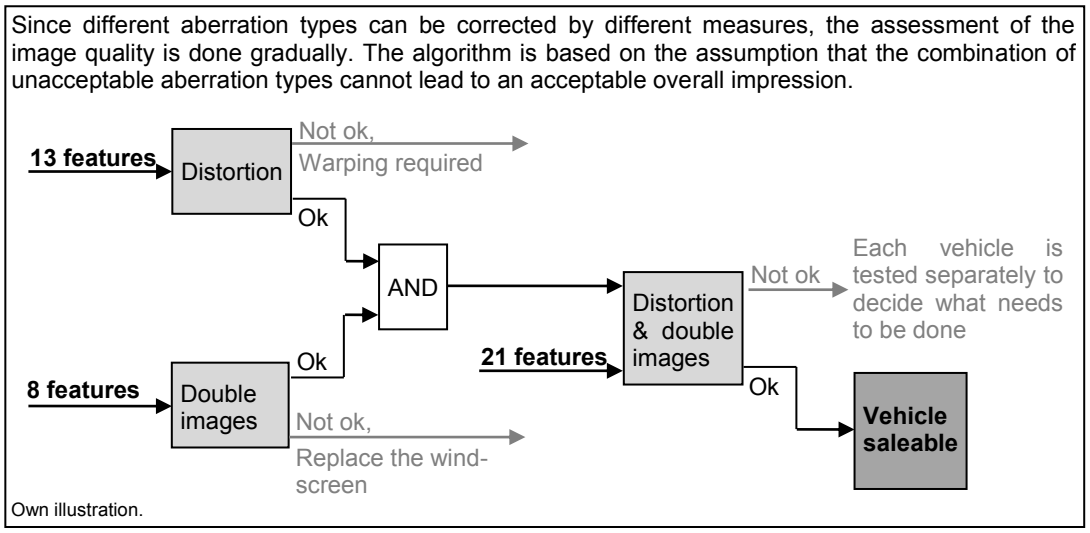


Figure 51: flowchart to assess the image quality

The flowchart shows that the assessment of the perceived quality is done gradually. Distortions, double images and the combination of distortions and double images are taken into account. First, the perception of distortions and double images is analysed separately. The reason for this is that the different aberration types can be corrected by different measures. Resulting distortions that are not suitable for customers are corrected by image warping. Warping assumes that a suitable pre-distorted image that has passed through the optical system is converted into a straight undistorted image, see chapter *3.8.1*. In contrast, double images are visible if the wedge-shape PVB layer is not properly fitted between the 2 glass plates of the windscreen, see chapter *3.8.2*. Since there is no way to correct the wedge shape layer, the windscreen has to be replaced to avoid double images.

Only if the separate analysis of distortions and double images is successful, the combination of the 2 aberrations is investigated. The algorithm is based on the assumption that the combination of unacceptable aberration types cannot lead to an acceptable overall impression. If the combination of distortions and double images is acceptable, the vehicle can be delivered to the customer. Otherwise, each vehicle is checked separately to decide what needs to be done.
Input parameters for the algorithm are objective features that numerically describe the image quality. 21 features are used (13 features for detecting distortions and 8 features for capturing double images), see chapter *5.1*.

This chapter is divided into several parts. First, the images of the database are analysed in more detail and divided into appropriate training and test datasets. Clustering methods are mainly used to find representative images. Afterwards, it is investigated which method is best suited to assess the quality of HUD images. For this purpose, the standard limit value consideration is implemented first. Then it is checked whether classification methods can achieve better recognition accuracies. Finally, it is shown that the manual labelling effort can be reduced by a semi-supervised or active learning scenario.

7.1 Analysis of the images from the database

For the development of the assessment algorithm, 71895 distorted images and 64410 images with different double images are available, see chapter *4.4*. The objective features are calculated for each image to numerically describe the quality, see chapter *5.1*. This chapter analyses the objective feature values of the images. First, it is checked in which size ranges the objective features occur. Subsequently, a principal component analysis is performed to see if dimensionality reduction could be quite useful for the implementation of the classification methods. The PCA is performed with the free analysis software GNU PSPP[7].

7.1.1 Investigation of the images of the distortion dataset

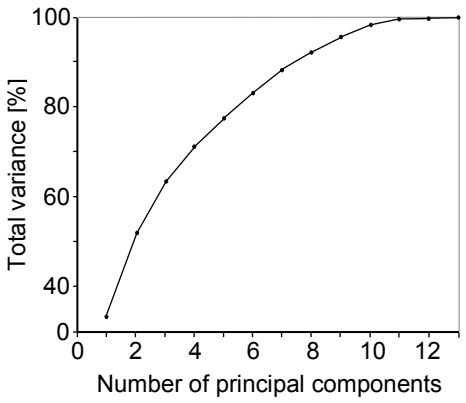
To get an overview of occurring distortion type quantities, frequency distribution diagrams are created for each feature, see appendix *A.2*. Figure 155 shows that the values of the feature no. [3.8] - *Smile vertical* and feature no. [1.1] - *Adjusted in width* are completely within the perception limits. All other features show clearly visible occurrences exceeding the limits. The perception limits of the features are calculated in chapter 5.2.1. The frequency diagrams show that assembly tolerances cause visible distortions.

[7] GNU PSPP is a program for the statistical analysis of sampled data. It is a free replacement for the proprietary program SPSS and is very similar to it. The program can be downloaded from the following link: https://www.gnu.org/software/pspp/.

In the next step, the feature values are investigated by the PCA. The result of the PCA is shown in Figure 52. The graph on the left shows how much information is lost by reducing the features to a smaller number of principal components. The x-axis represents the number of principal components and the y-axis the total variance covered. The table on the right shows this information numerically. The first column shows the resulting eigenvalues from the correlation matrix *R* of the feature vectors. The sum of the eigenvalues corresponds to the total variance, which is always equal to the number of features used. The second column contains the percentage of the total variance covered by the principal components and the last column the cumulative percentage.

The PCA is performed with 71895 images. The quality of each image is described by 13 features.

Total variance covered by the corresponding principal components



Principal component	Eigen-value	Total variance covered	
		Percentage of variance	Cumulative percentage
1	4.33	33.27%	33.27%
2	2.42	18.58%	51.86%
3	1.49	11.50%	63.35%
4	1.01	7.74%	71.09%
5	0.81	6.25%	77.34%
6	0.74	5.71%	83.05%
7	0.66	5.06%	88.11%
8	0.50	3.82%	91.93%
9	0.46	3.54%	95.47%
10	0.34	2.61%	98.08%
11	0.18	1.37%	99.45%
12	0.07	0.55%	99.99%
13	0.00	0.01%	100.00%

Own illustration. Results obtained by the program GNU PSPP.

Figure 52: results of the PCA for the distortion dataset

Here, a reduction to 10 principal components does not lead to a considerable loss of information, as more than 98% of the total variance is still covered. With a dimensionality reduction to 2 principal components, more than 50% of the total variance is covered. The PCA analysis shows that the features are well chosen and hardly correlated. For the implementation of an assessment algorithm, it is necessary to check whether a dimensionality reduction of the feature space could be useful.

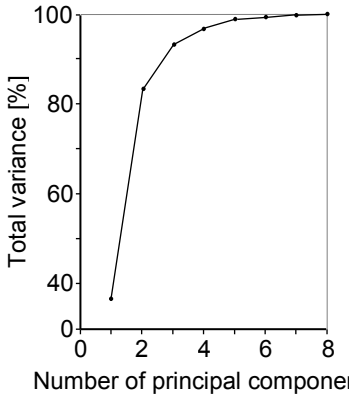
7.1.2 Investigation of the images of the double image dataset

Also for double images, the frequency distribution diagrams are created for each feature see, appendix *A.3*. Figure 156 shows that the values of the feature no. [*2.2*] - *Mean vertical* are completely within the perception limits. The values of all other features are well above the limits and are therefore visible. The perception limits of the features are calculated in chapter *5.2.2*.

Now, the feature values are analysed with the PCA. The results are summarised in Figure 53. The figure shows, as for the analysis of distortions, on the left side the graph and on the right side the corresponding values.

The PCA is performed with 64410 images. The quality of each image is described by 8 features.

Total variance covered by the corresponding principal components



Principal component	Eigen-value	Total variance covered	
		Percentage of variance	Cumulative percentage
1	3.66	45.74	45.74
2	3.13	39.16	84.90
3	0.72	9.05	93.95
4	0.24	3.05	97.01
5	0.16	1.97	98.98
6	0.04	0.55	99.53
7	0.02	0.31	99.84
8	0.01	0.16	100.00

Own illustration. Results obtained by the program GNU PSPP.

Figure 53: results of the PCA for the double image dataset

The analysis shows that the features are clearly correlated with each other. The more the features correlate with each other, the fewer principal components are needed to cover the total variance. Here, only 3 principal components are needed to cover more than 93% of the total variance. Even with a reduction to 2 principal components, almost 85% of the overall variance is covered. For the implementation of the assessment algorithm, a dimensionality reduction of the feature space can be quite useful.

7.2 Unsupervised learning: selection of representative images

To investigate the customer acceptance, the images of the database are assessed subjectively. The images of the database have already been prepared for the subjective assessment, see chapter *4.4.* Unfortunately, the rating of nearly 72000 distorted images and nearly 65000 images showing double images is very time consuming and is not workable. Thus, the rating is limited to a few representative images obtained by clustering methods. Clustering groups similar images together. The aim is that 1 group contains only images with the same subjective assessment. In addition, the rating of 1 single image should match the ratings of all other images in the same group. Consequently, the expenditure of the customer questionings can be reduced since only 1 image from each group needs to be evaluated.

7.2.1 Clustering the images in the distortion dataset

This chapter is divided into 4 parts. First, the cluster condition is determined numerically. Then, the images of the database are divided into several clusters. In the next step, the subjective perception of the images in the database is examined. Finally, the images of the database are split into training and test datasets.

Determine the cluster condition numerically:

The aim is to find the minimum number of clusters that exhibits no subjective difference between the images on a group. The requirement of subjective equality in a cluster is investigated and numerically quantified in a preliminary investigation. For this, the 71895 images are analysed in different cluster solutions. Due to a large amount of data and the temporal expenditure, the investigation is executed with only 6 test persons. For clustering, the simple k-means algorithm is used, where the number of clusters is known in advance. The images are then displayed successively for each cluster on an external monitor, see chapter *4.5*. As soon as a difference between the images in 1 cluster can be perceived, the questioning is aborted and restarted with modified parameters. The resulting clusters depend on the selected features and the chosen cluster number. Due to the different subjective relevance, see Table 6, not all evaluation features are needed to describe the subjective difference. By choosing subjectively relevant features, a reduction of the dimension is obtained because subjectively irrelevant features are ignored. Until a suitable arrangement of the images is found, either a further feature is taken for clustering or the cluster number is increased. It must be ensured that the start allocation of the k-means algorithm remains the same. The cluster condition is reached if the participants confirm that there is no subjective difference between the images in 1 cluster. Then the required cluster condition can be obtained based on the involved features [KÖPPL et al. 16].

The preliminary investigation shows that only 5 evaluation features (no. [3.1] - *Rotation*, no. [3.2] - *Misalignment horizontal*, no. [3.3] - *Misalignment vertical*, no. [3.6] - *Smile horizontal top*, no. [3.7] - *Smile horizontal bottom*) are needed to describe the subjective perceptible difference. The images are properly sorted if the maximum difference of the feature values Δ_{rate} inside the clusters is less than 1 HUD pixel (0.58 mm x 0.58 mm), as shown in Equation 21. The distribution of the values of the remaining 8 features does not affect the subjective perception. It is shown that images with similar objective feature values (no. [3.1], no. [3.2], no. [3.3], no. [3.6], no. [3.7]) cause the same quality impression [KÖPPL et al. 16].

$$\Delta_{\text{rate}} < 1\,\text{HUD pixel} \begin{cases} \text{feature no. [3.1], [3.2], [3.3]} \\ \text{feature no. [3.6], [3.7]} \end{cases}$$

The images are properly sorted, if the maximum difference of the feature values Δ_{rate} inside the clusters is smaller than 1 HUD pixel [KÖPPL et al. 16].

Equation 21: subjective equality for clustering the images in the distortion dataset

Clustering results:

Once the objective of clustering is known, the 71895 images are divided into subjective groups. All 13 features are used for this. The minimum number of clusters is determined where no subjective difference between the images can be perceived. The images in a cluster are all annotated with the same subjective label. Likewise, the values

of the objective features no. [3.1], no. [3.2], no. [3.3], no. [3.6], no. [3.7] vary by less than 1 HUD pixel, as shown in Equation 21.

To obtain the best solution, the results of the ward, the k-means, a combination of ward and k-means and mean-shift clustering are compared. The used clustering methods are implemented for this thesis in the free software environment *R*[8]. For ward and k-means, clustering functions of the package *cluster* (function *kmeans* and *hclust*) and for the mean-shift clustering, a function of the package *LPCM* (function *ms*) is applied. After each cluster passage, it is checked whether a subjective difference between the images in 1 cluster exists. Since for of the k-means algorithm the cluster solution depends on the randomly selected starting condition, the results of 100 different initial conditions are considered in more detail. The resulting numbers of clusters are shown in Figure 54. Here, the number of clusters without subjective difference is plotted against the number of different starting positions.

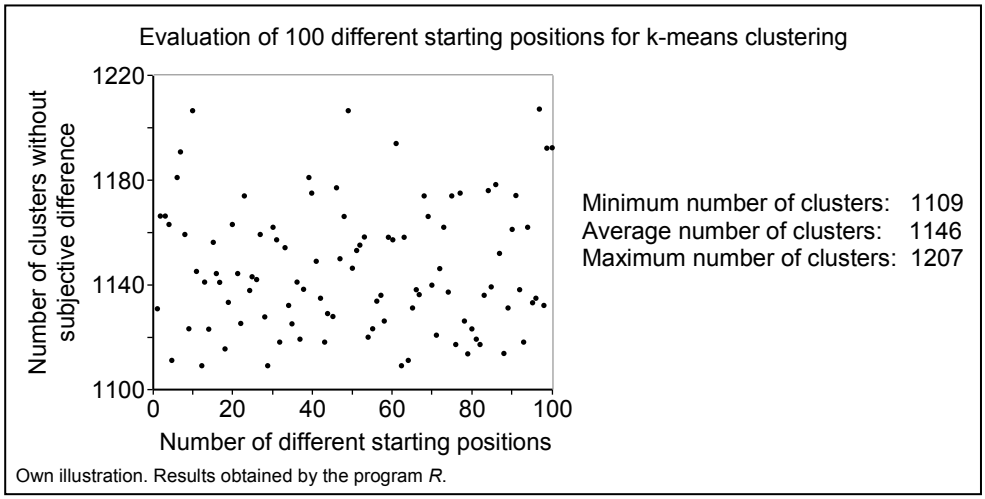


Figure 54: k-means: resulting numbers of clusters for different starting positions

The evaluation of the resulting numbers of clusters shows that the difference between the minimum (1109) and the maximum number of clusters (1207 clusters) amounts 98 clusters. The average number of clusters is 1146 clusters. This clearly shows that the resulting cluster number depends strongly on the chosen starting position.

The results of the clustering methods are summarised together with the average cluster number of the k-means procedure in Table 13. Comparing the solutions shows that the combination of ward and k-means results in 1007 groups. This is the smallest possible number of clusters and is therefore considered the best solution. In contrast, the largest number of clusters is obtained from the mean-shift algorithm. The results of the ward method and the k-means algorithm are between these values [KÖPPL et al. 16].

[8] *R* is a free software environment for statistical calculations and graphics. It compiles and runs on a variety of UNIX platforms, Windows and MacOS. The program can be downloaded from the following link: http://www.r-project.org/.

Clustering method	Number of clusters
ward	1239
k-means (average of 100 starting positions)	1146
ward and k-means	1007
mean-shift	4288

Table 13: clustering results for the distortion dataset

By combining ward and k-means clustering, the 71895 images are now distributed to 1007 clusters. Subsequently, the found cluster solution is confirmed by 6 test persons. For this, the images of each cluster are summarised in a film. The images are visible for 100 ms. Each distorted image is followed by a black image that is displayed for 10 ms. This results in 1007 individual films with a total length of 10 hours and 11 minutes. The investigation is carried out in the special test environment. After seeing the film, the test persons confirm that there is no subjective difference between the images in each cluster. Thus, the images of the dataset can be reduced to only 1007 representative images.

Looking more closely at the individual clusters, a frequency diagram is generated with the number of images in each cluster, as shown in Figure 55. 83 clusters contain only a single image. On the other side, there are 8 clusters with more than 500 images. The maximum number of images in a cluster is 898 images.

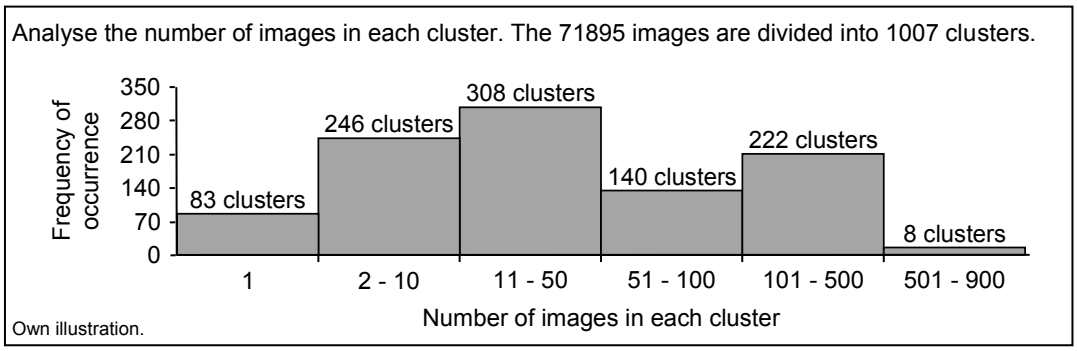


Figure 55: evaluation of the found cluster solution for the distortion dataset

Subjective evaluation of the images in the database:

The previously found cluster solution now serves as the basis for a further customer survey. In which only 1007 representative images are subjectively assessed. The images that are next to the theoretical cluster centres are shown to 12 test persons in the special test environment. The time required is about 8 hours per person. Since all images in each cluster cause the same quality perception, the ratings of 1007 cluster images allow conclusions to be drawn about the 71895 images. Thus, a rough overview of the subjective ratings of the images in the database is obtained, as shown in Figure 56.

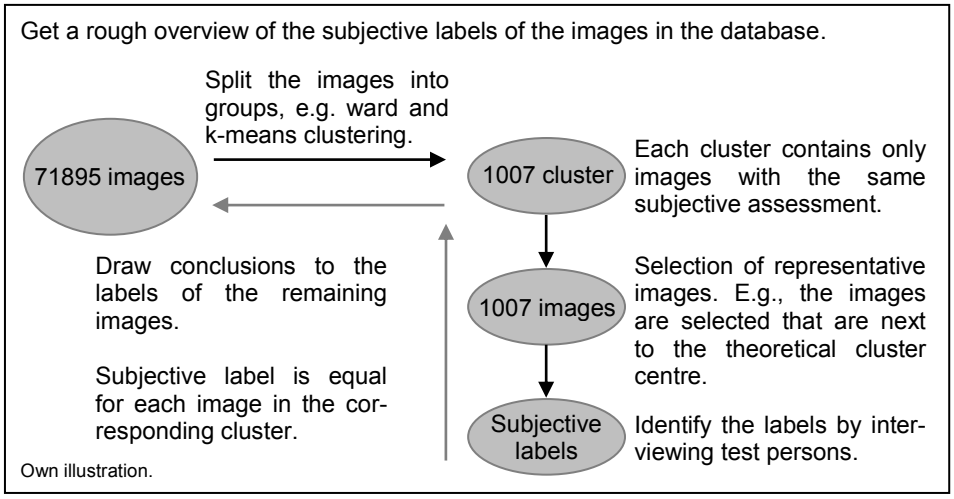


Figure 56: label the images of the distortion dataset roughly

The images are evaluated on a scale from 1 (very annoying) to 5 (imperceptible), according to the ITU-R 500 directive; see Table 1. During the questioning, it is remarkable, that after labelling 300-360 images, the assessment of the subsequent images is stressful and the test persons need many breaks. The results of the survey are summarised in Table 14 [KÖPPL et al. 16].

Subjective rating class	1	2	3	4	5
Number of clusters	407	551	35	9	5
Number of images in each class	20333	45983	3054	1869	656
Number of images AC / UAC	66020		5875		

Table 14: rough labelling of the images in the distortion dataset

The evaluation shows that only 5 clusters with the highest score exist. On the contrary, there are 407 clusters with the lowest score. It is remarkable that even more images are classified as annoying or very annoying, as imperceptible or not annoying. By the way of example, an image for each evaluation class can be found in the appendix *A.4*.

Dividing the images into train and test data:
In order to train and test the assessment algorithm, the images of the database are split into 2 independent datasets. First, the test dataset is extracted. Since the images in the test dataset need to be subjectively labelled, the maximal number of images is set to 360 images. Since the images of the database are roughly labelled, test images can be selected that are evenly distributed among all 5 classes. Thus, 72 images are selected for each class. All images except the images closest to the cluster centres are allowed. To determine the exact subjective labels, the 360 selected test images are evaluated by the test persons. The procedure of the test image extraction is shown in Figure 57.

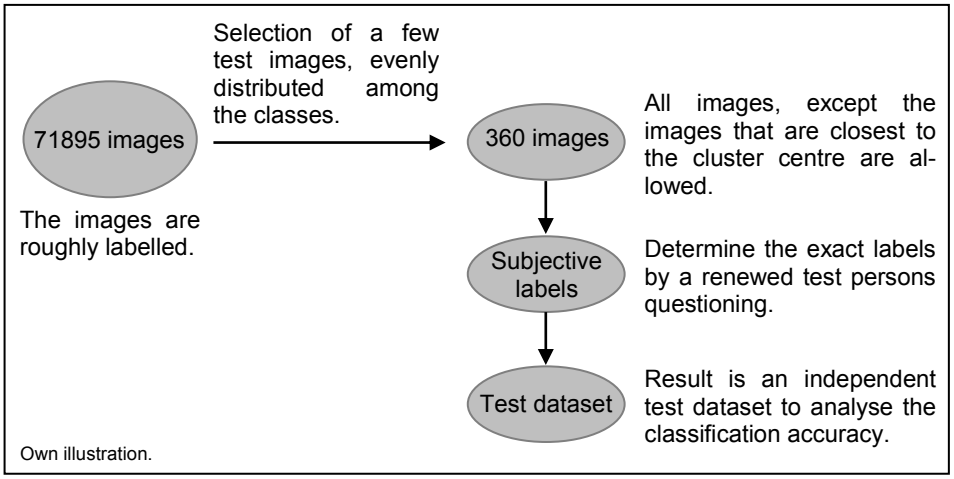


Figure 57: separation of the test dataset from the distortion dataset

The test images are evaluated by 12 test persons in the special test environment. The time required per test person is about 3 hours. The images are also labelled on the rating scale from 1 (very annoying) to 5 (imperceptible) rating points. The results of the survey are summarised in Table 15. The evaluation of the survey shows that the test dataset consists of 145 images of unacceptable quality (classes 1 and 2) and 215 images of acceptable quality (classes 3, 4 and 5).

Subjective rating class	1	2	3	4	5
Number of images in each class	89	56	77	87	51
Number of images AC / UAC	145		215		

Table 15: labelling of the images of the test dataset for distortion

In the next step, the training dataset is determined. The procedure for this is shown in Figure 58.

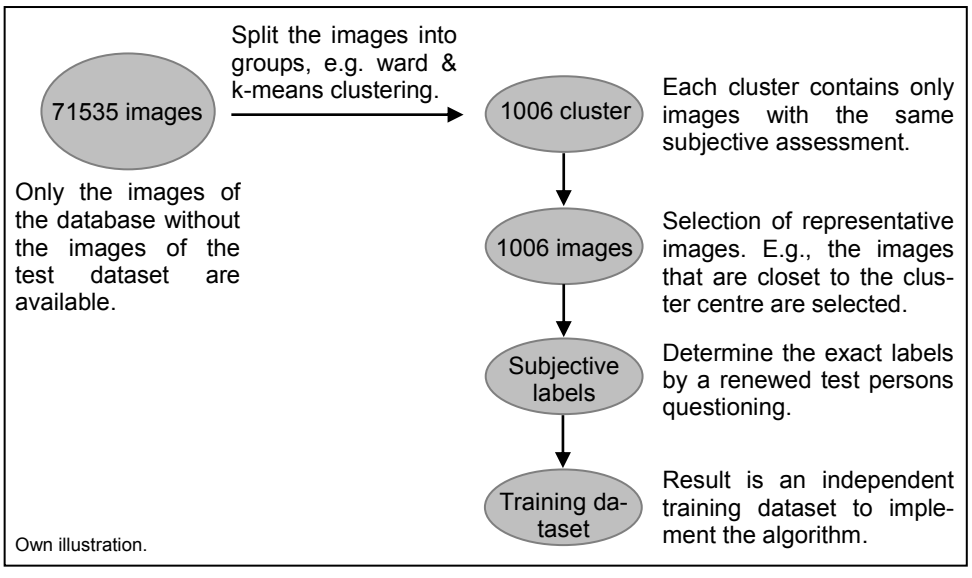


Figure 58: separation of the training dataset from the distortion dataset

71535 images are available after deduction of the test data. In order to train the evaluation algorithm as well as possible, the training data must be sufficiently representative. The representative training images are obtained by clustering. The cluster condition has already been determined numerically at the beginning of this chapter. Again, each cluster should contain only images that cause the same subjective quality impression. The combination of ward and k-means is applied as the clustering method. The cluster analysis shows that the remaining images can be divided into 1006 clusters. The images closest to the cluster centre are used as training data. Finally, the subjective labels are determined. Before the images of the training dataset are shown to the test persons, the 1006 images are compared to the 1007 representative images of the database. The comparison shows that 991 images are identical and already labelled by the test persons. Consequently, only 15 images are rated during the test persons questioning. The evaluation of the images takes place in the designated test environment. The labelling of the images takes about 8 minutes per person. The resulting labels of the training dataset are shown in Table 16.

Subjective rating class	1	2	3	4	5
Number of images in each class	408	550	33	10	5
Number of images AC / UAC	958		48		

Table 16: labelling of the images of the training dataset for distortion

The training dataset consists of 958 images of unacceptable quality (classes 1 and 2) and 48 images of acceptable quality (classes 3, 4 and 5). Here, 550 images are assigned to rating class 1 and only 5 images to rating class 5.

7.2.2 Clustering the images in the double image dataset

The clustering of the images with different double image distances is based on the same principle as the clustering of the distorted images in the previous chapter.

Determine the cluster condition numerically:

The aim is that the images in a cluster exhibit no subjective difference. In a preliminary investigation, the requirement of subjective equality is determined numerically. The 64410 images are analysed in different cluster solutions. Again, only 6 test persons participate in the survey. As cluster method, the k-means algorithm is used. The start allocation remains the same through the investigation. The survey takes place in the special test environment. As soon as the test persons perceive a difference between the images in 1 cluster, the questioning is restarted with modified parameters. The resulting clusters depend on the selected features and the number of clusters. Until the cluster condition can be determined, either another feature for clustering is selected or the number of clusters is increased. If all test persons confirm that, the images are arranged in the corresponding cluster, the cluster condition can be determined based on the involved features.

After executing the preliminary investigation, it can be seen that all 8 characteristic features are needed for clustering. The subjective perceptible difference can only be described if all features are used. The investigation shows that images with similar objective feature values cause the same quality impression. An appropriate cluster solution is found if the maximum difference of all feature values Δ_{rate} inside 1 cluster is less than 1 HUD pixel (0.58 mm x 0.58 mm), as shown in Equation 22.

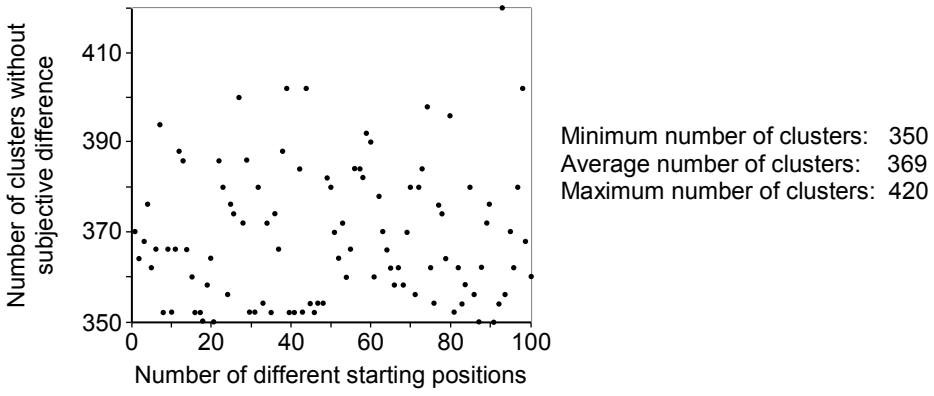
$$\Delta_{\text{rate}} < 1\,\text{HUD pixel} \begin{cases} \text{Horizontal: feature no. [1.1] ... [1.4]} \\ \text{Vertical: feature no. [2.1] ... [2.4]} \end{cases}$$

The images are properly sorted, if the maximum difference of all feature values Δ_{rate} inside the clusters is smaller than 1 HUD pixel.

Equation 22: subjective equality for clustering the images of the double image dataset

Clustering results:

According to the cluster condition, the images are subdivided into subjective clusters. Here, the results of the ward, the k-means, a combination of ward and k-means and the mean-shift clustering are compared. The minimum number of clusters is determined when the objective feature values vary by less than 1 HUD pixel. Since the randomly selected starting condition of the k-means algorithm affects the resulting cluster solution, the results of 100 different initialisations are analysed, as shown in Figure 59.



Own illustration. Results obtained by the program *R*.

Figure 59: k-means: resulting numbers of clusters for different starting positions

The graph shows that the average number of clusters is 369 clusters. The minimum cluster number is 350 and the maximum cluster number is 420 clusters. This results in a difference between the minimum and the maximum number of 70 clusters. Table 17 summarises the results of the remaining clustering algorithms together with the average cluster number of the k-means method. The mean-shift algorithm has the largest number of clusters with 1797 clusters. In contrast, combining ward and k-means yields with 422 clusters the smallest number of clusters. The results of the k-means method and the ward algorithm are in between.

Clustering method	Number of clusters
ward	425
k-means (average of 100 starting positions)	369
ward and k-means	422
mean-shift	1797

Table 17: clustering results for double image dataset

Examining the results shows that the k-means algorithm yields 350 clusters with a specific initialisation. This number is smaller than the number of clusters resulting from the combination of the ward and k-means. Thus, this solution is considered as the best possible solution.

The 64410 images are split into 350 clusters using the k-means algorithm. To confirm that there is no subjective difference between the images in each cluster, the cluster solution is assessed by 6 test persons. For this, the test persons watch a film that shows the summarised images of each cluster. The images are visible for 100 ms. Followed by a black image that is visible for 10 ms. The result is 350 individual films with a total length of 8 hours and 55 minutes. Again, the investigation is carried out in the special test environment. The result is that there is no subjective difference between the images in each cluster. Consequently, the images of the database can be reduced to 350 representative images.

The number of images in each cluster is summarised in the following frequency diagram, as shown in Figure 60. 1 cluster contains the maximum number of 4609 images. Similarly, 21 clusters contain more than 900 images. On the other hand, 83 clusters consist of only 1 single image.

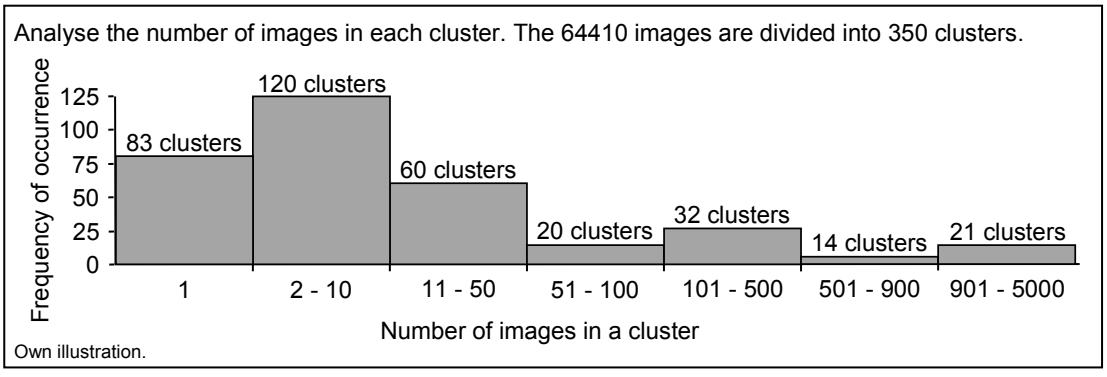


Figure 60: evaluation of the found cluster solution for the double image dataset

Subjective evaluation of the images in the database:
The 350 representative images of the cluster solution serve as the basis for a further customer survey. The images that are next to the theoretical cluster centre are labelled manually in the special test environment. In total, 12 test persons are asked. The time required is about 3 hours per person. The images are rated on a scale from 1 (very annoying) to 5 (imperceptible) according to the ITU-R 500 guideline. Since the images in 1 cluster cause all the same quality perception, the ratings of the 350 representative images allow conclusions to be drawn from the remaining 64410 images. Thus, a rough overview of the subjective labels of all images can be obtained. The resulting labels are summarised in Table 18.

Subjective rating class	1	2	3	4	5
Number of clusters	114	173	39	17	7
Number of images in each class	20665	33984	4930	4634	197
Number of images AC / UAC	54649		9761		

Table 18: rough labelling of the images in the double image dataset

The survey shows that there are 287 representative cluster images with unacceptable quality (114 images from class 1 and 173 images from class 2). On the other hand, 63 representative cluster images (39 images from class 3, 17 images from class 4 and 7 images from class 5) are accepted by the customers. In total, the database consists of 54649 images that are labelled as unacceptable, and of 9761 images labelled as acceptable. An example image for each evaluation class can be seen in the appendix *A.5*.

Dividing the images into train and test data:
Again, the images of the double image dataset are divided into test and training data. First, the test dataset is separated, as shown in Figure 61.

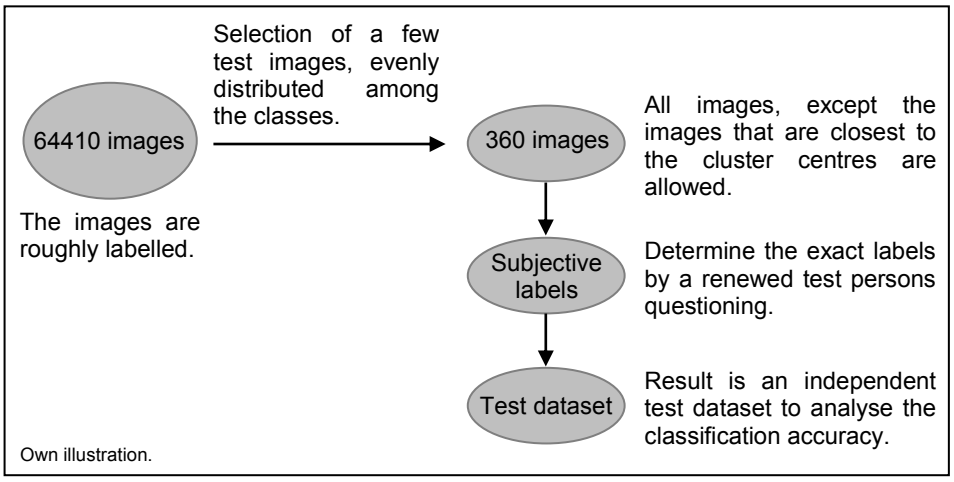


Figure 61: separation of the test dataset from the double image dataset

The test dataset consists of 360 images, which are evenly distributed over all 5 classes. Since the images of the database are roughly labelled, it is possible to select 72 imag-

es for each class. Any images could be selected except the images, which are closest to the cluster centre. Afterwards, the test images are evaluated by the test persons to find the exact labels. The 12 test persons evaluate the 360 test images in the special test environment. The time required is about 3 hours per person. The images are labelled on the rating scale from 1 (very annoying) to 5 (imperceptible). The results of the questioning are summarised in Table 19. The table shows that the test dataset consists of 147 images of unacceptable quality (classes 1 and 2) and 213 images of acceptable quality (classes 3, 4 and 5).

Subjective rating class	1	2	3	4	5
Number of images in each class	95	52	62	94	57
Number of images AC / UAC	147		213		

Table 19: labelling of the images of the test dataset for double images

The training dataset is determined in the last step. After deduction of the test dataset, there are still 64050 images available. Again, the representative training images are obtained by clustering. The cluster condition has already been determined at the beginning of this chapter. The k-means algorithm is used to find clusters, which contain images that cause the same subjective quality impression. The application of the algorithm reveals that the images could be split into 345 clusters. Consequently, the training dataset is composed of the representative images, which are closest to the cluster centre. The determination of the training dataset is illustrated in Figure 62.

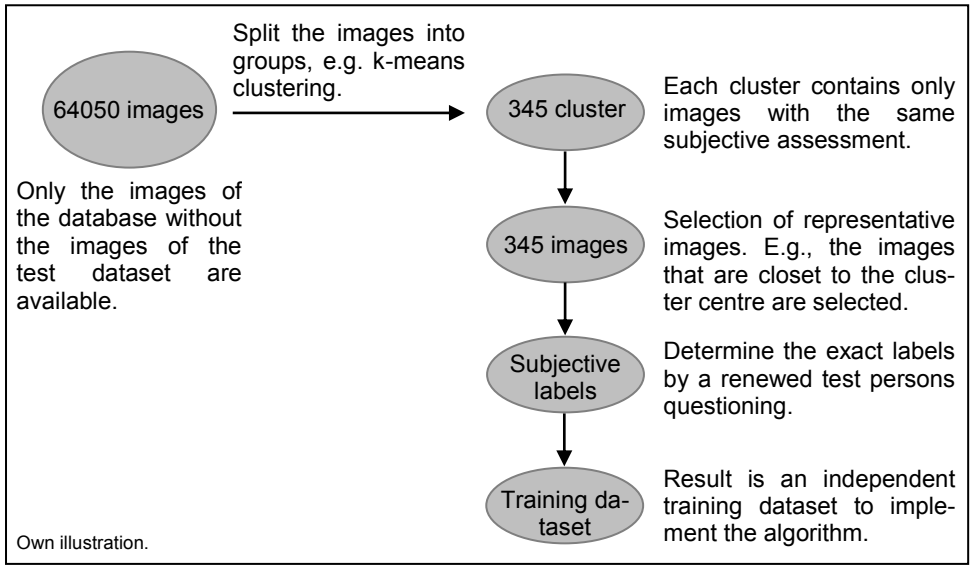


Figure 62: separation of the training dataset from the double image dataset

Finally, the images of the training dataset need to be provided with subjective labels. Therefore, the 345 train data are compared with the 350 representative images of the database. The comparison of the images shows that 307 images are identical and are already rated by the test persons. Consequently, only 38 images are labelled by 12 test persons. The survey takes place in the special test environment. Each test persons

takes about 20 minutes to label the remaining training images. The resulting labels are summarised in Table 20. The training dataset for double images consists of 284 images of unacceptable quality (rating classes 1 and 2) and 61 customer suitable images (rating classes 3, 4 and 5).

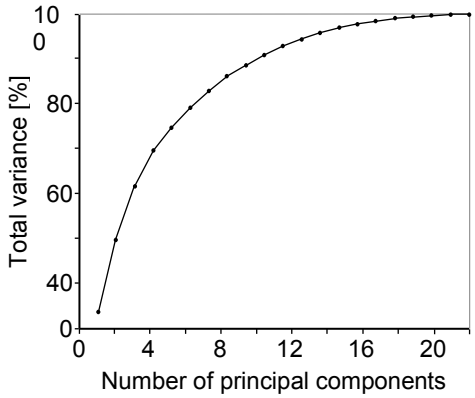
Subjective rating class	1	2	3	4	5
Number of images in each class	119	165	33	26	2
Number of images AC / UAC	284		61		

Table 20: labelling of the images of the training dataset for double images

7.2.3 Separation of the distortion and double image dataset

If the separate perception of distortions and double images is acceptable, the combination of the 2 aberration types is investigated. As the studies from the earlier chapters show, 5875 distorted images and 9761 images with different double image distances are classified as acceptable. From these images, 608 images exist that show both acceptable distortions and acceptable double images. Any of those images is described by 21 objective features.

Total variance covered by the corresponding principal components



The PCA is performed with 608 images. The quality of each image is described by 21 features.

Own illustration.
Results obtained by the program GNU PSPP.

Principal Component	Eigenvalue	Total variance covered	
		Percentage of variance	Cumulative percentage
1	7.05	33.57	33.57
2	3.36	16.02	49.58
3	2.52	12.00	61.58
4	1.69	8.05	69.63
5	1.09	5.18	74.81
6	0.92	4.37	79.18
7	0.77	3.66	82.84
8	0.70	3.35	86.19
9	0.53	2.51	88.70
10	0.48	2.28	90.98
11	0.39	1.87	92.85
12	0.33	1.58	94.42
13	0.28	1.35	95.77
14	0.25	1.21	96.98
15	0.17	0.81	97.80
16	0.16	0.75	98.54
17	0.11	0.52	99.07
18	0.10	0.45	99.52
19	0.06	0.30	99.82
20	0.04	0.18	100.00
21	0.00	0.00	100.00

Figure 63: results of the PCA for distortion and double image dataset

First, the 21 feature values of the 608 images are analysed by the PCA. The result of the PCA is shown in Figure 63. The figure shows graphically and numerically the loss of information by dimensionality reduction to a few principal components. The features are well chosen and they hardly correlate with each other. During a dimensionality re-

duction to 2 principal components, almost 50% of the total variance is covered. Here, 10 principal components are needed to cover more than 90% of the total variance. During implementing an assessment algorithm, it is necessary to check if a dimensionality reduction of the feature space is beneficial.

Since only a few images are available, the 608 images are manually labelled without previous clustering. All 608 images are labelled by the test persons and afterwards divided equally into training and test datasets, as shown in Figure 64.

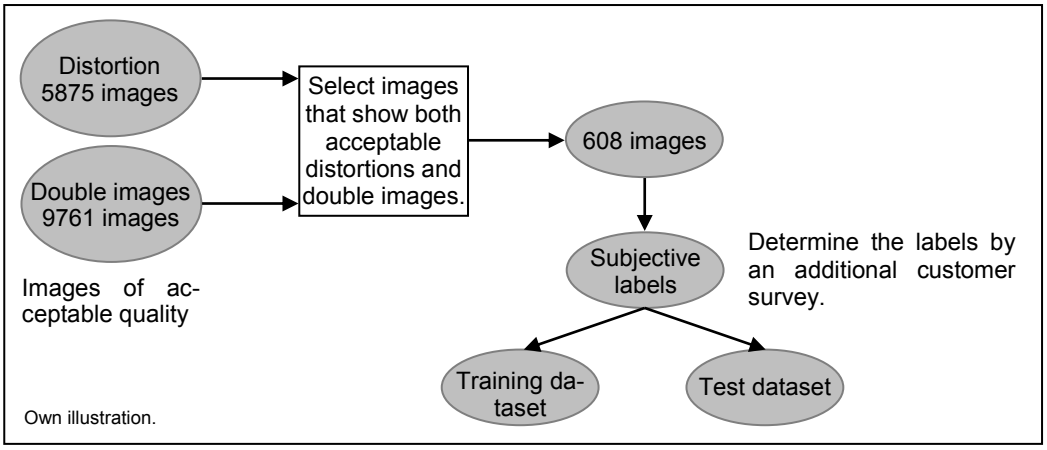


Figure 64: separation in training and test data for distortion and double images

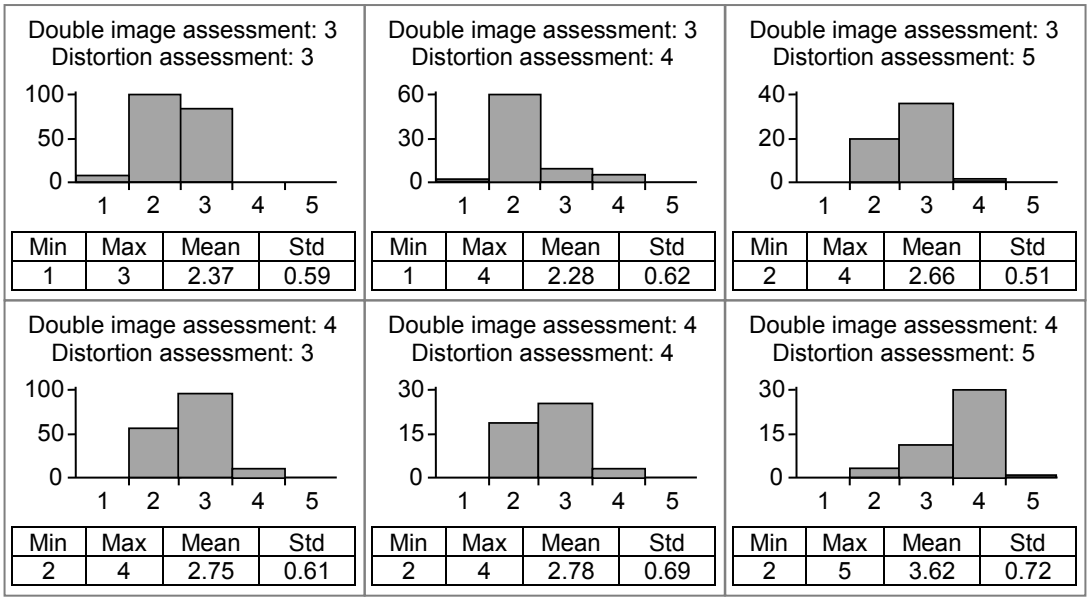
The images are labelled in the special test environment. Overall 12 test persons are asked (time involved approximately 5 hours per person). The images are evaluated on a scale from 1 (very annoying) to 5 (imperceptible), according to the ITU-R 500 directive. The results of the questioning are summarised in Table 21. The table shows that the 276 images of unacceptable quality (classes 1 and 2) and 332 images of acceptable quality (classes 3, 4 and 5) are available. An example image for each class is shown in the appendix *A.6*.

Subjective rating class	1	2	3	4	5
Number of images in each class	12	264	267	63	2
Number of images AC / UAC	276		332		

Table 21: labelling of the images in the distortion and double image dataset

To gain an overview of the relationship between the overall image quality and the separate labels for distortions and double images frequency distribution diagrams are generated, as shown in Figure 65. The diagrams show that the existing double images are evaluated separately with 3 or 4 rating points and occurring distortions with 3, 4 or 5 rating points. The first 2 diagrams show that the expectations towards the image quality are not fulfilled for the most part if images showing distortions rated with 3 or 4 points and double images labelled with 3 points. The overall impression of these images is mostly labelled with 2 rating points and partly with only 1 rating point. In contrast, the remaining combinations are mostly rated with 3 or 4 points. In addition, these combina-

tions can lead to an unacceptable overall impression. Therefore, an assessment system for the overall image quality is required because it is not possible to derive the label for the overall impression from the separate assessment of the individual aberration types.



The values are given in HUD pixels and rounded to full numbers. Own illustration.

Figure 65: relationship between the overall impression and the separate labels

To train and test the assessment algorithm, the 608 images are split into 2 independent datasets. For this purpose, the images of each rating class are divided equally into 2 parts. The distribution of the images is shown in Table 22.

Subjective rating class	1	2	3	4	5
Training dataset					
Number of images in each class	6	132	133	31	1
Number of images AC / UAC	138		165		
Test dataset					
Number of images in each class	6	132	134	32	1
Number of images AC / UAC	138		167		

Table 22: dividing the labelled images of the distortion and double image dataset into 2 parts

The training dataset consists of 138 images of unacceptable quality and 165 images of acceptable quality. Similarly, the test dataset consists of 138 images of unacceptable quality and 167 images of acceptable quality.

7.3 Limit value consideration

The limit value consideration is implemented in the first step to check whether supervised learning methods reach better classification results. The limit value consideration requires ergonomic limits for the objective features and assumes a good image quality if the objective feature values lie within these limits, see chapter *4.7*. The limit values are determined from the training dataset and verified by the test dataset. First, all images are extracted from the training dataset which are labelled with 3, 4 or 5 rating points. Since the feature values for these images are known, the minimum and maximum values for each objective feature are determined. The minimum occurring value is set as the lower limit and the maximum occurring value becomes the upper limit. This ensures that all training images that represent acceptable quality are within these limits. Then, the limit value consideration for the dataset for distortion, the dataset for double images, and the dataset for distortion and double images is carried out.

Limit value consideration for the distortion dataset:

To implement the limit value consideration for the distortion dataset, 48 training images representing an acceptable image quality and 360 test images are available. The quality of each image is described by 13 objective features. The limits that can be derived from the training images are summarised in Table 23.

Objective feature		Limits [HUD pixels] Lower	Upper
Adjusted in width	[1.1]	-4.63	5.24
Adjusted in height	[1.2]	-8.09	8.75
Aspect deviation	[1.3]	-6.27	6.85
Enlargement horizontal	[2.1]	0.54 $\xrightarrow{\text{Corrected}}$ 0	2.69
Enlargement vertical	[2.2]	0.95 $\xrightarrow{\text{Corrected}}$ 0	1.75
Rotation	[3.1]	0.15 $\xrightarrow{\text{Corrected}}$ 0	0.76
Misalignment horizontal	[3.2]	-0.97	0.06
Misalignment vertical	[3.3]	-1.95	0.72
Trapezoid horizontal	[3.4]	-1.48	0.54
Trapezoid vertical	[3.5]	0.03 $\xrightarrow{\text{Corrected}}$ 0	1.80
Smile horizontal top	[3.6]	-2.14	3.19
Smile horizontal bottom	[3.7]	-2.42	3.18
Smile vertical	[3.8]	0.22 $\xrightarrow{\text{Corrected}}$ 0	0.91

Table 23: obtained limit values for the distortion dataset

When analysing the obtained limit values, it becomes clear that the lower limit values of the features no. [2.1], [2.2], [3.1], [3.5], and [3.8] are greater than 0 HUD pixels. These limits are corrected to 0 because an undistorted HUD image has feature values of 0 HUD pixels. Thus, the quality of an ideal HUD image is recognised as acceptable. Now, the 360 test images are assessed by the limit value analysis. The image quality is classified as acceptable if the objective values for the 13 features are within the speci-

fied limits. Once a single feature value exceeds the limits, the image quality is classified as non-customisable. The result of the limit consideration is shown in Figure 66. The figure shows 2 histograms representing the prediction distribution of the limit analysis. The histogram on the left entitled “Manually UAC” shows how the test images, which are subjectively identified as unacceptable, are classified by the limit analysis. Likewise, the second histogram titled “Manually AC” shows how the limit consideration classifies images that are labelled as acceptable by the test persons.

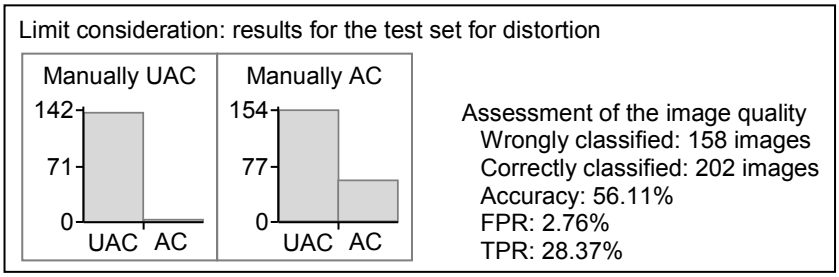


Figure 66: limit consideration: results for the distortion dataset

With 158 wrongly classified images, the classification accuracy is only 56.11%. The low accuracy value indicates that the limit value analysis is not suitable for evaluating combinations of different distortion types. The percentage of images showing an acceptable quality, which are correctly identified as customisable is only 28.37%. Here, images that are subjectively labelled as acceptable are mostly misclassified by the algorithm. Thus, the algorithm would direct many vehicles with acceptable image quality to post-processing. In contrast, images that are subjectively labelled as unacceptable are mostly considered unacceptable by the limit analysis. Thus, the proportion of manually unacceptable labelled test images that are wrongly classified as acceptable is 2.76%. Thus, only a few vehicles with an unacceptable image quality would reach the customers.

Limit value consideration for the double image dataset:

A total of 61 training images representing an acceptable image quality and 360 test images are available to implement the limit algorithm for the double image dataset. For each image, the occurring double images are captured by 8 objective features. Based on the labelled training data the limit values are determined which are summarised in Table 24. Looking closely the limit values, it is found that the upper limits are above 1 HUD pixel. As described in chapter *5.2.2*, double images with more than 1 HUD pixel could be perceived. The occurring double images of the training data are perceptible. Nevertheless, they are not perceived as annoying. On the other side, the lower limit values are greater than 0 HUD pixels. An exception is only the limit value of feature no. [1.2]. Since the feature values of an ideal HUD image without double images corresponds to 0 HUD pixels, these upper limits are set to 0. Thus, it is guaranteed that the quality of an ideal HUD image without double images is classified as acceptable.

Objective feature	Limits [HUD pixels] Lower	Upper
Maximal horizontal [1.1]	0.22 Corrected → 0	1.85
Maximal vertical [2.1]	0.47 Corrected → 0	1.67
Mean horizontal [1.2]	-1.38	2.04
Mean vertical [2.2]	0.41 Corrected → 0	1.06
95% quantile horizontal [1.3]	0.83 Corrected → 0	2.10
95% quantile vertical [2.3]	0.56 Corrected → 0	1.36
80% quantile horizontal [1.4]	0.10 Corrected → 0	1.89
80% quantile vertical [2.4]	0.91 Corrected → 0	1.82

Table 24: obtained limit values for the double image dataset

These limits are used to assess the quality of the 360 test images. The obtained results are summarised in Figure 67. The limit value consideration achieves an accuracy of 89.44% for the given classification task. In total, 38 images are classified incorrectly. The algorithm has a false positive rate of 8.16%. This means that 8.16% of the unacceptable labelled test images receive an acceptable label by the algorithm. 87.79% of all test images showing an acceptable quality are also classified as acceptable by the limit value consideration.

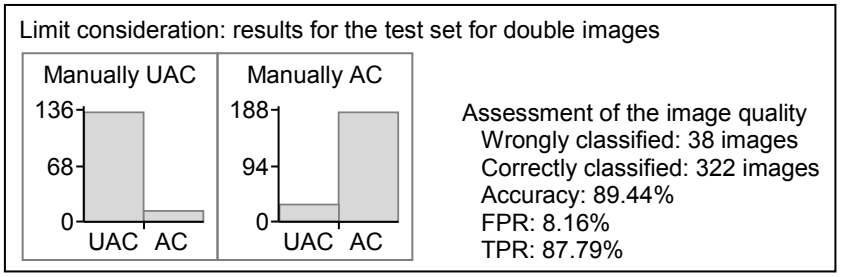


Figure 67: limit consideration: results for the double image dataset

Limit value consideration for the distortion and double image dataset:

To implement the limit value consideration, 165 training images representing an acceptable image quality and 305 test images are available. The quality of each image is described by 21 objective features. The limits that can be derived from the labelled training images are summarised in Table 25. Again, the lower limit values greater than 0 HUD pixels are matched to an ideal HUD image. The lower limit values of the features no. [2.1], [2.2] and [3.5] describing distortions, as well as the lower limit values of the features no. [1.1], [2.1], [2.2], [1.3], [2.3], [1.4], and [2.4] describing double images are set to 0. Thus, an ideal image without double images receives a positive label. The corrected limit values are used to assess the quality of the test images. The achieved classification results are summarised in Figure 68. Based on the given classification task, the limit value consideration reaches an accuracy of 67.21%. The low accuracy value indicates that the combination of various aberration types cannot be properly assessed. The problem with this algorithm is the high FPR, which indicates that 55.80% of the images with unacceptable quality receive a positive label.

Objective feature		Limits [HUD pixels] Lower	Upper
Distortion			
Adjusted in width	[1.1]	-0.71	8.32
Adjusted in height	[1.2]	-5.14	2.24
Aspect deviation	[1.3]	-4.49	0.80
Enlargement horizontal	[2.1]	0.95 Corrected → 0	3.40
Enlargement vertical	[2.2]	0.03 Corrected → 0	2.65
Rotation	[3.1]	-0.19	0.38
Misalignment horizontal	[3.2]	-0.38	0.00
Misalignment vertical	[3.3]	-0.38	1.34
Trapezoid horizontal	[3.4]	-0.57	0.40
Trapezoid vertical	[3.5]	0.31 Corrected → 0	1.15
Smile horizontal top	[3.6]	-1.65	1.22
Smile horizontal bottom	[3.7]	-1.55	2.12
Smile vertical	[3.8]	-0.67	1.39
Double image			
Maximal horizontal	[1.1]	0.52 Corrected → 0	1.46
Maximal vertical	[2.1]	1.19 Corrected → 0	1.72
Mean horizontal	[1.2]	-0.28	0.02
Mean vertical	[2.2]	0.05 Corrected → 0	0.33
95% quantile horizontal	[1.3]	0.44 Corrected → 0	1.22
95% quantile vertical	[2.3]	1.07 Corrected → 0	1.66
80% quantile horizontal	[1.4]	0.05 Corrected → 0	0.62
80% quantile vertical	[2.4]	0.67 Corrected → 0	1.07

Table 25: obtained limit values for the distortion and double image dataset

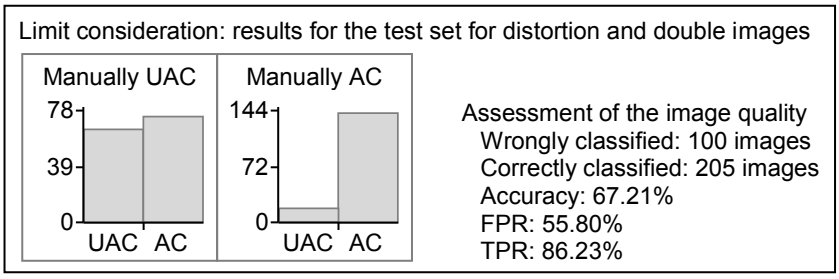


Figure 68: limit consideration: results for the distortion and double image dataset

Thus, many vehicles that are not customer suitable would leave the manufacturing process because the limit analysis confirms acceptable image quality. On the other hand, the TPR value is 86.23%. This high value indicates that nearly 90% of vehicles with acceptable image quality would be correctly classified as customer suitable.

7.4 Supervised learning: implementation of classifiers

It is examined whether classification methods can achieve a better classification performance than the standard limit value consideration. Here, supervised learning (SL) methods such as the polynomial classifier[9], the nearest neighbour classifier, and the learning vector quantisation are used. The classification algorithms are implemented especially for this thesis. The polynomial classifiers are based on the equations introduced in chapter *6.4.3*. The classification task is implemented in Matlab. Classifiers of 1st, 2nd, 3rd and 4th polynomial order are realised. The k-nearest neighbour classification is introduced in chapter *6.4.1*. Here k-nearest neighbour classifiers for $k = 1$, $k = 3$, and $k = 5$ are implemented in Matlab. In addition, for the PC and the kNN, the feature space is reduced to different dimensions (PCA) to find the best number of principal components for the given classification problem[10]. The theory of the learning vector quantisation is described in chapter *6.4.2*. The LVQ classification is implemented in the free software environment *R*. Here, the prototypes are determined by the learning rules LVQ1, LVQ2.1, and LVQ3. Functions of the existing package *LVQTools* are used. In the following is checked how many prototypes are required and according to which learning rule these prototype vectors have to be determined to achieve the best classification result for the given classification problem.
Finally, the classification problem is reduced to 2 possibilities and the ROC curve of each classifier type is determined by varying the classification parameters.
Using the training dataset, the various classifiers are trained to find the relationship between the objective features and the customer acceptance. Subsequently, the quality of the trained classifiers is checked by applying the independent test dataset. Successively, assessment algorithms for the distortion dataset, the double image dataset and the distortion and double image dataset are implemented.

7.4.1 SL: assessment algorithms for the distortion dataset

To implement an assessment algorithm for distortions, 1006 training images and 360 test images are available. The image quality is described by 13 objective feature values. The classification results are compared with the results of the limit value analysis. The limit value analysis reaches the accuracy of 56.11%, the TPR of 28.37% and the FPR of 2.76%, see chapter *7.3*.

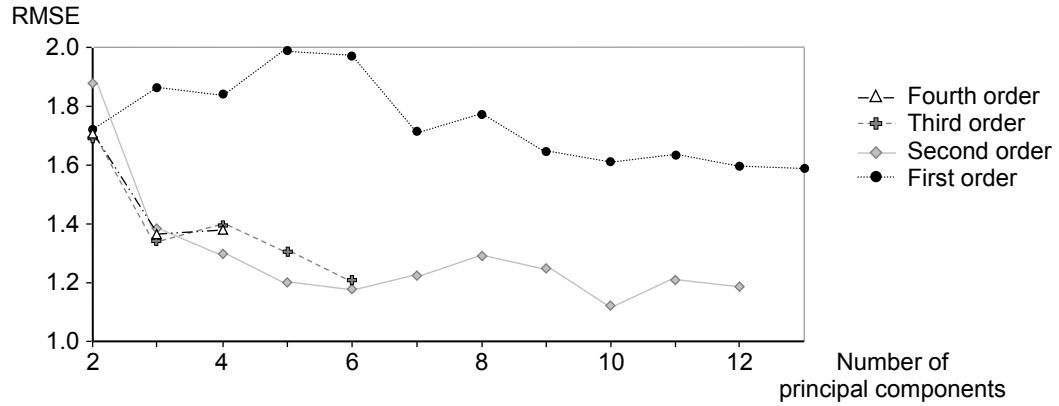
Polynomial classification:

The number of training images should be 10 times greater than the number of polynomial terms [MARSLAND 11]. Thus, it is first examined which complexity of the classifiers can be realised with the given number of training images. With 1006 training images it is possible to implement 1st order classifiers up to all 13 feature dimensions. The 2nd order classifiers can be implemented up to 12 principal components and the 3rd order

[9] The possibility to evaluate the image quality with a polynomial classifier was already realised in a preliminary investigation by [HELLMANN 12].
[10] In the preliminary investigation of [HELLMANN 12] the dimensions are also reduced by the PCA.

classifiers are analysed up to 6 principal components. Finally, the 4th order classifiers can only be realised up to 4 feature dimensions.
The assessment quality of the trained classifiers is checked with the test images. The trained classifiers assign the test image to the class with the highest probability. First, the resulting RMSE values of the classifiers are analysed for different polynomial orders up to the maximum possible feature dimensions. The values are summarised in the upper diagram of Figure 69.

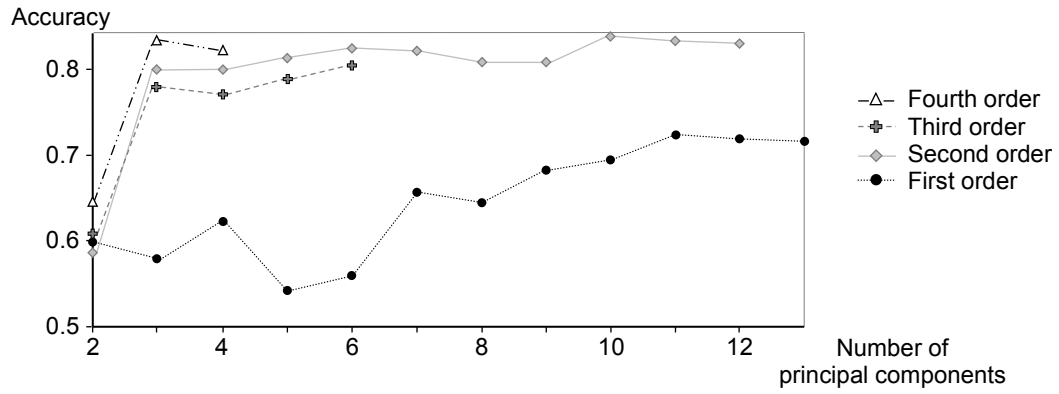
Development of the RMSE values for different numbers of principal components



Resulting minimal RMSE values for the different classifier types

	1st order	2nd order	3rd order	4th order
RMSE	1.59	**1.12**	1.21	1.36
no. of principal components	13	**10**	6	3

Development of the classification accuracy for different numbers of principal components



Resulting maximal accuracy values for the different classifier types

	1st order	2nd order	3rd order	4th order
Accuracy	72.50%	**83.89%**	80.56%	83.33%
no. of principal components	11	**10**	6	3

Only the meaningful RMSE values and the accuracy values are shown which result from classifiers for which the number of training samples is 10 times greater than the number of polynomial terms, according to [MARSLAND 11]. Own illustration.

Figure 69: SL, PC for the distortion dataset: resulting RMSE and accuracy values

The x-axis shows the number of principal components, namely the dimension of the feature space. The resulting RMSE values are plotted on the y-axis. It can be seen that

the development of the RMSE values is very similar for the 2nd, 3rd and 4th order classifiers. As the number of principal components increases, the RMSE values decrease. The decrease of the RMSE values when increasing the dimension of the feature space is typical. The increasing number of principal components reduces information loss caused by dimensionality reduction. It also shows that the RMSE values of 1st order classifiers differ significantly from the others. The values raise first and hardly decrease afterwards. In addition, the RMSE values are significantly higher than the values for the 2nd, 3rd and 4th order classifiers. This also shows that the simple 1st order classifier is not suitable for handling the complex classification task.
The resulting minimal RMSE values of the classifiers of different polynomial orders are summarised in the table below the upper diagram of Figure 69. It can be seen that the 2nd order PC for 10 principal components results in the lowest RMSE of 1.12. This very high dimension of the feature space can be explained by the fact that the features are hardly correlated. Since the features contain only slightly redundant information, a dimensionality reduction is only possible with loss of information, see chapter *7.1.1*. The 3rd order PC achieves the second-best value (RMSE = 1.21) for 6 principal components. The third-best result (RMSE = 1.36) is achieved for a classifier of 4th order in the 3-dimensional feature space. Far behind is the RMSE value (RMSE = 1.59) for the 1st order classifier that requires all 13 features without dimensionality reduction.

In the next step, in addition to the resulting RMSE values, the development of the classification accuracies is also analysed. The accuracy values are summarised in the second diagram of Figure 69. Again, only the values up to the maximum possible feature dimensions are shown. Comparing the development of the RMSE values with the development of the classification accuracies shows that the values have an inverse trend. If the RMSE values decrease, the accuracy values increase and vice versa. It can be seen that by increasing the dimension of the feature space, the classification accuracy increases. The maximal classification accuracies are summarised in the table below the second diagram of Figure 69. Again, the 2nd order PC achieves the highest accuracy of 83.89% in the 10-dimensional feature space. The second-best classification accuracy of 83.33% is reached by the 4th order classifier for 3 principal components. The maximal accuracy of 80.56% reaches the 3rd order classifier in the 6-dimensional feature space. Far behind is the accuracy of the 1st order classifier for 11 principal components. This classifier only achieves an accuracy of 72.50%. Compared with the obtained accuracy of the limit value analysis (56.11%), the maximum accuracies of all 4 classifiers are significantly higher.

Since the 2nd order polynomial classifier achieves both the minimal RMSE value and the maximal classification accuracy for 10 principal components, the classification result is investigated in more detail, as shown in Figure 70. The figure shows 7 histograms representing the prediction distribution of the PC. The top row contains 5 histograms. Each diagram represents the class in which the test images are sorted manually. The diagram bars show how the classifier rates the images. For example, the left

histogram titled “Manually 1” shows how the PC rates the images that are labelled by the test persons with 1 rating point. This kind of presenting the classification results has already proven itself at the Daimler AG [HELLMANN 12]. The lower histograms can be derived from the top 5 histograms. The left-hand histogram shows how the subjectively unacceptable test images (rating class 1 and 2) are classified. The right-hand histogram shows the labelling of acceptable images (rating class 3, 4 and 5).
The histograms show that the classifier assigns images that are manually labelled with 1 rating point to the appropriate rating class almost without errors. Images that are manually labelled with 3 rating points are also largely assigned to this class by the classifier. In contrast, images that are manually rated with an average of 2 rating points are mostly miscategorised by the PC. Images that are manually labelled with 4 or 5 rating points are assigned by the classifier to rating classes 3, 4 and 5. However, these misclassifications are considered uncritical, since the rating classes 3, 4 and 5 represent an acceptable image quality. The resulting TPR value is very high. Here, 99.53% of the test images with an acceptable quality receive a positive label. Thus, less than 1% of the vehicles with customer suitable HUD images would be wrongly sent to rework. The percentage of subjectively unacceptable labelled test images that are wrongly classified as being acceptable is 39.31%. Thus, many vehicles with an unacceptable image quality would reach the customers.

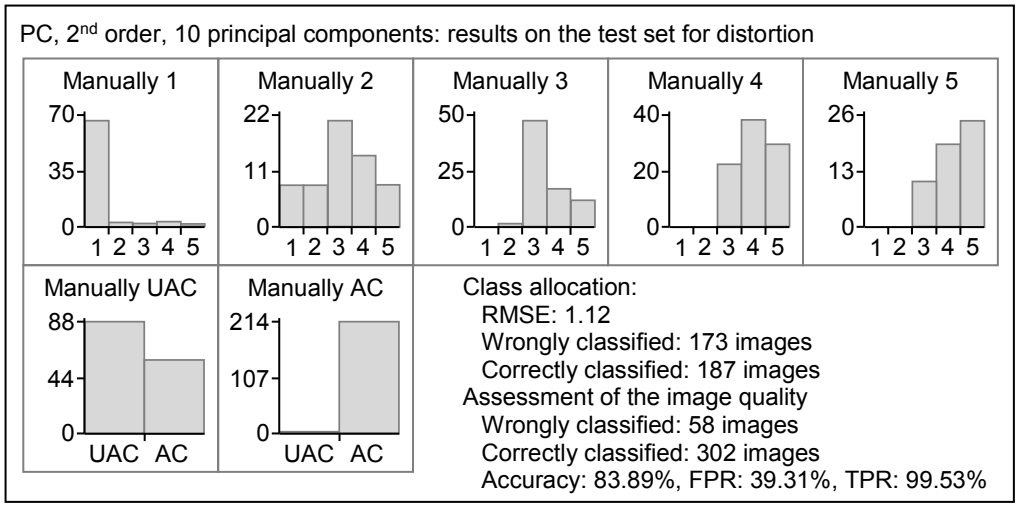


Figure 70: SL, PC for the distortion dataset: possible labelling of the test images

Next, it is checked which classifier type obtains the minimal FPR and the maximal TPR for the given classification task. For the maximum possible structure dimension, the resulting values are summarised in Table 26. The minimal FPR of 39.31% is reached by the 2nd order PC in the 10-dimensional component space. However, this classifier achieves only a TPR of 99.53%. The classifiers, which achieve a TPR of 100% (2nd order classifier for 11 principal components and 4th order classifier for 3 components) reach a minimal FPR of 41.38%. Thus, these 2 algorithms would direct no vehicles with an acceptable image quality to rework. Overall, the PC tends to label the test images as acceptable.

Comparing these values with the results of the limit value consideration (FPR = 2.76%, TPR = 28.37%) shows that the limit consideration reaches a significantly lower FPR than the polynomial classifiers. Conversely, the classifiers achieve a significantly higher TPR than the limit consideration. These results show clearly that the choice of the assessment method is a compromise between a low FPR and a high TPR.

PC	no. of principal components	FPR	TPR
	Minimal FPR values		
1st order	2	45.52%	63.72%
2nd order	10	39.31%	99.53%
3rd order	6	48.28%	100.00%
4th order	3	41.38%	100.00%
	Maximal TPR values		
1st order	12	53.79%	89.30%
2nd order	11	41.38%	100.00%
3rd order	6	48.28%	100.00%
4th order	3	41.38%	100.00%

Table 26: SL, PC for the distortion dataset: min FPR and max TPR

Finally, the allocation rule of the PC is modified and reduced to a 2-class problem. Up to now, the test image is assigned to the class with the greatest probability. This is now changed that the decision for an acceptable quality is only made if the probability that the feature vector belongs to an acceptable quality is higher by a multiple ϕ than the probability of belonging to an unacceptable quality [KRISTIAN et al. 11: p. 203]. The constant ϕ varies between 0.1 and 10. The increase takes place in steps of 0.1. For each value of ϕ, the resulting classifier is applied to the test images and the FPR and TPR values are calculated. This investigation is performed on the classifiers of different polynomial orders up to the maximum possible feature dimensions. For each classifier, the resulting ROC curve is determined and the area under the curve (AUC) is calculated. The curves that result for each polynomial order in the maximum AUC are shown in Figure 71.

The maximum AUC value of 0.87 is reached by the 2nd order classifier in the 7-dimensional component space. The 3rd order classifier achieves the second-best value of 0.81 for 4 principal components. The values of 1st and 4th order classifiers in the 12 and 3-dimensional component space are lower. These classifiers reach an AUC value of 0.68 and 0.67, respectively. The ROC diagrams show that the choice of the most suitable classifier type is always a compromise between a high TPR and low FPR. An increase in the TPR is accompanied by an increase in the FPR.

The resulting maximum values of the area under the curve are shown. The classification ability of the classifier is better if the AUC value is higher.

	1st order	2nd order	3rd order	4th order
AUC	0.68	**0.87**	0.81	0.67
no. of principal components	12	**7**	4	3

ROC-curves: these diagrams illustrate the compromise between the matches (correctly positive classifications) and the costs (false positive classifications).

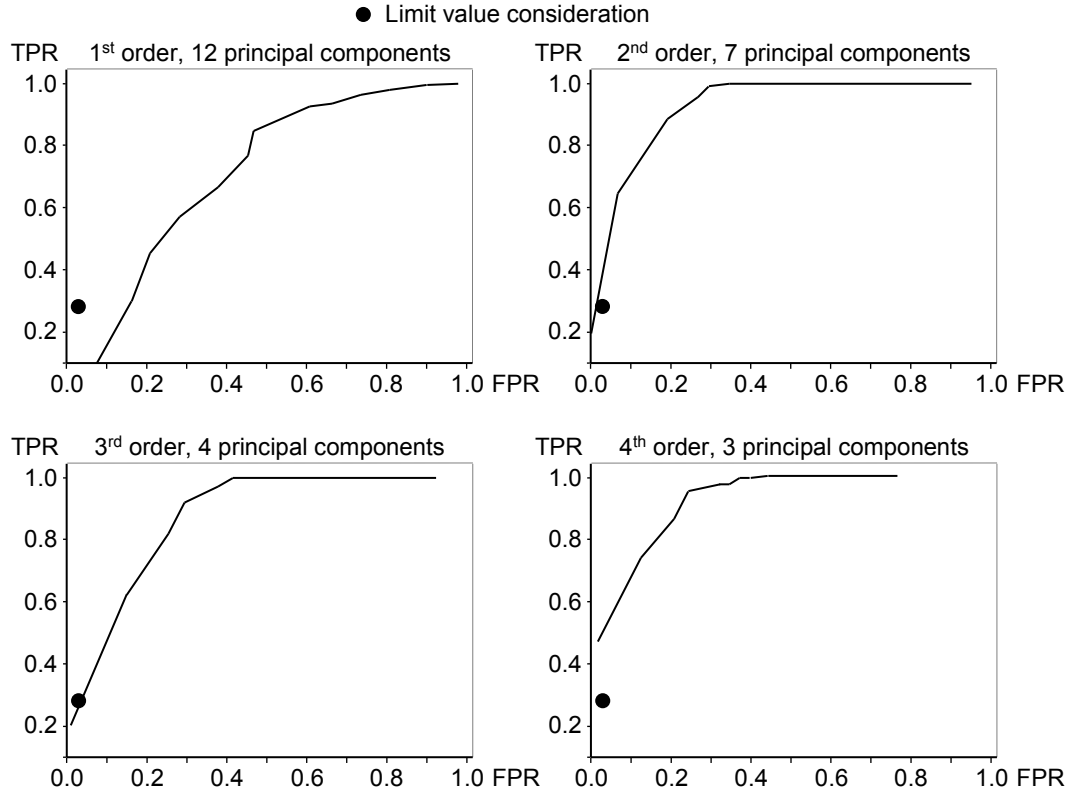


Figure 71: SL, PC for the distortion dataset: ROC curves for 2-class classifiers

For example, the 4th order classifier in the 3-dimensional component space with a constant ϕ of 0.1 is applied to the test images. A possible classification result of the test images is shown in Figure 72. For the given classification task, this classifier seems to be better suited than the limit value consideration. Here, a higher accuracy, a higher TPR and a lower FPR are achieved. The evaluation of the test data gives an accuracy of 67.78%, a false positive rate of 1.38% and a true positive rate of 46.98%. Thus, only 1.38% of the test images that are manually labelled as unacceptable receive a positive label. Likewise, 46.98% of the test images that are manually labelled as acceptable are also recognized as such. Overall, the 2-class PC tends to label the test images as unacceptable. This could be because the classifier is trained with a lot more unacceptable images than with acceptable images. Consequently, the classifier is too well adapted to the training dataset and the generalisation is not completely given.

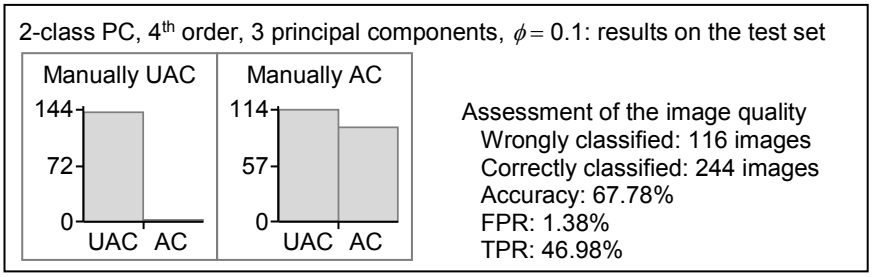


Figure 72: SL, 2-class PC for the distortion dataset: possible labelling of the test images

K-nearest neighbour classification:

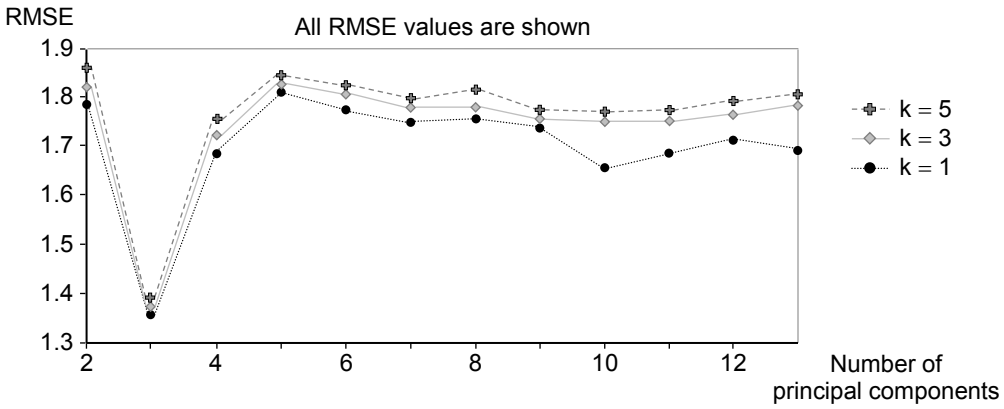
For $k = 1$, the test image is assigned to the class of the most similar test image. For $k = 3$ and $k = 5$, the class distribution of the *k*-nearest neighbours is analysed. If the majority of the nearest neighbours are classified as AC (class 3, 4 and 5) or UAC (class 1 and 2), the test image is assigned to the same class as the nearest training image of the majority belongs.

The following investigates which classifier type obtains the best prediction results for the given classification problem. First, the development of the RMSE values for different feature dimensions is summarised in the upper part of Figure 73. Here, the RMSE values are plotted on the y-axis of the diagram. The number of used principal components is plotted on the x-axis. The diagram shows that the development of the RMSE values is very similar for $k = 1$, $k = 3$, and $k = 5$. Over the entire principal component space, the RMSE values are lowest for $k = 1$. The RMSE values for $k = 3$ are slightly above the values for $k = 1$. For $k = 5$, the highest RMSE values are achieved. It is remarkable that the curves show a decrease in the RMSE values for 3 principal components. Thereafter, the RMSE values increase for 4 and 5 principal components. From 6 to 13 principal components, the values oscillate slowly.

Since the RMSE values in the 3-dimensional component space differ from the other values, it is investigated why this might be the case. It is assumed that the proportion of the variance explanation of the subjectively relevant features by the principal components matters. Therefore, the rotated component-loading matrix for 3 principal components is shown in the middle table of Figure 73. Looking at the principal components 1 and 2, it can be seen that they mainly represent the features no. [1.1] - *Adjusted in width*, no. [1.2] - *Adjusted in height* and no. [1.3] - *Aspect deviation*. To be seen in the last 3 rows of the table. Despite the high variance explanation of these features, the first 2 principal components are not relevant for the assessment of the subjectively perceived image quality. The first 3 rows of the table show that the variance of the subjectively relevant features (no. [3.1] - *Rotation*, no. [3.2] - *Misalignment horizontal* and no. [3.3] - *Misalignment vertical*) is strongly covered by the 3rd principal component. Therefore, this principal component is most important to describe the subjectively perceived image quality. This may explain the good agreement between the predicted class labels and the subjectively perceived quality of the test data. When more than 3 principal components are selected, more and more variance is covered by features that are less

relevant to describe the perceived quality. Thus, the further increase in the RMSE values could be explained.

Development of the RMSE values for different numbers of principal components



Rotated component loading matrix – try to explain why the RMSE values are lowest for 3 dimensions

The following table contains the rotated component-loading matrix for 3 dimensions. The features are sorted according to the subjective relevance; see Table 6.

Objective feature		Principal components 1	2	3
Rotation	[3.1]	0.65	-0.12	0.62
Misalignment horizontal	[3.2]	-0.16	0.02	0.88
Misalignment vertical	[3.3]	-0.08	-0.09	0.75
Smile horizontal top	[3.6]	0.08	0.71	0.14
Smile horizontal bottom	[3.7]	0.24	0.76	-0.09
Smile vertical	[3.8]	0.43	0.17	-0.24
Enlargement horizontal	[2.1]	0.29	0.01	-0.42
Enlargement vertical	[2.2]	0.31	-0.52	0.02
Trapezoid horizontal	[3.4]	0.73	0.54	0.10
Trapezoid vertical	[3.5]	0.62	0.20	-0.26
Adjusted in width	[1.1]	0.04	-0.84	0.19
Adjusted in height	[1.2]	-0.96	0.00	0.14
Aspect deviation	[1.3]	-0.96	0.15	0.10

The variance of the subjectively relevant features is strongly covered by the third component.

Despite the high variance explanation of these features, the principal components 1 and 2 are not relevant to evaluate the perceived image quality.

Resulting minimal RMSE values for the different classifier types

	k = 1	k = 3	k = 5
RMSE	**1.37**	1.39	1.40
no. of principal components	**3**	3	3

Figure 73: SL, kNN for the distortion dataset: resulting RMSE values

The table shown in the lower part of Figure 73 summarises the resulting minimal RMSE values of each classifier type. It can be seen that the NN classifier achieves the minimum RMSE value of 1.37 for $k = 1$ and a dimension of 3 principal components. The classifier for $k = 3$ has the second-best RMSE value of 1.39 in the 3-dimensional component space. The worst result (RMSE = 1.40) is reached by $k = 5$ and a dimension of 3 principal components.

The labels of the test images that result in the lowest RMSE value are shown in Figure 74. The second histogram in the upper row shows that the classifier is able to assign images that are manually labelled with 2 rating points to the appropriate rating class

almost without errors. Images that are manually labelled with 1 rating point are assigned to rating classes 1 or 2 by the classifier. However, these misclassifications are considered uncritical, since rating classes 1 and 2 both represent an unacceptable image quality. In contrast, images that are manually labelled with 3, 4 or 5 rating points are assigned by the NN classifier primarily to rating class 2. These assignments are critical because the classifier assigns images of acceptable quality to rating class 2, which represents an unacceptable image quality. Overall, the classifier tends to assign the test images to rating classes 1 and 2, which both represent an unacceptable quality.

The last 2 histograms show that the percentage of customer suitable images that are correctly labelled as customer suitable is only 32.56%. Here, many vehicles with an acceptable image quality would be wrongly sent to rework. On the other hand, images that are manually labelled as unacceptable are usually labelled properly by the classifier. The FPR value is only 6.21%. Thus, only a few vehicles with an unacceptable image quality would reach the customers. Overall, the NN classifier achieves an accuracy of 57.22%.

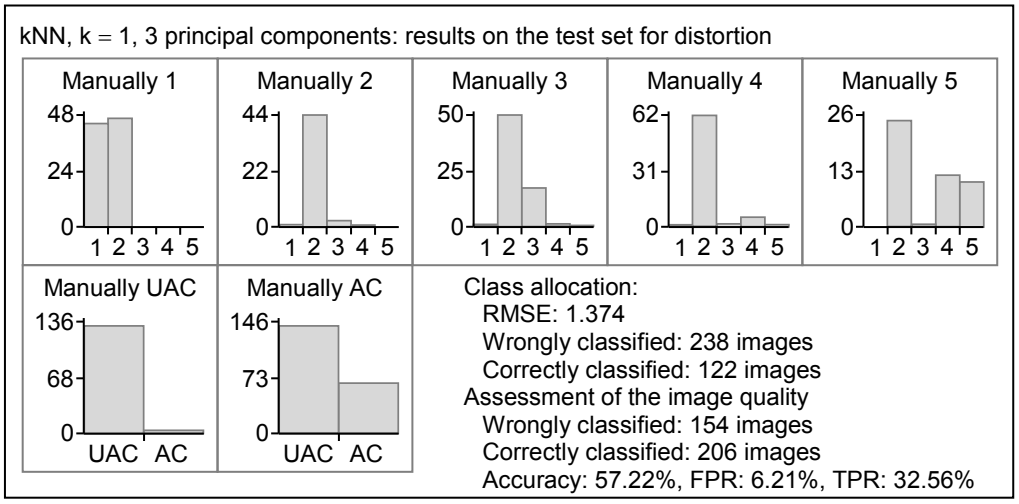
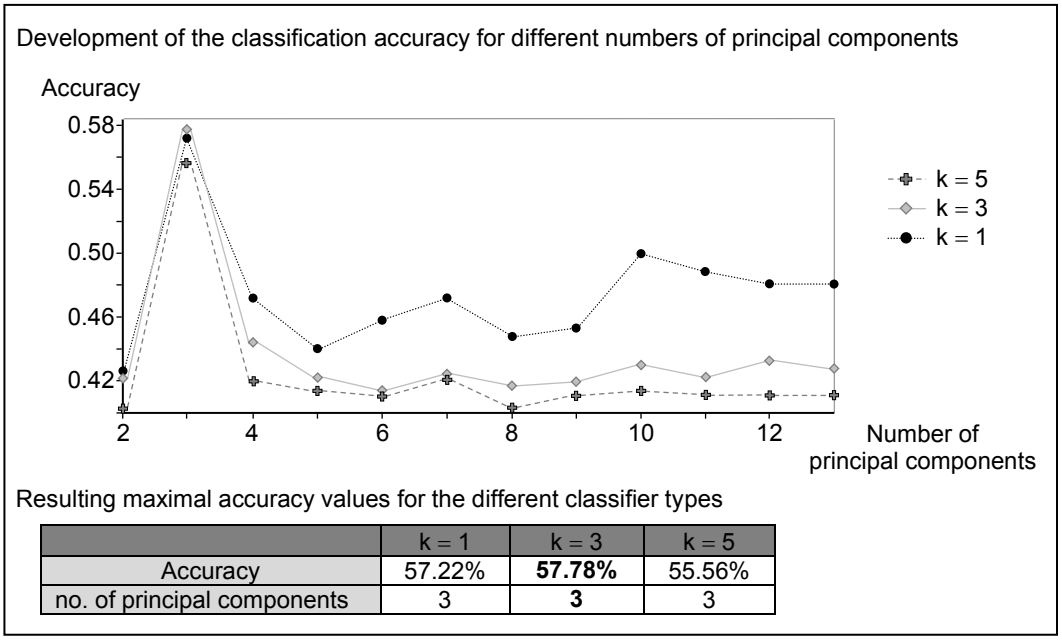


Figure 74: SL, kNN for the distortion dataset: possible labelling of the test images

The next step is to examine the classification accuracies across the different feature dimensions. The development of the classification accuracies is shown in Figure 75. The x-axis shows the different dimensions of the feature space and the y-axis the resulting classification accuracies. By comparing the development of the RMSE values with the development of the accuracy values, it can be seen that the development is nearly inverse. In the 3-dimensional feature space, the classifiers reach the maximal accuracy. Besides, the classifier that considers 3 nearest neighbours achieves the highest value of 57.78%. Except for the result of the classifier for $k = 5$, the other 2 classifiers reach a higher accuracy than the limit value consideration.



Resulting maximal accuracy values for the different classifier types

	k = 1	k = 3	k = 5
Accuracy	57.22%	**57.78%**	55.56%
no. of principal components	3	**3**	3

Figure 75: SL, kNN for the distortion dataset: resulting accuracy values

Now, it is investigated which classifier reaches the lowest FPR and the highest TPR The values are summarised in Table 27.

NN	no. of principal components	FPR	TPR
	Minimal FPR values		
k = 1	11	4.83%	17.67%
k = 3	10	0.00%	4.65%
k = 5	7	0.00%	3.26%
	Maximal TPR values		
k = 1	3	6.21%	32.56%
k = 3	3	2.07%	30.70%
k = 5	3	2.07%	26.98%

Table 27: SL, kNN for the distortion dataset: min FPR and max TPR

The k-nearest neighbour classifiers for k = 3 and k = 5 reach the ideal FPR of 0% in the 10 and 7-dimensional feature space, respectively. No vehicles with an unacceptable image quality would reach the customer. Unfortunately, the TPR values are only 4.65% and 3.26%, respectively. This means that almost all vehicles with an acceptable quality would be checked in the rework process. When comparing these values with the results of the limit consideration (accuracy = 56.11%, TPR = 28.37%, FPR = 2.76%), it can be seen that the FPR values of the classifiers are well below the FPR of the limit consideration. However, the limit consideration reaches a TPR value that is significantly better than the TPR values of the classifiers. The maximal TPR value of 32.56% is reached by the NN classifier for k = 1. However, this classifier type only achieves the FPR of 6.21%. Thus, the TPR is higher than the value of the limit consideration, while

the FPR is also higher. This shows again that the choice of the appropriate classification algorithm is a compromise between a low FPR and a high TPR.

Finally, the allocation rule of the kNN is modified and reduced to a 2-class problem. So far, the test image is assigned to that class to which the training image is most similar. Now the test image is assigned as acceptable if at least *m* of the *k*-nearest training images are classified as AC. Otherwise, the test image is assigned as UAC. The number of nearest training images to be considered is set to k = 30, k = 20, and k = 10. At the same time, the number of images *m* required to classify a test image as AC is varied between 1 and k. Each classifier is applied to the test images and the resulting ROC curve is determined. This investigation is also done for different feature dimensions. The curve, which results in the maximum AUC value, is shown in Figure 76.

The resulting maximum values of the area under the curve are shown. The classification ability of the classifier is better if the AUC value is higher.

	k = 10	k = 20	k = 30
AUC	0.87	0.86	**0.88**
no. of principal components	3	3	**3**

ROC-curve: this diagram illustrates the compromise between the matches (correctly positive classifications) and the costs (false positive classifications).

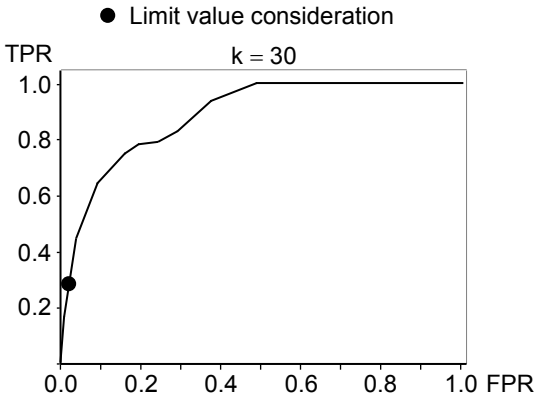


Figure 76: SL, kNN for the distortion dataset: ROC curve for a 2-class classifier

The maximal possible AUC values are reached by the classifiers in the 3-dimensional feature space. The maximum achievable area under the curve is in the range of 0.86 and 0.88. This shows again that the choice of the most suitable classifier type is a compromise between a high TPR and low FPR. Overall, the results of the 2-class classifiers are very similar to the results of the limit value consideration.

Learning vector quantisation:

The prototypes are determined by the learning rules LVQ1, LVQ2.1 and LVQ3. A test image is assigned to the same class as the nearest prototype vector. The quality of the LVQ classifier depends on the prototypes. Thus, LVQs with different numbers of prototypes are determined from the training dataset. The resulting classifiers are applied to

the test dataset and the RMSE values and the accuracy values are calculated as shown in Figure 77.

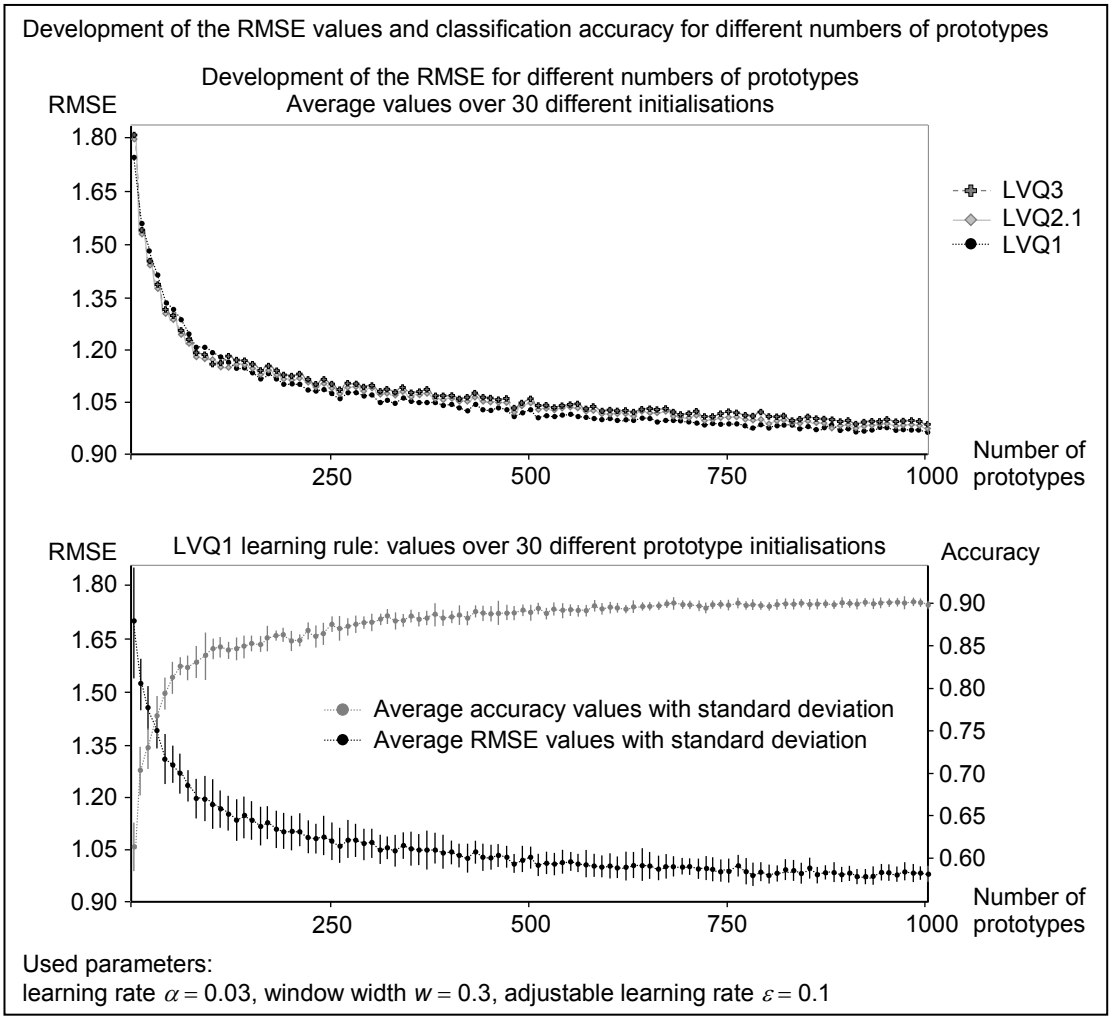


Figure 77: SL, LVQ for the distortion dataset: evaluate different numbers of prototypes

For the determination of the prototypes, a learning rate α of 0.03, a window width *w* of 0.3 and an adjustable learning rate ε of 0.1 are used. Since the resulting prototypes depend on the random initialisation, the average values are calculated over 30 different initialisations and the number of prototypes is increased in intervals of 10. The x-axis shows the number of prototypes. The RMSE values and the classification accuracies are plotted on the y-axis. The upper diagram of Figure 77 shows that the development of the RMSE values is very similar for the learning rules LVQ1, LVQ2.1, and LVQ3. Between 5 and 155 prototypes, the RMSE values are lowest for the LVQ2.1 learning rule. The curve of the LVQ3 learning rule is slightly above the curve of LVQ2.1 and the curve of the LVQ1 learning rule is the highest. From 165 prototypes, the curve changes for the LVQ1 learning rule. Now, the RMSE values are lowest for this learning rule. Between 5 and 65 prototypes, the RMSE curves decrease rapidly and almost linearly. After that, the decline is much lower. The lower diagram of Figure 77 shows the development of the RMSE values and the development of the classification accuracies of the

LVQ1 learning rule. In addition to the average values, the standard deviations over 30 different initialisations are plotted for each value. It can be seen that the resulting RMSE value and the classification accuracy depend on the random initialisation of the prototypes. The fewer prototypes are initialised, the greater the standard deviation of the RMSE values and the accuracy values.
It is now investigated how the FPR and the TPR develop for different numbers of prototypes. Again, the prototypes are determined by the learning rule LVQ1. The average values and the standard deviation over 30 different initialisations are calculated. The number of prototypes is increased in intervals of 10. The results are summarised in Figure 78. The x-axis shows the number of prototypes. The percentages of the FPR and TPR are plotted on the y-axis.

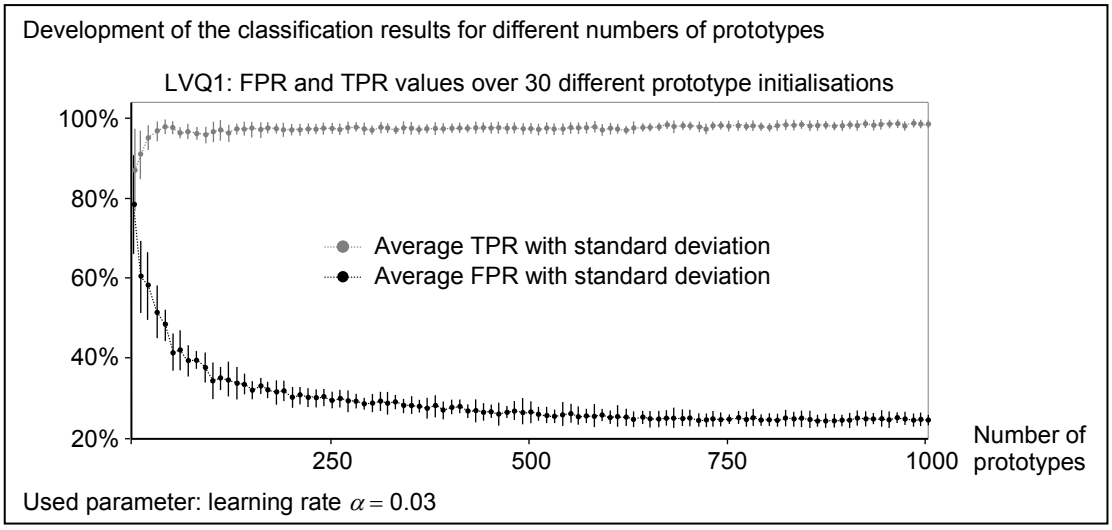
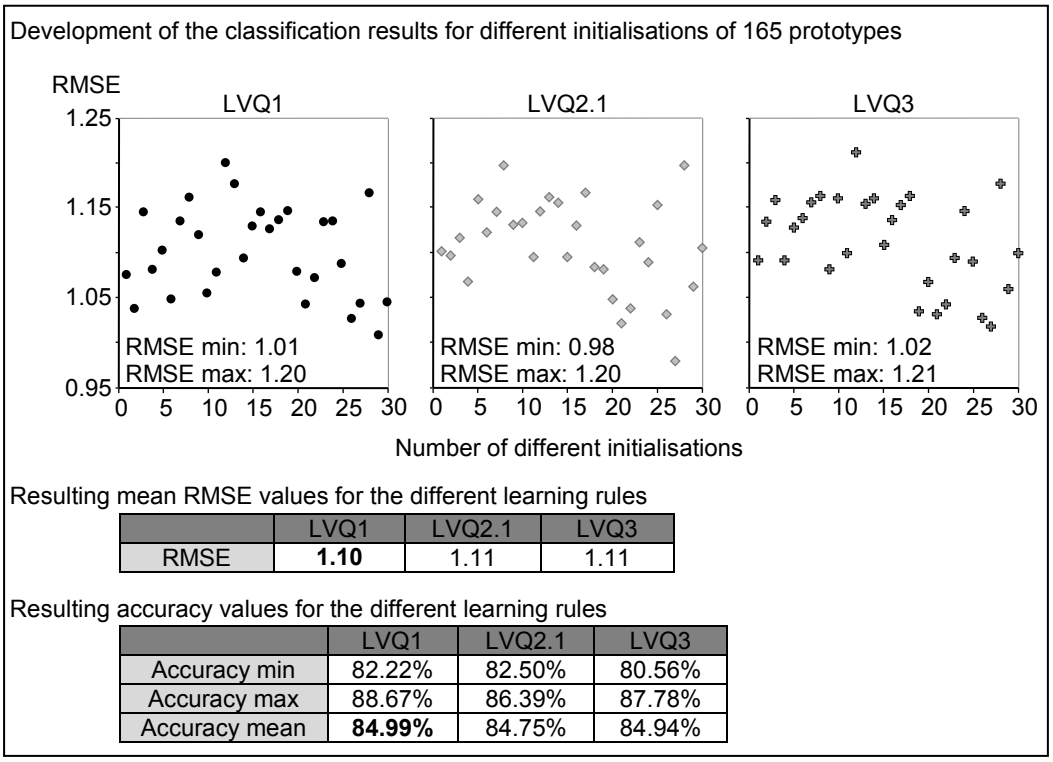


Figure 78: SL, LVQ for the distortion dataset: resulting values for learning rule LVQ1

The figure shows that the TPR is very high for all prototype numbers. For 5 prototypes, the average TPR is already 86.99%. If the number of prototypes is increased, the average TPR increases above 99% in the first 13 iterations. For all numbers of prototypes, the LVQ reaches a higher TPR than the limit value analysis (28.37%). The FPR values decrease by increasing the number of prototypes. The average FPR falls on average from 78.47% to 24.72%. Here, the largest reduction is achieved within the first 17 iterations. Thereafter, the FPR only decreases slowly, although the number of prototypes is further increased. However, the low FPR of the limit value consideration (2.76%) is not achieved.
For the following investigations, 165 prototypes are used. The number of prototypes results from the diagrams in Figure 77. This prototype number is located in the middle of the bend of the RMSE and accuracy curves. Thus, the best relationship between the number of prototypes, the RMSE values and the classification accuracy is expected. If fewer prototypes are used, the RMSE values increase sharply and the accuracies decrease equally. In contrast, if more prototypes are used, the decrease of the RMSE values is low and the accuracies increase only slowly. Since the RMSE value and the

classification accuracy depend on the random initialisation of the prototypes, the results of 30 different initialisations are considered in more detail, as shown in Figure 79.



Resulting mean RMSE values for the different learning rules

	LVQ1	LVQ2.1	LVQ3
RMSE	**1.10**	1.11	1.11

Resulting accuracy values for the different learning rules

	LVQ1	LVQ2.1	LVQ3
Accuracy min	82.22%	82.50%	80.56%
Accuracy max	88.67%	86.39%	87.78%
Accuracy mean	**84.99%**	84.75%	84.94%

Figure 79: SL, LVQ for the distortion dataset: resulting values for 165 prototypes

The upper part of Figure 79 shows 3 diagrams representing the resulting RMSE values of the learning rules LVQ1, LVQ2.1 and LVQ3. The number of different initialisations is plotted against the resulting RMSE values. The diagrams show that the random initialisation of the prototypes affects the resulting RMSE values. The average RMSE values are summarised under the diagrams of Figure 79. On average, the LVQ1 learning rule shows the lowest RMSE value of 1.10, followed by LVQ2.1 and LVQ3 with the RMSE value of 1.11. Likewise, the resulting accuracies are summarised in the last table of Figure 79. The resulting average accuracy values of all 3 learning rules are very close together. Nevertheless, the highest average value of 84.99% is achieved by the LVQ1 learning rule. Here, all 3 learning methods lead to classifiers, which achieve a much higher accuracy than the limit analysis (56.11%).

The classification result for an arbitrary prototype initialisation is further investigated for learning rule LVQ1, as shown in Figure 80. Images that are manually labelled with 1 rating point are mainly assigned by the neural net to rating class 1. Likewise, the neural net can mainly assign images correctly, which are labelled with 3, 4 or 5 rating points. In contrast, the classifier assigns images labelled manually with 2 rating points to all 5 classes. Thus, the classifier provides the worst prediction accuracy for class 2.

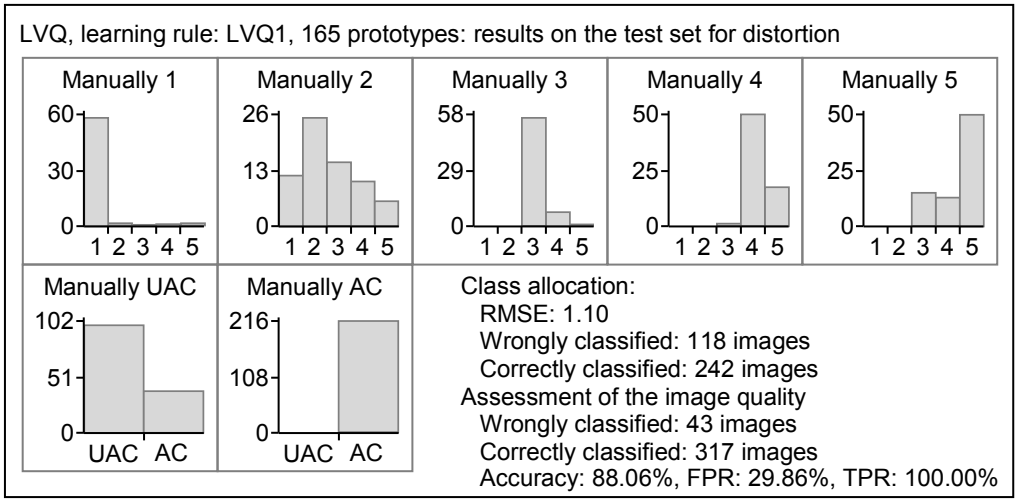


Figure 80: SL, LVQ for the distortion dataset: possible labelling of the test images

Based on the given classification task, the neural net archives the accuracy of 88.06%. On the positive side, this algorithm reaches the TPR of 100%. This ideal percentage value indicates that all test images with an acceptable image quality are classified customer suitable. However, the neural net achieves the FPR of 29.86%. This value indicates that 29.86% of the test images with an unacceptable quality receive a positive label. Thus, vehicles that are not customer suitable could reach the customers because the classifier confirms an acceptable image quality. In comparison to the results of the limit analysis, the FPR reached here is significantly above the low value of 2.76%. In contrast, the ideal TPR of 100% is not reached by the limit analysis.

At the end of the investigation, the classification problem is reduced to a 2-class problem. In total, 165 prototype vectors that are determined by the learning rule LVQ1 form the basis for this investigation. Now, a test image is classified as AC only if at least *m* of the *k*-nearest prototype vectors represent a customer suitable image quality. On the other hand, the test image is rejected and classified as UAC. The number of considered nearest prototype vectors is set to $k = 30$, $k = 20$, and $k = 10$. The number of nearest prototype vectors *m* required to classify a test image as AC is varied between 1 and k. Finally, the resulting ROC curves are determined. The curve, which results in the maximum AUC value, is shown in Figure 81.
The maximum AUC value of 0.86 is reached if the number of nearest prototype vectors k is set to 10. Even if k is increased, the area under the resulting ROC curve cannot be increased. The ROC curve shows that even at low FPR values TPR rates higher than 60% are reached. Although this classifier type reaches high TPR values, the FPR values fall not below the value of the limit analysis.

The resulting maximum values of the area under the curve are shown. The classification ability of the classifier is better if the AUC value is higher.

	k = 10	k = 20	k = 30
AUC	**0.86**	0.85	0.82
no. of prototypes	**165**	165	165

ROC-curve: this diagram illustrates the compromise between the matches (correctly positive classifications) and the costs (false positive classifications).

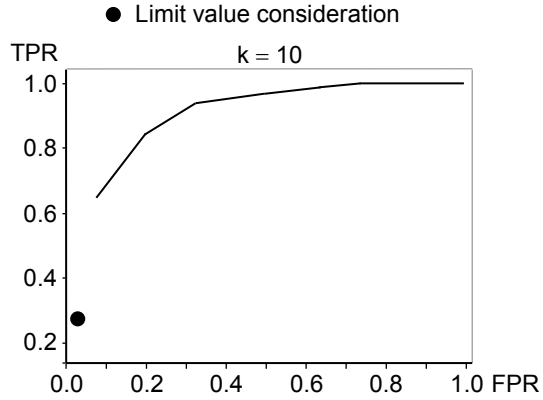


Figure 81: SL, LVQ for the distortion dataset: ROC curve for a 2-class classifier

Summary of the obtained results:

The polynomial classifiers tend to label many test images as customer suitable. Thus, the classifiers reach very high TPR values, which are clearly above the TPR of the limit analysis (limit value analysis: accuracy = 56.11%, TPR = 28.37%, FPR = 2.76%). Since many test images receive a positive label, the FPR values of the classifiers are also high. For example, the 2nd order PC for 10 principal components achieves the accuracy of 83.89%, the FPR of 39.31%, and the TPR of 99.53%. In contrast, when the PC's assignment rule is reduced to a 2-class problem, it is possible to implement classifiers that receive a low FPR value. These classifiers tend to label the test images as unacceptable. The maximum area under the ROC curve of 0.87 is reached by the 2nd order classifier in the 7-dimensional component space. For example, the 4th order classifier in the 3-dimensional component space reaches the accuracy of 67.78%, the FPR of 1.38%, and the TPR of 46.98%.

Like the 2-class PC, the kNN classifiers achieve low FPR values. These classifiers label many test images as not customer suitable, and therefore low TPR and FPR values are achieved. For example, the NN classifier for k = 1 and a dimension of 3 principal components reaches the accuracy of 57.22%, the FPR of 6.21%, and the TPR of 32.56%. If the mapping rule is reduced to a 2-class problem, the maximum achievable area under the ROC curve is in the range of 0.86 and 0.88. The results of the 2-class kNN algorithm are similar to the results of the limit analysis.

The results of the LVQ depend on the number of chosen prototypes and their random initialisation. Here, the ideal TPR of 100% can be achieved by 165 prototypes, which are trained by the learning rule LVQ1. This classifier obtains the accuracy of 88.06% and the FPR of 29.86%. This indicates that some test images wrongly receive a positive label. The LVQ achieves similar results when the classification problem is reduced

to a 2-class problem. The ROC diagram shows that even at low FPR values, TPR rates of more than 60% are achieved. This classifier type reaches high TPR values, but the FPR values do not fall below the value of the limit analysis
The accuracy value of the limit analysis is exceeded by all classifier types. This suggests that these classifiers are better suited to assess the combination of different types of distortions. Overall, it is found that the choice of the appropriate assessment algorithm is always a compromise between a high TPR and a low FPR.

7.4.2 SL: assessment algorithms for the double image dataset

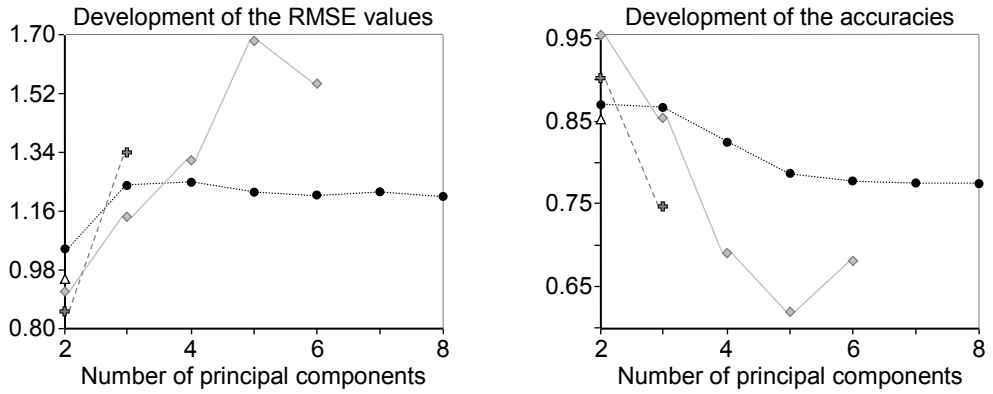
In order to implement an assessment algorithm for double images, 345 labelled training images and 360 test images are available. The image quality is uniquely described by 8 objective features. The classification results are compared with the results of the limit value analysis. The limit value analysis reaches the accuracy of 89.44%, the TPR of 87.79% and the FPR of 8.16%, see chapter *7.3*.

Polynomial classification:
The number of training images should be 10 times greater than the number of polynomial terms [MARSLAND 11]. By using 345 training images, it is possible to implement 1st order polynomial classifiers up to all 8 feature dimensions. The 2nd order classifiers can only be implemented up to 6 principal components. For the 3rd order polynomials, only classifiers up to 3 components can be implemented. The 4th order classifier can only be analysed up to 2 feature dimensions.

First, a test image is assigned to the class with the highest probability. The resulting RMSE values and the classification accuracies for various feature dimensions and polynomial orders are shown in Figure 82. The different dimensions of the feature space are displayed on the x-axis. The y-axis displays the resulting RMSE values and the classification accuracy respectively.
The left diagram in Figure 82 shows the development of the RMSE values. It can be seen that the RMSE values are lowest for all polynomial orders in the 2-dimensional component space. By increasing the number of principal components, the RMSE values initially increase significantly. The RMSE values for 1st order polynomials are well below the values of the of 2nd and 3rd order polynomials. The RMSE values of 2nd order polynomials increase significantly higher than the RMSE values for the other polynomials. Minimum achievable RMSE values are summarised in the in the table below the diagrams in Figure 82. As mentioned earlier, the RMSE values are lowest for all implemented polynomial classifiers in the 2-dimensional feature space. Since the 8 objective features are highly correlated, 84.90% of the total variance is covered by the 2 components. The 3rd order PC provides the lowest value with the RMSE of 0.85. In contrast, the 1st order classifier gives the highest RMSE value of 1.04. The values of 2nd (RMSE = 0.91) and 4th (RMSE = 0.94) order classifiers are in between.

The right diagram in Figure 82 shows the development of the classification accuracies. The development of the accuracy values is inverse to the development of the RMSE values. As the number of principal components increases, the classification accuracies decrease first and then increase slightly. For all polynomial orders, the 2-dimensional feature space results in the maximal accuracy values. These values are summarised in the lower table of Figure 82. The 2nd order classifier achieves the highest accuracy of 95.56%. It follows by the 3rd order PC with an accuracy of 90.00%. These 2 classifiers achieve a higher accuracy than the limit value consideration (89.44%). The 1st and the 4th order PCs only achieve an accuracy of 86.94% and 85.00%, respectively. These values are below the accuracy of the limit value analysis.

Development of the RMSE values and the accuracies for different numbers of principal components



Only the meaningful RMSE values and the accuracy values are shown which result from classifiers for which the number of training samples is 10 times greater than the number of polynomial terms, according to [MARSLAND 11]. Own illustration.

Resulting minimal RMSE values for the different classifier types

	1st order	2nd order	3rd order	4th order
RMSE	1.04	0.91	**0.85**	0.94
no. of principal components	2	2	**2**	2

Resulting maximal accuracy values for the different classifier types

	1st order	2nd order	3rd order	4th order
Accuracy	86.94%	**95.56%**	90.00%	85.00%
no. of principal components	2	**2**	2	2

Figure 82: SL, PC for the double image dataset: resulting RMSE and accuracy values

The labelling of the test images giving the lowest RMSE value is examined more closely. For this, the classification results of the 3rd order PC in the 2-dimensional feature space are summarised in Figure 83. It can be seen that test images that are manually labelled with 1 rating point are assigned to the appropriate rating class with almost no errors. Images that are manually labelled with 2 rating points are classified primarily in class 2 and class 1. The wrong allocation to rating class 1 is uncritical because this class also stands for an unacceptable image quality. On the contrary, the misclassifications of images labelled by the test persons with 3 rating points are more problematic.

Some of these images are assigned to rating class 2 which, unlike to class 3, represents an unacceptable image quality. The remaining images, which are manually labelled with 3 points, are assigned by the PC to rating classes 3, 4 and 5. Images that are manually labelled with 4 or 5 rating points are mostly assigned by the classifier in classes 4 and 5. These misallocations are non-critical, as classes 3, 4 and 5 represent an acceptable image quality. The classifier has a low FPR of 7.48%. This means that 7.48% of unacceptable labelled test images receive an acceptable label from the classifier. 88.26% of all test images showing an acceptable quality are also classified as acceptable by the PC. This classifier reaches a lower FPR value and a higher TPR value than the limit consideration analysis (TPR = 87.79%, FPR = 8.16%). This is also reflected in the higher classification accuracy. The 2nd order classifier in the 2-dimensional feature space is more cost-effective. The classifier would send fewer vehicles with an acceptable quality to rework. Likewise, fewer vehicles with an unacceptable image quality would reach the customers.

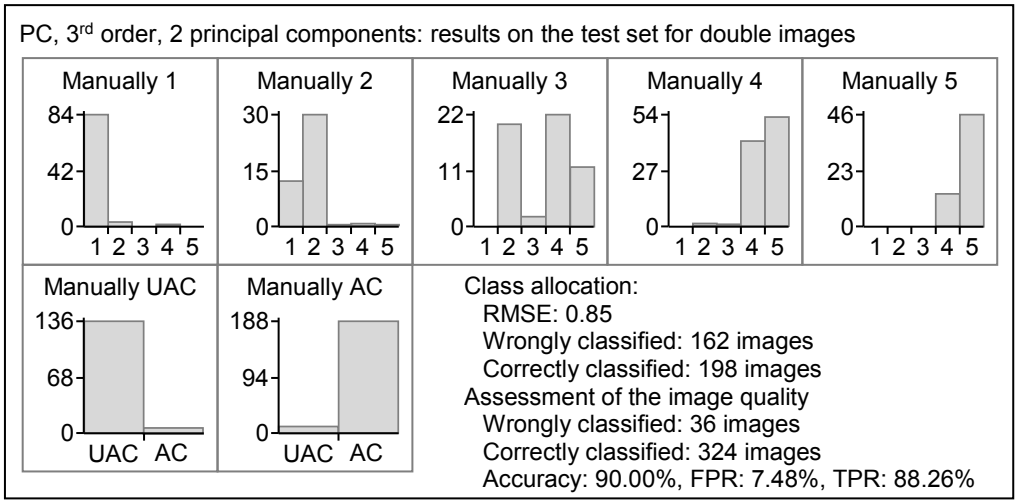


Figure 83: SL, PC for the double image dataset: possible labelling of the test images

In addition, the classification result with the highest accuracy is examined in more detail. The labelling of the test images is shown in Figure 84. The classifier assigns images, which are manually labelled with 1 rating point, to rating classes 1 and 2. Likewise, images that are manually labelled with 4 or 5 rating points are assigned to classes 3, 4 and 5. These misclassifications are uncritical, as these rating classes represent the customer suitable image quality. In contrast, test images that are manually perceived as unacceptable and therefore labelled with 2 rating points are assigned by the classifier to all 5 rating classes. Similarly, the classifier assigns images that are manually labelled with 3 points to rating classes 2, 3, 4 and 5. Here, problems are caused by test images that are close to the boundary between an acceptable and an unacceptable quality. Especially for these images, it is difficult to assign them correctly. Based on the given classification task, the classifier achieves the TPR of 98.59%. This high percentage indicates that almost all test images with an acceptable quality receive a positive label. The TPR value of the classifier is also 10.80% higher than the TPR value of the

limit value consideration. Thus, the classifier would cause fewer rework costs. In contrast, the FPR value is 0.68% higher than the FPR of the limit consideration. Here, insignificant more vehicles with an unacceptable quality would reach the customers.

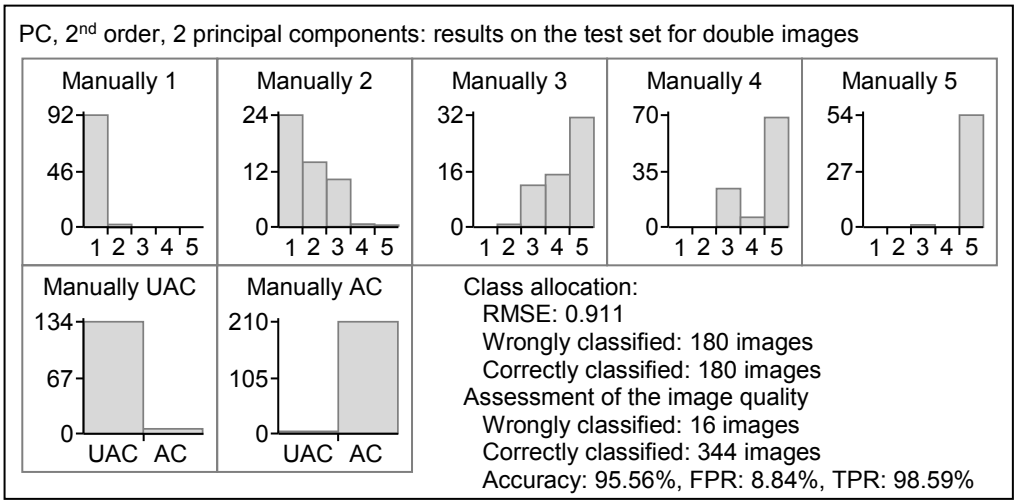


Figure 84: SL, PC for the double image dataset: possible labelling of the test images II

The next step is to investigate which classifier type achieves the lowest false positive rate and the highest true positive rate for the given test images. The values, resulting from classifiers of different polynomial orders, are compared up to the maximal possible feature dimensions, as shown in Table 28.

PC	no. of principal components	FPR	TPR
	Minimal FPR values		
1st order	6	31.29%	84.04%
2nd order	2	8.84%	98.59%
3rd order	2	7.48%	88.26%
4th order	2	14.29%	84.51%
	Maximal TPR values		
1st order	2	31.97%	100.00%
2nd order	2	8.84%	98.59%
3rd order	2	7.48%	88.26%
4th order	2	14.29%	84.51%

Table 28: SL, PC for the double image dataset: min FPR and max TPR

The lowest FPR of 7.48% is reached by the PC of 3rd order in the 2-dimensional component space. This classifier type obtains the TPR of 88.26%. Both values are better than the values of the limit consideration. The 1st order classifier achieves the maximum possible TPR of 100% in the 2-dimensional feature space. Thus, there would be no additional costs because no vehicles with an acceptable image quality would be mistakenly sent to rework. Unfortunately, the classifier shows the highest FPR of

31.97%. This value is far above the value of the limit value consideration (8.16%). The resulting values show that choosing the most appropriate classification method is always a compromise between a low FPR and a high TPR.

Finally, the allocation task of the classifier is limited to 2 possibilities, namely acceptable or unacceptable. The test image receives a positive label if the probability of belonging to an acceptable quality is higher by a multiple ϕ than the probability of belonging to an unacceptable quality [KRISTIAN et al. 11: p. 203]. The constant ϕ is varied between 0.1 and 10. The increase is done in steps of 0.1. The analysis is performed for classifiers of the various polynomial orders up to the maximum possible feature dimensions. For each classifier, the resulting ROC curve is determined and the area under the curve is calculated. The curves resulting in the maximum AUC value are shown in Figure 85.

The resulting maximum values of the area under the curve are shown. The classification ability of the classifier is better if the AUC value is higher.

	1st order	2nd order	3rd order	4th order
AUC	0.996	**0.998**	0.981	0.964
no. of principal components	2	**2**	2	2

ROC-curves: these diagrams illustrate the compromise between the matches (correctly positive classifications) and the costs (false positive classifications).

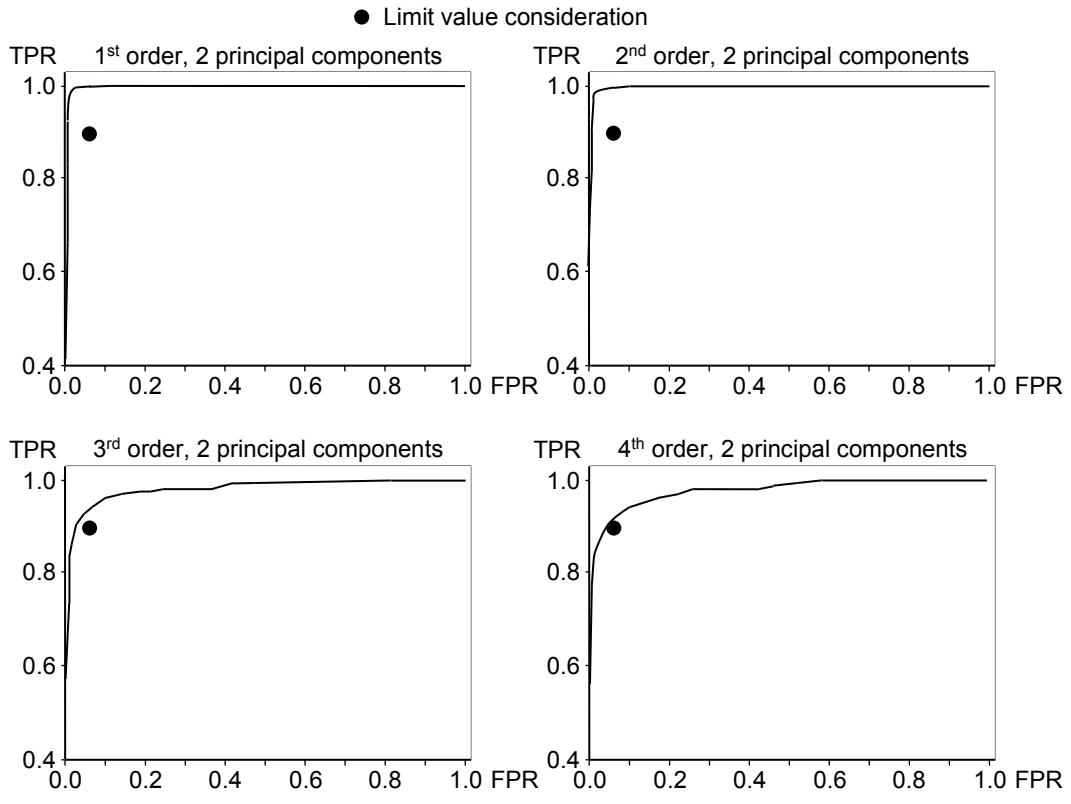


Figure 85: SL, PC for the double image dataset: ROC curves for 2-class classifiers

As the classifier works more accurately the closer the curve approaches the upper left corner of the diagram, the resulting ROC curves are nearly perfect. The ROC curves initially rise almost vertically. The TPR is close to 100%, while the FPR remains close to 0%. Thereafter, the FPR values increases. The highest AUC value of 0.998 is reached by the 2nd order classifier in the 2-dimensional feature space. Then immediately follows the AUC value of 0.996, which is achieved by the 1st order classifier in the 2-dimensional feature space. Both AUC values barely miss the ideal value of 1. In addition, 3rd and 4th order classifiers show nearly perfect ROC curves. These classifiers achieve AUC values of 0.981 and 0.964.

A possible classification result is shown in Figure 86. For this, the 2nd order classifier in the 2-dimensional feature space with a constant ϕ of 0.6 is applied to the test images.

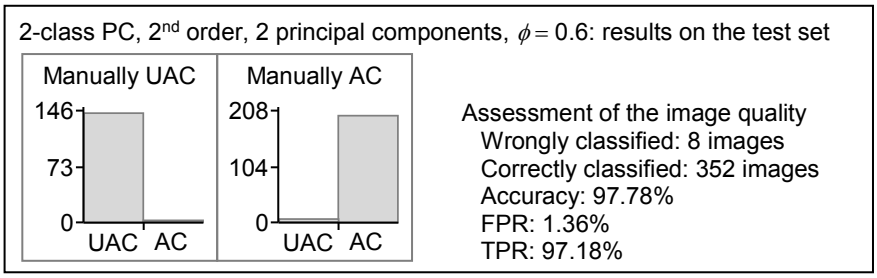


Figure 86: SL, 2-class PC for the double image dataset: possible labelling of the test images

The evaluation of the given test data achieves the accuracy of 97.18%. The FPR is 1.36%. Thus, only 1.36% of the test images that are manually labelled as unacceptable receive a positive label. An equally good value shows the TPR. Here, 97.18% of all images showing an acceptable quality are also classified as acceptable by the classifier. Thus, this classifier works more accurately than the limit value consideration.

Nearest neighbour classification:

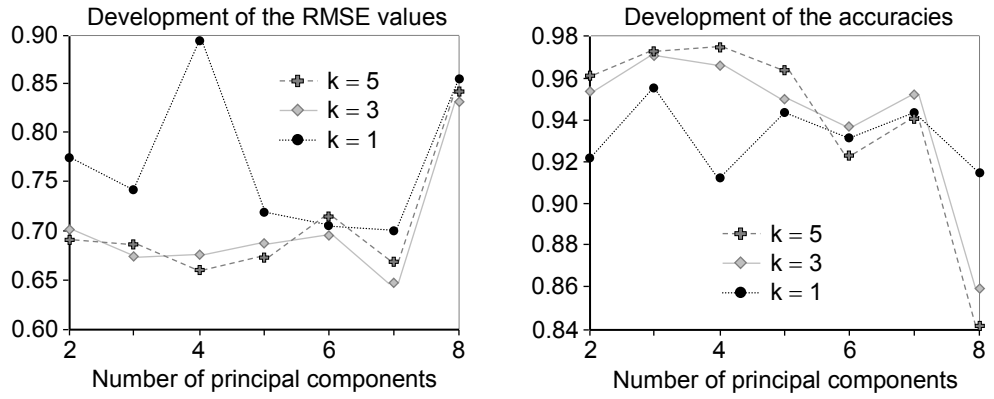
The majority of the *k*-nearest neighbours are classified either as AC or as UAC. Regardless, the test image is assigned to the class of the next training image from that majority. The investigation is carried out for $k = 1$, $k = 3$ and $k = 5$.

Figure 87 shows the development of the RMSE values and the development of the classification accuracies for different feature dimensions. The x-axis shows the number of components and the y-axis the RMSE values and the accuracy values, respectively. It can be seen that the development of the RMSE values is very similar for $k = 3$ and $k = 5$. By increasing from 2 to 6 principal components, the RMSE values hardly change. Thereafter, the RMSE values decrease slightly for 7 principal components. In the end, the RMSE values increase for 8 principal components. The development for $k = 1$ looks different. First, the RMSE values decrease. The peculiarity is that the RMSE values increase significantly for 4 main components and decrease almost equally for 5 principal components. Overall, the RMSE values for $k = 1$ are higher (except for the RMSE value for 6 principal components) than the values for the other classifier types. The first

table in Figure 87 summarises the minimum RMSE values for the kNN classification. The lowest RMSE value of 0.65 reaches the classifier for k = 3 in the 7-dimensional feature space. The second best RMSE value of 0.66 is achieved by considering 5 neighbours and 4 principal components. The worst result of 0.70 is achieved by the NN classifier for k = 1 and 7 feature dimensions.

For k = 3 and k = 5, the development of the classification accuracies is inverse to the development of the RMSE values. Initially, the values increase slightly and fall off again significantly. By increasing the number of principal components, the accuracy values for k = 1 follow a zigzag course. The maximal accuracy values are shown in the second table of Figure 87. The nearest neighbour classifier in the 4-dimensional component space achieves the highest accuracy of 97.50% for k = 5. Taking into account 3 principal components, the classifiers for k = 3 and k = 1 achieve accuracy values of 97.22% and 95.56%, respectively. The maximum accuracies are very high and are only slightly below the ideal value of 100%. These accuracy values also exceed the accuracy achieved by the limit analysis (89.44%).

Development of the RMSE values and the accuracies for different numbers of principal components



Resulting minimal RMSE values for the different classifier types

	k = 1	k = 3	k = 5
RMSE	0.70	**0.65**	0.66
no. of principal components	7	**7**	4

Resulting maximal accuracy values for the different classifier types

	k = 1	k = 3	k = 5
Accuracy	95.56%	97.22%	**97.50%**
no. of principal components	3	3	**4**

Figure 87: SL, NN classification for double images, development of the results

The class assignment resulting in the lowest RMSE value is analysed more closely. The results are shown in Figure 88. The allocation of images that are manually labelled with 1 or 5 rating points is uncritical. These images are assigned by the classifier to the classes 1 and 2 or 4 and 5, respectively. Images labelled with 2 rating points by the test persons are mainly assigned to this class. In contrast, images subjectively belong to rating class 3 are assigned to classes 2, 3, 4 and 5. Here, the assignment to class 2 is

problematic because images that are manually labelled as acceptable receive an unacceptable label. Images that are manually labelled with 4 rating points are assigned primarily to rating classes that represent an acceptable image quality. Based on the given classification task, the classifier reaches the accuracy of 95.28%, the FPR of 0.68% and the TPR of 92.49%. Thus, the values are better than the values of the limit value consideration (accuracy = 89.44%, FPR = 8.16%, TPR =87.79%). Since 92.49% of all vehicles with an acceptable image quality are recognized as such, only 7.51% of the vehicles would be mistakenly sent to rework. Likewise, only 0.68% of the vehicles with an unacceptable image quality would reach the customers.

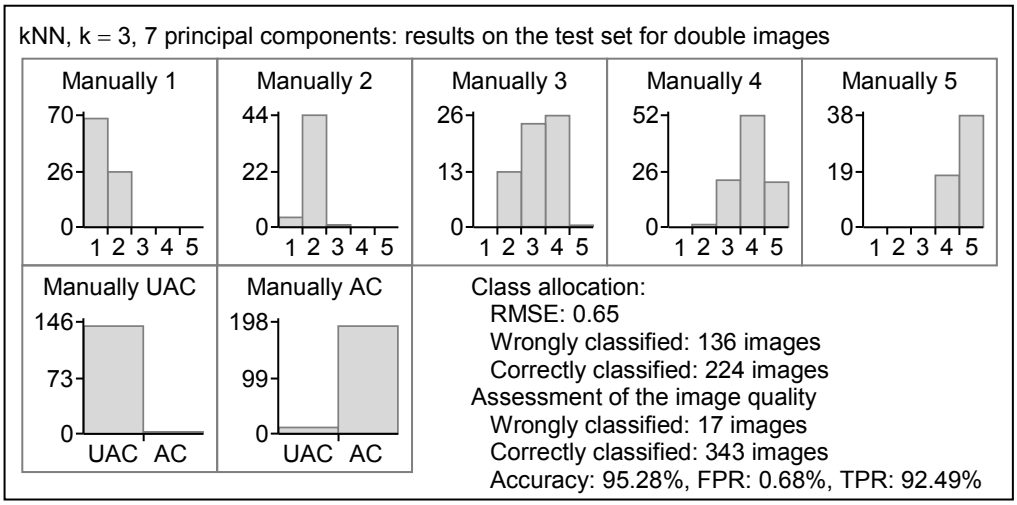


Figure 88: SL, kNN for the double image dataset: possible labelling of the test images

In addition, the classification result with the highest accuracy is further investigated, as shown in Figure 89. Test images that are manually labelled with 1 or 5 rating points are assigned to rating classes 1 and 2 or 3, 4 and 5, respectively. The quality of these images is thus assessed correctly. Likewise, the quality of images that are manually labelled with 2 or 4 rating points is mainly correctly assessed. Difficult is the assessment of test images, which are manually labelled with 3 rating points. These images are mostly assigned to rating classes 3 and 4, this is known to be unproblematic, but also to rating class 2, which causes problems. On the positive side, the classifier achieves the accuracy of 97.50%, the FPR of 0.68% and the TPR of 96.24%. For the given classification problem the classifier is thus better suited than the limit value consideration (accuracy = 89.44%, FPR = 8.16%, TPR = 87.79%). 96.24% of the test images, which are manually labelled as acceptable, also receive a positive label from the classifier. Thus, less than 4% of the vehicles would be mistakenly sent to rework. The percentage of non-customisable images that are falsely labelled as acceptable is only 0.68%. Here, less than 1% of vehicles with an unacceptable quality would reach the customers.

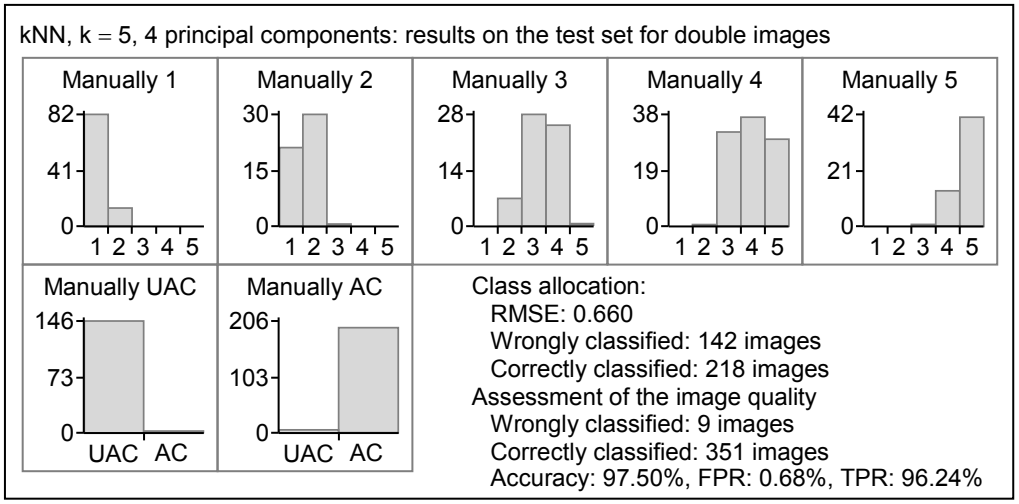


Figure 89: SL, kNN for the double image dataset: possible labelling of the test images II

This section examines, which classifier type achieves the lowest FPR and the highest TPR. The values obtained based on the given test images are shown in Table 29. Looking more closely at this table, the resulting FPR values are below and the TPR values are above the values of the limit consideration. Thus, the NN classifiers appear to be more suitable for the given assessment task than the standard method. The lowest FPR of 0.68% and the highest TPR of 96.24% are achieved by the classifier in the 4-dimensional feature space where 5 nearest neighbours are taken into account. This low FPR is also achieved by classifiers for $k = 1$ and $k = 3$ in the 6- or 3-dimensional feature space. However, the maximal TPR of 88.73% and 95.77%, respectively, is lower than the reference value.

NN	no. of principal components	FPR	TPR
	Minimal FPR values		
$k = 1$	6	0.68%	88.73%
$k = 3$	3	0.68%	95.77%
$k = 5$	4	0.68%	96.24%
	Maximal TPR values		
$k = 1$	3	5.44%	96.24%
$k = 3$	3	0.68%	95.77%
$k = 5$	4	0.68%	96.24%

Table 29: SL, kNN for the double image dataset: min FPR and max TPR

In the final step, the classification algorithm is modified to a 2-class problem. A test image is assigned as AC only if at least *m* of the *k*-nearest training images are classified as AC. Otherwise, the test image is rejected. The number of considered nearest training images is set to $k = 30$, $k = 20$, and $k = 10$. At the same time, the number of nearest training images *m* representing the acceptable image quality is varied between 1 and k. Each classifier is applied to the test images and the resulting ROC curve is

determined. The investigation is carried out for all feature dimensions. The curve giving the maximum AUC value is shown in Figure 90.

The resulting maximum values of the area under the curve are shown. The classification ability of the classifier is better if the AUC value is higher.

	k = 10	k = 20	k = 30
AUC	0.88	0.94	**0.96**
no. of principal components	8	8	**7**

ROC-curve: this diagram illustrates the compromise between the matches (correctly positive classifications) and the costs (false positive classifications).

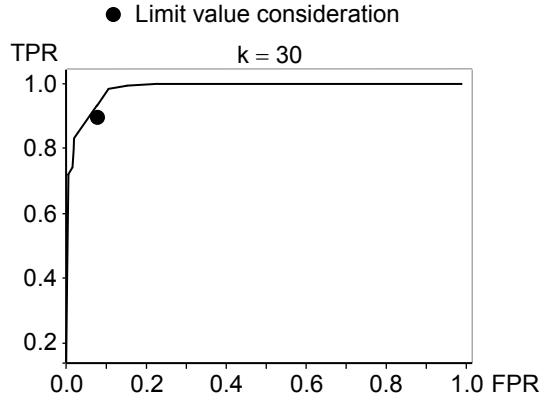


Figure 90: SL, kNN for the double image dataset: ROC curve for a 2-class classifier

The maximum AUC value of 0.96 is achieved in the 7-dimensional component space when the number of nearest prototype vectors k is set to 30. Again, the resulting ROC curve is nearly identical to the curve of an ideal classifier. The resulting AUC value is almost equally to 1. A possible classification result is shown in Figure 91. This result is again slightly better than the result of the limit analysis.

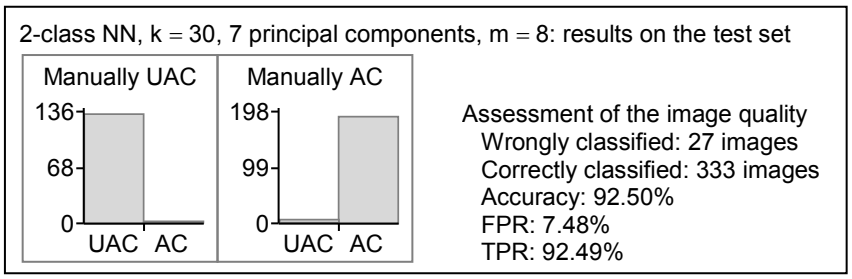


Figure 91: SL, 2-class kNN for the double image dataset: possible labelling of the test images

The evaluation of the given test data achieves the accuracy of 92.50%. The FPR is 7.48%. The classifier would thus only send 7.48% of the vehicles with an unacceptable image quality to the customers. An equally good value shows the TPR. Here, 92.49% of all acceptable images are also classified as acceptable. Therefore, the classifier would cause almost no additional costs, since only 7.51% of the vehicles with an acceptable quality would be mistakenly sent to rework.

Learning vector quantisation:

The prototypes are determined by the learning rules LVQ1, LVQ2.1 and LVQ3. A learning rate α of 0.03, a window width *w* of 0.3, and an adjustable learning rate ε of 0.1 are used. A test image is assigned to the same class as the nearest prototype vector belongs. Based on the available test dataset, it is checked which combination of learning rule and prototype number achieves the best result for the given classification problem. First, the RMSE values and the classification accuracies are determined. The results are summarised in Figure 92.

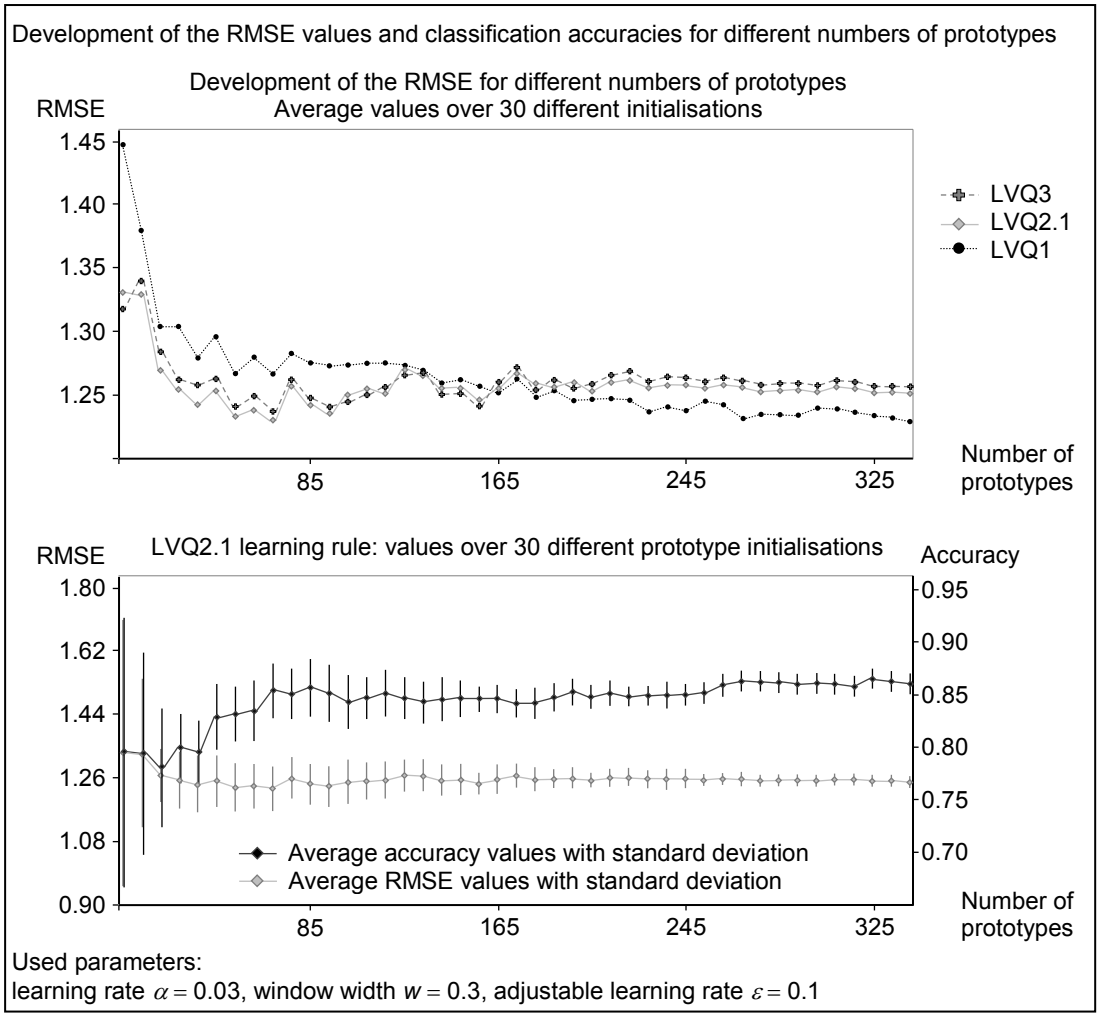


Figure 92: SL, LVQ for the double image dataset: evaluate different numbers of prototypes

The average values and the standard deviations for 30 different prototypes initialisations are shown. The number of prototypes is increased at intervals of 8. For each learning rule, the average RMSE values are shown in the upper part of the figure. The lower diagram shows in addition to the RMSE values the classification accuracies that are determined by the learning rule LVQ2.1. The RMSE values resulting from the classifier, which is trained by learning rule LVQ1, first decrease strongly. However, up to 165 prototypes, the RMSE values are significantly higher than the values of the classifiers, which are determined by learning rules LVQ2.1 or LVQ3. For more than 165 pro-

totypes, the RMSE values of the classifier trained according to LVQ1 are below the other corresponding values. The development of the RMSE values of the classifiers trained by learning rule LVQ2.1 and LVQ3 are very similar to each other. As the number of prototypes is increased, the RMSE values decrease first and then remain nearly constant. For a small number of prototypes, the lowest RMSE values are achieved by classifiers trained by learning rule LVQ2.1.

For learning rule LVQ2.1, the development of the RMSE values and the classification accuracies generally shows an opposite trend. As the number of prototypes increases, the recognition accuracy of the classifier slowly increases. The standard deviation of the accuracy is highest for a small number of prototypes. For the investigated prototype numbers, the classifier achieves averaged recognition accuracies of just 85%. Thus, the classifier does not achieve the recognition accuracy of the limit value consideration (89.44%).

In the next step, the FPR and the TPR for classifiers determined by the learning rule LVQ2.1 are analysed for different numbers of prototypes. Again, averages and standard deviations are calculated for 30 different prototype initialisations. The results are summarised in Figure 93. The diagram shows that the TPRs are increasing only slowly for an increasing number of prototypes. The values rise from 87.28% to 95.45%. For 69 prototypes, on average, a higher TPR is already determined than for the limit value analysis. The average FPR values fall from 62.67% to 43.33% for increasing the number of prototypes. Here, the FPR is significantly higher than the FPR achieved by the limit value analysis.

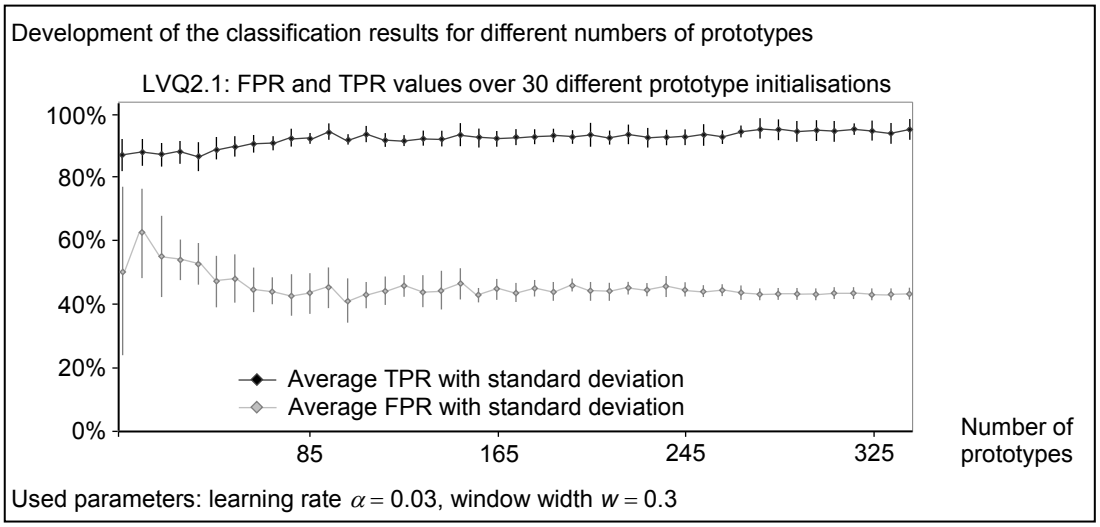


Figure 93: SL, LVQ for the double image dataset: resulting values for learning rule LVQ2.1

Now the number of prototypes is set to 69. The prototype number is in the middle of the bending of the RMSE curve, shown in Figure 92. Thus, the best relationship is expected between the number of prototypes and the RMSE value. A possible classification result is shown in Figure 94. For this, a classifier is used, which is trained by learning rule LVQ2.1.

The first histogram shows that test images that are manually labelled with 1 rating point are assigned to rating class 1 and 4. Test images that are manually labelled with 2 points are assigned mainly to rating classes 1, 2, and 4. Class 4 assignments are problematic because, unlike rating classes 1 and 2, this class represents an acceptable image quality. This explains the high FPR of 33.33%. The FPR is much higher than the value of the limit analysis (8.16%). This indicates that $^1/_3$ of all images that are manually rated as unacceptable get a positive label from the classifier. Thus, the classifier would send 25.17% more vehicles of unacceptable quality to the customers than the limit value consideration. The labelling of test images that are manually labelled with 3 rating points is the hardest. These images are assigned to rating classes 2, 3, 4 and 5. Test images that are manually labelled with 4 rating points are sorted mainly into this class, but also into the classes 2, 3, and 5. Only images that are manually labelled with 5 points are classified into classes 4 and 5 and are therefore considered customisable. The classifier achieves a TPR of 94.37%. Here, 94.37% of customer-friendly images are also recognised as such. Thus, the TPR is much higher than the value of the limit value consideration (87.79%) and is very close to the ideal value of 100%. This classifier would be less costly than the limit analysis because fewer vehicles of acceptable quality would be mistakenly sent to rework. Overall, the classifier achieves a recognition accuracy of 83.06%, which is below the accuracy (89.44%) of the limit analysis.

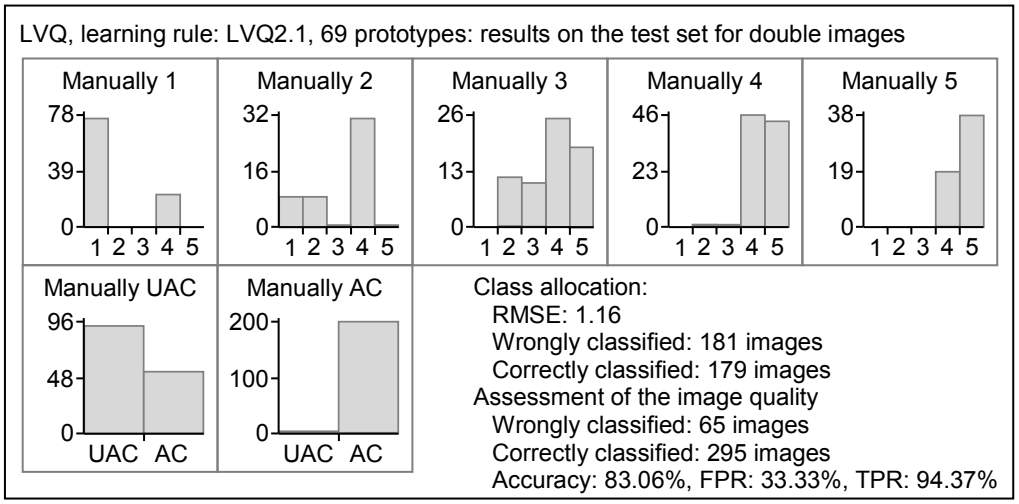


Figure 94: SL, LVQ for the double image dataset: possible labelling of the test images

Finally, the LVQ mapping rule is reduced to a 2-class problem. Here, the image quality is classified only as AC or UAC. A test image is considered acceptable if at least *m* of the *k*-nearest prototype vectors represent a customer-friendly image quality. 69 prototype vectors are used, which are determined by the learning rule LVQ2.1. Again, the ROC curve of the classifier is determined by varying the number of relevant prototype vectors *m*. The number of considered nearest prototype vectors is set to k = 30, k = 20, and k = 10. The ROC curve leading to the maximum possible AUC value is shown in Figure 95.

The resulting maximum values of the area under the curve are shown. The classification ability of the classifier is better if the AUC value is higher.

	k = 10	k = 20	k = 30
AUC	0.70	0.97	**0.98**
no. of prototypes	69	69	**69**

ROC-curve: this diagram illustrates the compromise between the matches (correctly positive classifications) and the costs (false positive classifications).

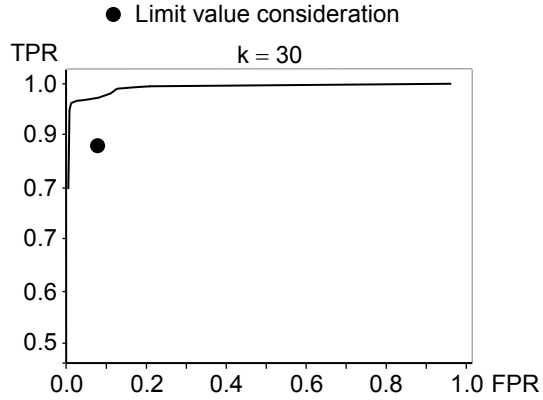


Figure 95: SL, LVQ for the double image dataset: ROC curve for a 2-class classifier

The maximum AUC value of 0.98 is reached when the number of considered nearest prototype vectors k is set to 30. Again, the ROC curve is almost perfect. The TPR values are close to 100%, while the FPR values remain close to 0%. Thereafter, the FPR values increase. If k is set to 20, a good result is also achieved. If k is further reduced, the area under the ROC curve is sustainably reduced. A possible classification result for k = 30 and m = 5 is shown in Figure 96.

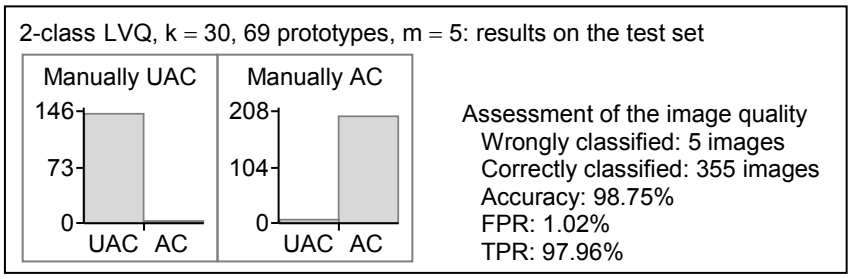


Figure 96: SL, 2-class LVQ for the double image dataset: possible labelling of the test images

With this classifier, an accuracy of 98.75% can be achieved. The FPR is 1.02%. Thus, only 1.02% of the test images that are manually labelled as unacceptable receive a positive label. Likewise, the TPR has a high value. Here, 97.96% of all images showing an acceptable quality are also considered as acceptable by the classifier. Thus, this classifier type works more accurately than the limit analysis (accuracy = 89.44%, FPR = 8.16%, TPR = 87.79%).

Summary of the obtained results:
Since the limit analysis achieves a good classification result, it is difficult to implement a classification algorithm that achieves a better result. Polynomial classifiers can be implemented that achieve a higher TPR values than the limit analysis. Since these classifiers tend to label the test images as customer suitable, the FPR values are usually higher than the reference value. Thus, more vehicles with an unacceptable image quality would reach the customers. For example, the 2nd order PC in the 2-dimensional component space achieves the accuracy of 95.56%, the FPR of 8.84%, and the TPR of 98.59%. Only if the classification problem is limited to 2 possibilities, classifiers can be implemented that achieve better results than the limit analysis. The resulting ROC curves of the 2-class classifier are nearly perfect. Thus, 2nd order classifier achieves the accuracy of 97.78%, the FPR of 1.36%, and the TPR of 97.18%.
The kNN classifiers reach FPR values below and TPR values above the values of the limit analysis. Thus, the kNN classifiers seem to be better suited than the standard method. For example, the kNN classifier for $k = 3$ and a dimension of 7 principal components achieves an accuracy the 95.28%, the FPR of 0.68%, and the TPR of 92.49%. Also as a 2-class classifier, the NN algorithm achieves a nearly perfect ROC curve and the classification results are better than the values of the limit analysis.
If the number of prototypes is greater than 69, the LVQ may reach TPR values that are higher than the value of the limit analysis. Since these classifiers tend to label many test images as customer suitable, the FPR values obtained are also high. Thus, the classifiers would send more vehicles with an unacceptable image quality to the customers than the standard method. For example, if the learning rule LVQ2.1 is used to train 69 prototypes, the recognition accuracy of 83.06%, the FPR of 33.33%, and the TPR of 94.37% is achieved. Only the reduction to a 2-class problem, where the image quality is classified as AC or UAC, leads to better classification results. The 2-class LVQ algorithm can achieve better results than the limit value analysis. For example, the 2-class LVQ can achieve the accuracy of 98.75%, the FPR of 1.02%, and the TPR of 97.96%.

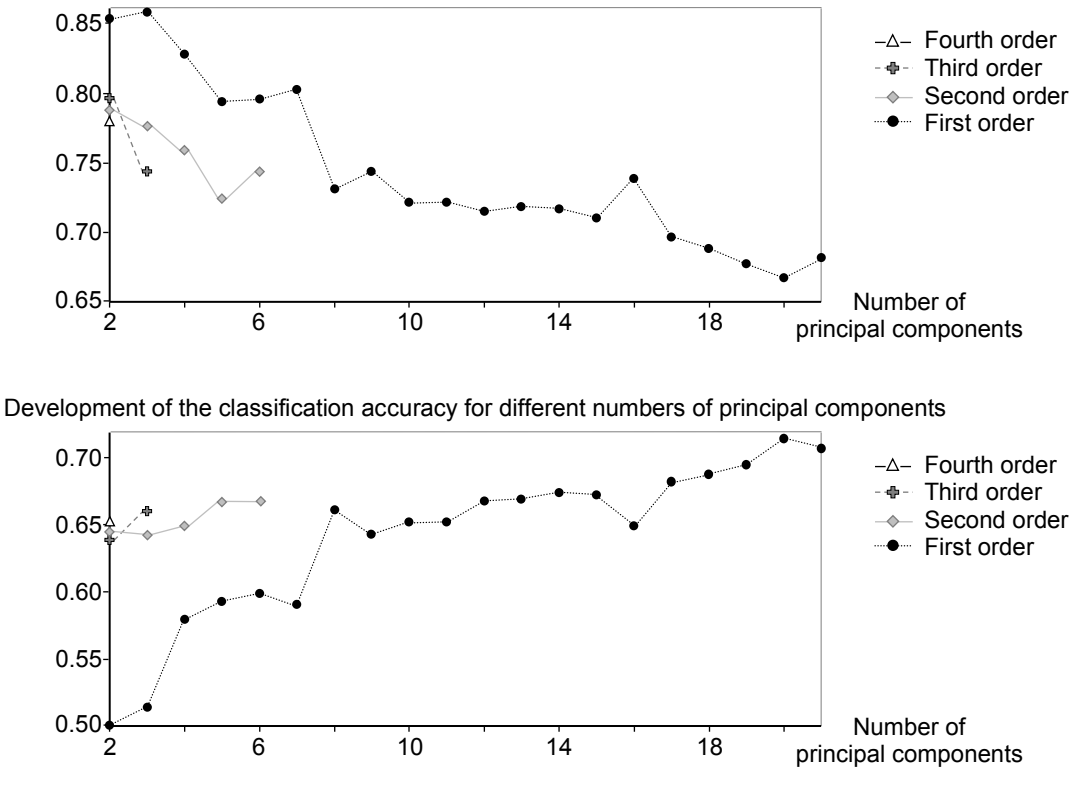
7.4.3 SL: assessment algorithms for the distortion and double image dataset

This chapter examines the perception of combinations of distortions and double images. The image quality is clearly described by 21 objective features. The available dataset consists of 303 training images and 305 test images. The classification results are compared with the results of the limit analysis. The limit value analysis reaches the accuracy of 67.21%, the TPR of 86.23% and the FPR of 55.80%, see chapter *7.3*.

Polynomial classification:
Due to a large number of objective criteria and the limited number of training images, it is first checked for which classifier dimensions the number of training images is 10 times greater than the number of polynomial terms [MARSLAND 11]. With 303 training

images, it is possible to implement 1st order classifiers up to all 21 feature dimensions. The 2nd order classifiers can be implemented up to 6 feature dimensions and the 3rd order classifiers are analysed up to 3 dimensions. The 4th order classifier can only be analysed up to the 2-dimensional feature space.

First, the test image is assigned to the class with the greatest probability. The resulting RMSE values and the accuracies for different polynomial orders and feature dimensions are shown in Figure 97.



Resulting minimal RMSE and maximal accuracy values for the different classifier types

	1st order	2nd order	3rd order	4th order
RMSE	**0.67**	0.72	0.73	0.78
Accuracy	**71.48%**	68.82%	66.23%	65.25%
no. of principal components	**20**	5	3	2

Figure 97: SL, PC for the distortion and double image dataset: resulting values

The x-axes show the different dimensions of the feature space and the y-axis the resulting RMSE values and the classification accuracies. The upper diagram shows the development of the RMSE values up to the maximal possible feature dimensions. While the number of principal components increases, the loss of information caused by the PCA decrease. Thereby the RMSE values also decrease. The RMSE values of 1st order classifiers are highest. The higher the polynomial orders, the lower the RMSE values of the classifiers. The lower diagram shows the development of the associated recognition accuracies. A reverse development compared to the RMSE values can be

seen. While the number of principal components is increased, the recognition accuracy increases as well.
The resulting minimum RMSE values and the maximum recognition accuracies for each polynomial order are summarised in the table of Figure 97. Based on the given classification task, the 4th order classifier in the 2-dimensional feature space achieves the RMSE value of 0.78 and the accuracy of 65.25%. Better results are obtained from the 3rd order classifier in the 3-dimensional feature space and the 2nd order classifier in the 5-dimensional feature space. These classifiers achieve RMSE values of 0.73 and 0.72, respectively as well as accuracies of 66.23% and 68.82% respectively. It is assumed that the 1st order classifier in the 20-dimensional feature space achieves the best possible result with the RMSE value of 0.67 and the accuracy of 71.48%. The recognition accuracies of the 2 most suitable classifiers are higher than the accuracy of the limit value consideration.

The assignment result of the 1st order classifier in the 20-dimensional component space is investigated more closely, as shown in Figure 98. Since the 21 objective features are hardly correlated with each other, a dimensionality reduction of the component space is only scarcely possible. In the 20-dimensional component space, the entire variance of the 21 features is 100% covered. This is only possible because the eigenvalue of the 21st principal component is exactly 0. For details, see chapter *7.2.3*.

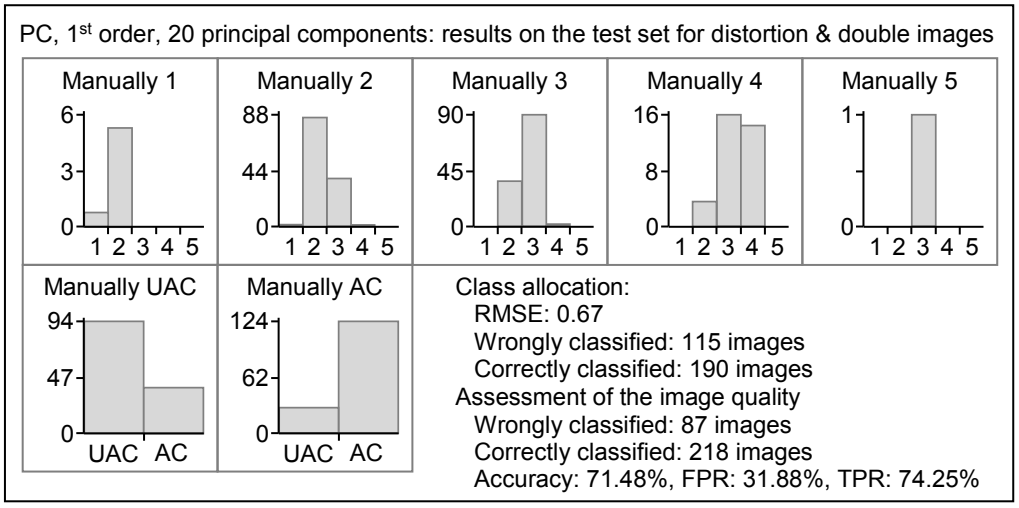


Figure 98: SL, PC for the distortion and double image dataset: possible result

The classifier assigns test images, which are manually labelled with 1 rating point, mainly to class 2. This misclassification is unproblematic since both rating classes represent an unacceptable image quality. Test images labelled with 2 rating points are assigned by the classifier into this class, but also to class 3. This misclassification classifies images of unacceptable quality as suitable for customers. Thus, the FPR is 31.88%. This FPR is 23.92% lower than the rate of the standard method. Thus, fewer non-custom vehicles would be wrongly sent to the customers. On the other hand, test images that are manually labelled with 3 or 4 rating points are assigned to rating classes 2, 3 and 4 by the classifier. Therefore, the classifier reaches the TPR of 74.25%.

This value is 11.98% lower than the TPR of the limit value consideration. The classifier would cause more costs than the standard method because more vehicles of acceptable quality would be mistakenly sent to rework. Finally, the test image labelled with 5 points is assigned to rating class 3.

The next step is to find classifiers that achieve the lowest FPR and the highest TPR for the given classification task, shown in Table 30. The values from classifiers of different polynomial orders are compared with each other up to the maximal possible feature dimensions. All classifier types listed in the table achieve lower FPR values, but also lower TPR values than the standard method. The lowest FPR and the highest TPR are reached by the 1st order classifier in the 20-dimensional component space. The second lowest FPR of 34.06% is achieved from the 2nd order classifier in the 6-dimensional component space. The maximum TPRs of 74.25%, 72.46%, 71.86% and 70.66% are achieved from classifiers of the 1st, 4th, 3rd and 2nd order in the 20, 2, 3 and 5-dimensional feature space.

PC	no. of principal components	FPR	TPR
	Minimal FPR values		
1st order	20	31.88%	74.25%
2nd order	6	34.06%	70.06%
3rd order	3	40.58%	71.86%
4th order	2	43.48%	72.46%
	Maximal TPR values		
1st order	20	31.88%	74.25%
2nd order	5	37.68%	70.66%
3rd order	3	40.58%	71.86%
4th order	2	43.48%	72.46%

Table 30: SL, PC for the distortion and double image dataset: min FPR and max TPR

Finally, the 5-class problem is reduced to a 2-class problem. Now the decision for an acceptable quality is only made if the probability belonging to an acceptable quality is higher by a multiple ϕ than the probability of belonging to an unacceptable quality [KRISTIAN et al. 11: p. 203]. The constant ϕ is varied between 0.1 and 10; the increase is done in steps of 0.1. The investigation is carried out for classifiers of different polynomial orders up to the maximum possible feature dimensions. For each classifier, the resulting ROC curve is determined and the area under the curve is calculated. The curves, which result in the maximum AUC value, are shown in Figure 99.

The obtained values for the area under the curve are very close to each other and vary between 0.63 and 0.69. The highest AUC of 0.69 is achieved by the 2-class PC of 4th order in the 2-dimensional component space. The resulting AUC values are therefore far away from the ideal value of 1. By varying the constant ϕ, classifiers can be imple-

mented that achieve either a higher TPR or a lower FPR than the limit value consideration. Thus, the choice of the most suitable classifier is a compromise between a high TPR and low FPR.

The resulting maximum values of the area under the curve are shown. The classification ability of the classifier is better if the AUC value is higher.

	1st order	2nd order	3rd order	4th order
AUC	0.68	0.68	0.63	**0.69**
no. of principal components	14	2	2	**2**

ROC-curves: these diagrams illustrate the compromise between the matches (correctly positive classifications) and the costs (false positive classifications).

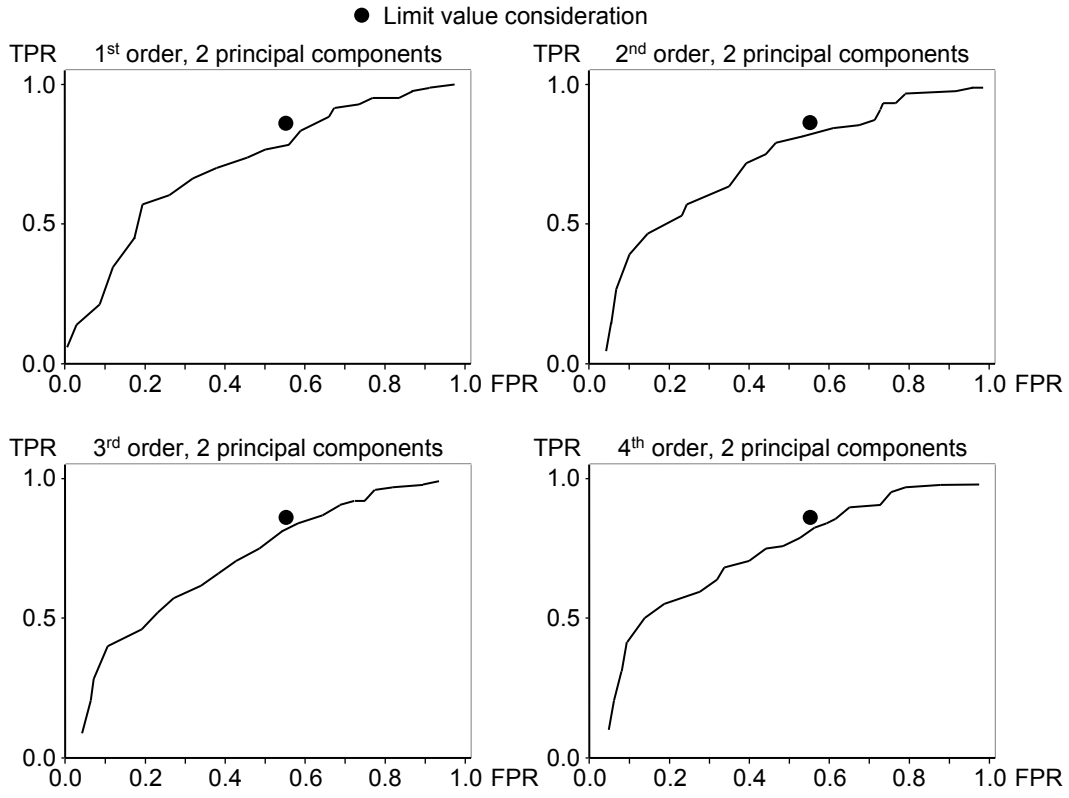
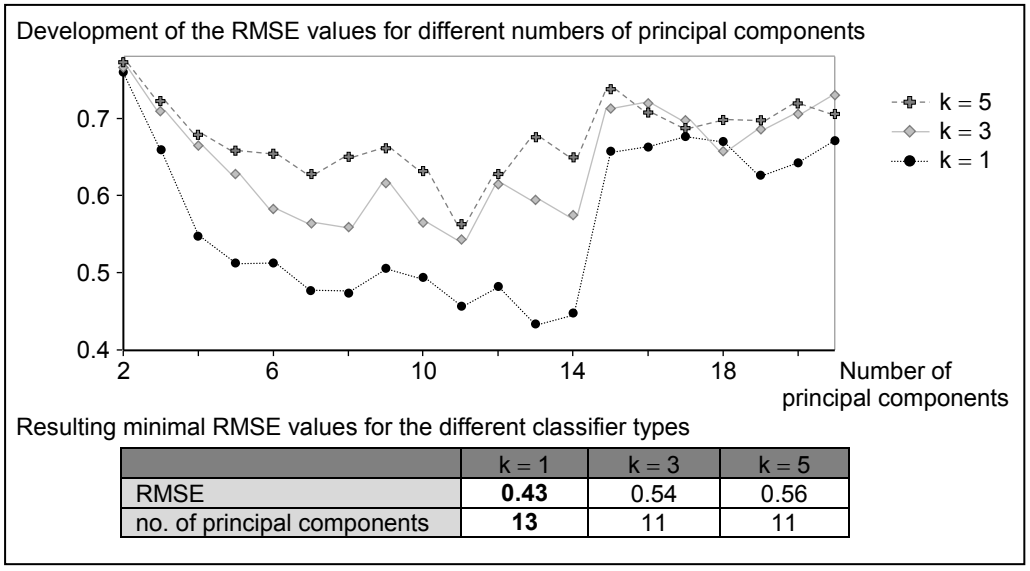


Figure 99: SL, PC for the distortion and double image dataset: ROC curves for 2-class classifiers

Nearest neighbour classification:

The class distribution of the *k*-nearest neighbours is analysed. If the majority of the nearest neighbours are classified as AC (class 3, 4 and 5) or UAC (class 1 and 2), the test image is assigned to the same class as the nearest training image of the majority belongs. First, it is examined for which dimension of the feature space the classifier achieves the lowest RMSE value for $k = 1$, $k = 3$, and $k = 5$. The development of the RMSE values for various feature dimensions is shown in Figure 100. The x-axis shows the number of principal components used and the y-axis the resulting RMSE values.



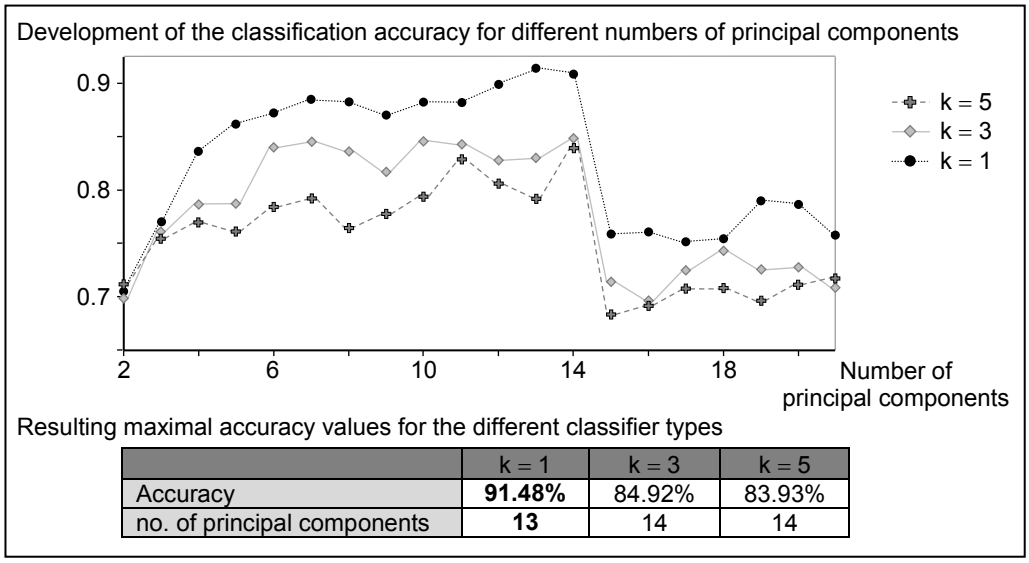
Resulting minimal RMSE values for the different classifier types

	k = 1	k = 3	k = 5
RMSE	**0.43**	0.54	0.56
no. of principal components	**13**	11	11

Figure 100: SL, kNN for the distortion and double image dataset: resulting RMSE values

The development of the RMSE values is very similar for classifiers that consider 1, 3 or 5 nearest neighbours. For k =1, the classifiers achieve the lowest RMSE values. The highest RMSE values are achieved for classifiers that consider 5 nearest neighbours. If the number of principal components is increased from 2 to 13, the RMSE values decrease. For 14 principal components, the RMSE values increase sharply again. If the feature dimension is further increased, the RMSE values hardly decrease again. The minimal possible RMSE values of the classifiers are summarised in the table of Figure 100. The lowest RMSE value of 0.43 is achieved by the classifier for $k = 1$ in 13-dimensional component space. The second best RMSE value of 0.54 is achieved by considering 3 nearest neighbours and 11 principal components. The highest RMSE value of 0.56 reaches the classifier for $k = 5$ and 11 principal components.

The development of the recognition accuracies for various feature dimensions is shown in the diagram of Figure 101. The x-axis shows the used dimensions of the feature space and the y-axis the accuracies. Comparing the development of the RMSE values with the development of the recognition accuracies, an opposing trend can be found. If the RMSE values increase, the accuracies decrease and vice versa. For all feature dimensions, the accuracy values are highest for classifiers that consider only 1 nearest neighbour. From 2 to 13 principal components, the accuracy values increase. For 14 dimensions, the recognition accuracies decrease sharply. After that, the accuracies hardly increase again. The maximum possible recognition accuracies of the classifiers are summarised in the table below the diagram of Figure 101. The highest accuracy of 91.48% reaches the classifier for $k = 1$ in the 13-dimensional component space. This value is not far from the ideal value of 100%. The accuracy values of the classifier for $k = 3$ and $k = 5$ in the 14-dimensional component space follow with values of 84.92% and 83.93%. All 3 classifier achieve recognition accuracies that are significantly above the value of the standard method.



	k = 1	k = 3	k = 5
Accuracy	**91.48%**	84.92%	83.93%
no. of principal components	**13**	14	14

Figure 101: SL, kNN for the distortion and double image dataset: resulting accuracies

In the 13-dimensional component space, the 1st order classifier for k = 1 achieves both the lowest RMSE value and the highest accuracy. Therefore, the labelling of the test images by this classifier is examined in more detail, as shown in Figure 102.

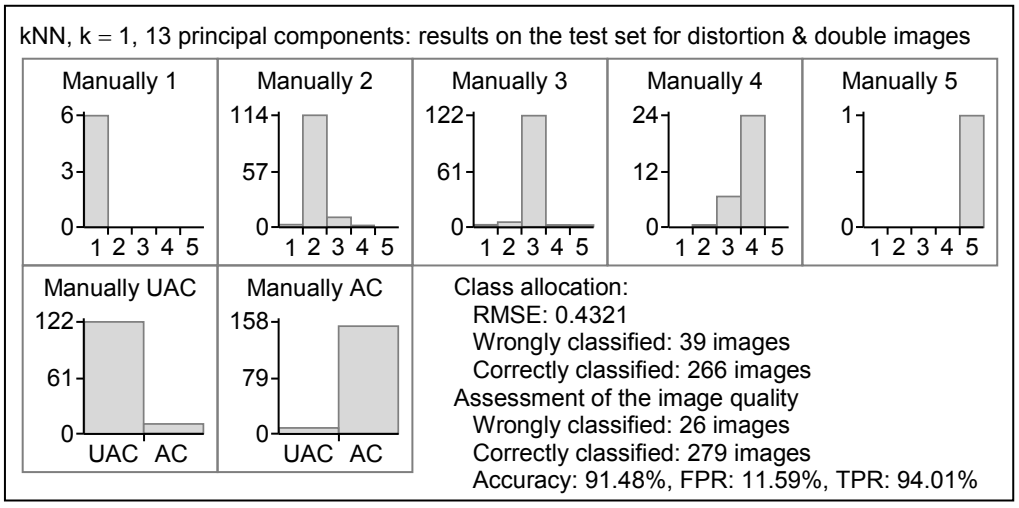


Figure 102: SL, kNN for the distortion and double image dataset: possible result

The first and the last histogram of the top row show that test images, which are manually labelled with 1 and 5 rating points, respectively, are 100% assigned to these classes by the classifier. In addition, test images belonging to the other rating classes are mainly marked with the corresponding class label. However, there are some misclassifications. Images that are manually labelled with 2 rating points are assigned to rating classes 1, 2, 3 and 4. The misclassifications to rating classes 3 and 4 are problematic because images of an unacceptable quality receive a positive label. Likewise, test images, which are manually labelled with 3 and 4 rating points, are assigned by the classifier to rating classes 1, 2, 3, 4 and 5. The difficulty here is the assignment to rating

classes 1 and 2. Thereby, a negative label is wrongly assigned to test images which show a customer suitable quality. The classifier reaches the accuracy of 91.48%, the FPR of 11.59%, and the TPR of 94.01%. The TPR value is higher than the value determined by the standard method (86.23%). Here 94.01% of the test images that are manually labelled as acceptable receive a positive label. Thus, this classifier would be less costly than the standard method, since fewer vehicles of acceptable quality would be mistakenly sent to rework. Likewise, the classifier achieves a much lower FPR than the standard method (55.80%). Here, only 11.59% of the test images with an unacceptable quality receive a positive label. The number of vehicles with an unacceptable quality that would reach the customer would be 44.21% lower. The classifier seems to be better suited than the standard method.

The next step is to investigate which classifier types achieve the lowest FPR and the highest TPR for the given classification task. The results are summarised in Table 31. All FPR values listed here are lower than the value of the limit consideration. The lowest FPR of 10.14% is reached by the NN classifier for $k = 1$ in the 14-dimensional feature space. This classifier also achieves the TPR of 91.62%, which is higher than the TPR of the standard method. The highest TPR of 94.01% is achieved by the classifier for $k = 1$ in the 13-dimensional component space. This classifier reaches the FPR of 11.59%. Similarly, the classifiers for $k = 3$ and $k = 5$ for 14 principal components achieve the TPR of 88.62% and FPR values of 11.59%, 19.57%. All these classifiers achieve better values than the standard method and are considered more suitable for the given classification problem.

NN	no. of principal components	FPR	TPR
	Minimal FPR values		
$k = 1$	14	10.14%	91.62%
$k = 3$	6	15.94%	83.83%
$k = 5$	11	19.57%	85.03%
	Maximal TPR values		
$k = 1$	13	11.59%	94.01%
$k = 3$	14	19.57%	88.62%
$k = 5$	14	21.74%	88.62%

Table 31: SL, kNN for the distortion and double image dataset: min FPR and max TPR

Finally, the mapping rule of the kNN is reduced to a 2-class problem where the image quality is classified only as AC or UAC. A test image is considered as acceptable if at least *m* of the *k*-nearest training images represent a customer suitable image quality. The ROC curve of the classifier is determined by varying the number of required acceptable training images *m*. Here, the number of considered nearest training images is set to $k = 30$, $k = 20$, and $k = 10$. The ROC curve giving the maximum AUC value is shown in Figure 103.

The resulting maximum values of the area under the curve are shown. The classification ability of the classifier is better if the AUC value is higher.

	k = 10	k = 20	k = 30
AUC	**0.86**	0.82	0.76
no. of principal components	**12**	11	2

ROC-curve: this diagram illustrates the compromise between the matches (correctly positive classifications) and the costs (false positive classifications).

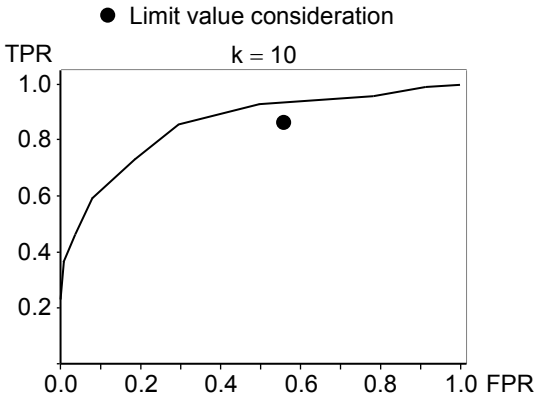


Figure 103: SL, kNN for the distortion and double image dataset: ROC curve for a 2-class classifier

The maximum AUC value of 0.86 is determined in the 12-dimensional component space when the number of nearest training images k is set to 10. The ROC curve shows that high TPR values also entail high FPR rates. Again it can be shown, that the choice of the most suitable classifier type is a compromise between a high TPR and low FPR. A possible classification result is shown in Figure 104.

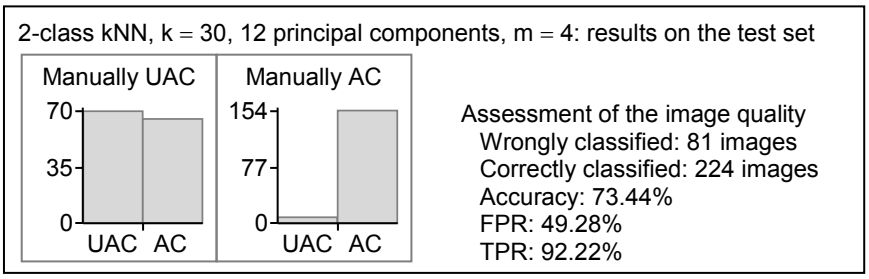


Figure 104: SL, 2-class kNN for the distortion and double image dataset: possible labelling of the test images

The classifier achieves a high FPR rate of 49.28%. Almost half of all test images that are manually rated as unacceptable receive a positive label from the classifier. Nevertheless, the classifier would send 6.52% fewer vehicles with an unacceptable image quality to the customers than the limit value consideration. On the other hand, the classifier achieves the TPR of 92.22%. The qualities of test images, which are manually considered customisable, are also classified as 92.22% as such. Thus, the TPR is higher than the value of the limit value consideration. Therefore, the classifier would be less costly than the limit analysis because fewer vehicles with an acceptable quality

would be mistakenly sent to rework. Overall, the classifier achieves the recognition accuracy of 73.44%, which is above the accuracy of the limit analysis.

Learning vector quantisation:

A learning rate α of 0.03, a window width *w* of 0.3, and an adjustable learning rate ε of 0.1 are applied to determine the prototypes. The prototype vectors are determined by the learning rules LVQ1, LVQ2.1 and LVQ3. The test image is assigned to the same class as the nearest prototype vector belongs. In the following, it is checked which combination of learning rule and prototype number achieves the best result for the given classification task. First, the development of the RMSE values is investigated for all 3 learning rules and different prototype numbers. The result can be seen in the upper diagram of Figure 105.

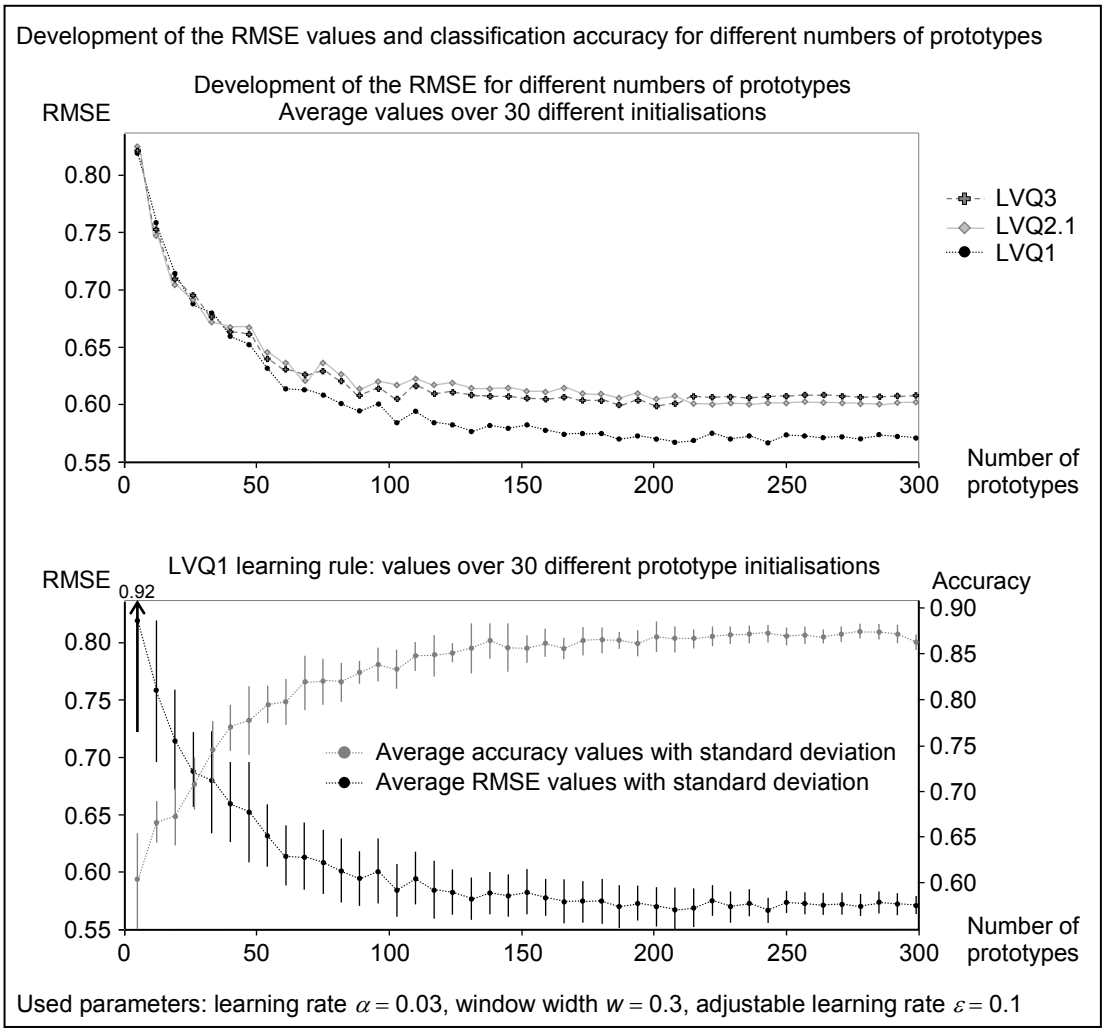
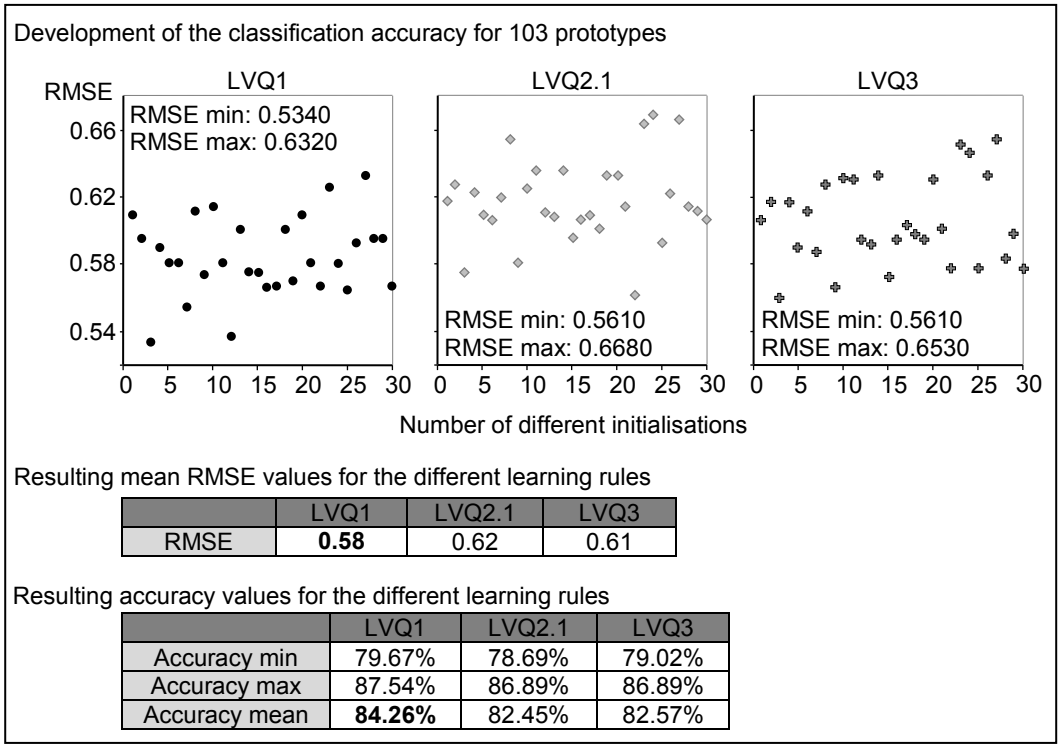


Figure 105: SL, LVQ for the distortion and double image dataset: evaluate different numbers of prototypes

The average values over 30 different initialisations of prototypes are shown. The number of prototypes in increased in intervals of 7. The x-axis shows the number of prototypes and the RMSE values are plotted on the y-axis. For the 3 learning rules, the de-

velopment of the RMSE values is very similar. As the number of prototypes is increased, the RMSE values decrease. For small numbers of prototypes, the decline of the RMSE values is greatest. The larger the number of prototypes is chosen, the lower the decrease in the RMSE values. Up to 54 prototypes, the RMSE values hardly differ for the 3 learning rules. From then on, the resulting RMSE values of the classifiers determined by learning rule LVQ1 are a little bit lower than the RMSE values of the other 2 classifiers.
Therefore, the RMSE values and the recognition accuracies resulting from these classifiers are more closely investigated, as shown in the second diagram of Figure 105. The RMSE values and the recognition accuracies are plotted against the number of prototypes. In addition to the average values, the standard deviations are shown for 30 different prototype initialisations. From the standard deviations shown, it can be seen that the resulting RMSE values and the classification accuracies depend on the random initialisation of the prototypes. The development of the classification accuracies is inversely related to the development of the RMSE values. The more prototypes are used, the higher the classification accuracies.

For the following investigation, 103 prototypes are used. For this number of prototypes, the best relationship is assumed between the number of prototypes and the classification quality. As Figure 105 shows, if fewer prototypes are used, the RMSE values increase and the accuracies decrease. In contrast, if more prototypes are used, the decrease in the RMSE values is low and the accuracies increase slowly.
Since the RMSE values and the recognition accuracies depend on the random initialisation of 103 prototypes, the results of 30 different initialisations are examined in more detail, as shown in Figure 106. The upper 3 diagrams show the resulting RMSE values of classifiers determined by the 3 learning rules. The x-axis shows the number of different initialisations and the y-axis the RMSE values. Again, it becomes clear that the RMSE values depend on the random initialisation of the prototypes. The resulting recognition accuracies are summarised in the lower table of Figure 106. For 30 different initialisations, the recognition accuracies of the classifiers determined by learning rule LVQ1 are between 79.67% and 87.54%. For the learning rule LVQ2.1 between 78.69% and 86.89% and for LVQ3 between 79.02% and 86.89%. The classifier determined by learning rule LVQ1 achieves on average the lowest RMSE value of 0.58 and the maximum classification accuracy of 84.26%.



Resulting mean RMSE values for the different learning rules

	LVQ1	LVQ2.1	LVQ3
RMSE	**0.58**	0.62	0.61

Resulting accuracy values for the different learning rules

	LVQ1	LVQ2.1	LVQ3
Accuracy min	79.67%	78.69%	79.02%
Accuracy max	87.54%	86.89%	86.89%
Accuracy mean	**84.26%**	82.45%	82.57%

Figure 106: SL, LVQ for the distortion and double image dataset: resulting values for 103 prototypes

Thus, the evaluation of the test images by this classifier is examined more closely. The labels resulting from a classifier with an arbitrary prototype initialisation are shown in Figure 107.

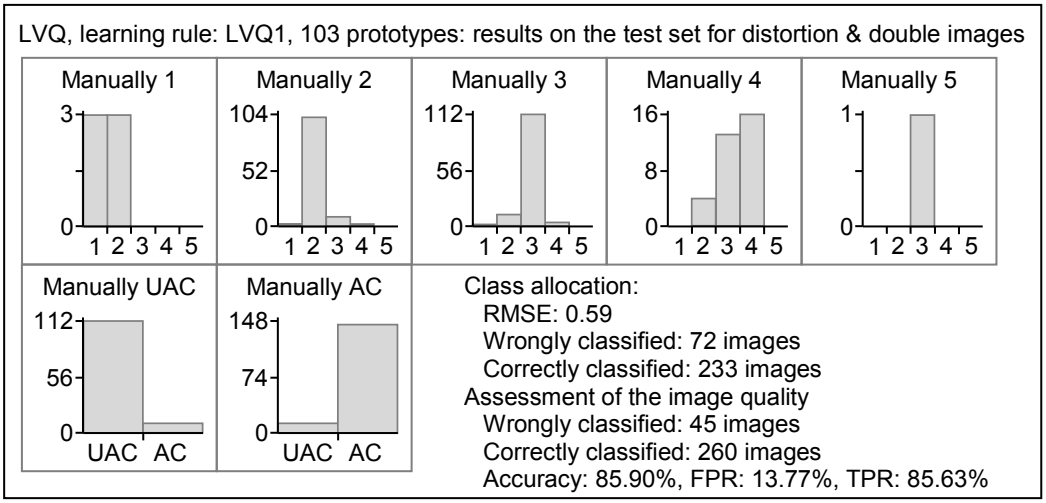


Figure 107: SL, LVQ for the distortion and double image dataset: possible labelling of the test images

The first histogram shows that test images that are manually labelled with 1 rating point are assigned by the classifier equally to rating classes 1 and 2. Since both rating classes represent an unacceptable image quality, the misclassification has no impact on the assessment of the image quality. Likewise, the last histogram of the first row shows that the test image, which is manually labelled with 5 rating points, is assigned to rating

class 3. This misclassification has also no influence on the assessment of the image quality. In contrast, test images, which are manually labelled with 2 rating points, are assigned to classes 1 and 2 as well as to rating classes 3 or 4. The misclassifications to classes 3 and 4 are problematic because vehicles with unacceptable image quality would reach the customers. Test images that are manually labelled with 3 or 4 rating points are assigned by the classifier to rating classes 1, 2, 3, and 4. The misclassifications to rating classes 1 and 2 would lead to vehicles being incorrectly sent to rework despite an acceptable image quality. Overall, this classifier achieves the recognition accuracy of 85.90%, which clearly exceeds the value of the limit consideration (67.21%). In addition, this classifier achieves the TPR of 85.63%. This value is slightly below the value of the standard method. Here, 85.63% of all test images of acceptable quality receive a positive label. As well, the classifier reaches the FPR of 13.77%, which is well below the FPR of the standard method. Only 13.77% of the test images with an unacceptable quality receive a positive label. Thus, up to 42.03% fewer vehicles with an unacceptable quality would reach the customers.

In the next step, the FPR and the TPR of classifiers determined by the learning rule LVQ1 are analysed for different numbers of prototypes, as shown in Figure 108. Again, the averages and standard deviations are shown for 30 different initialisations.

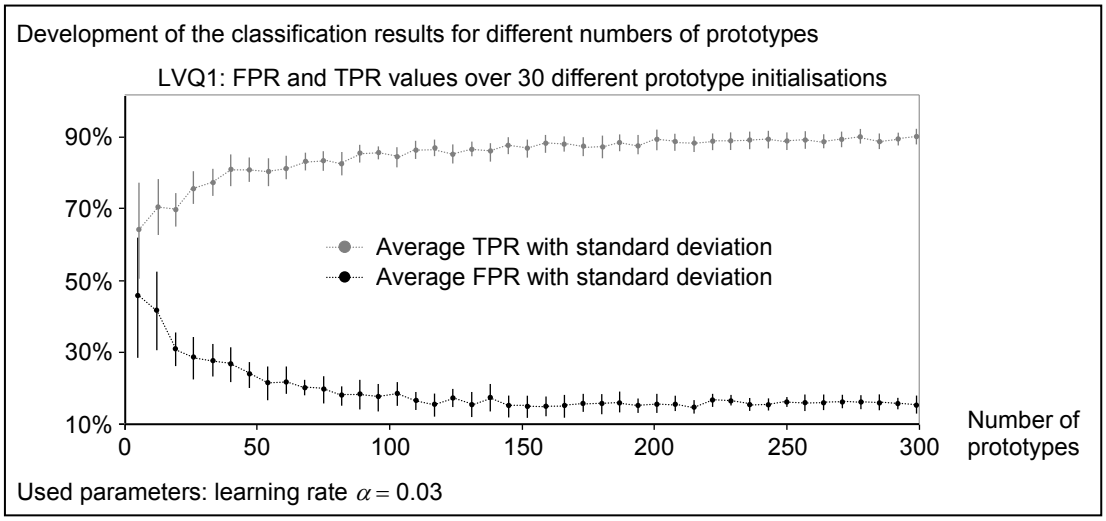


Figure 108: SL, LVQ for the distortion and double image dataset: resulting values for LVQ1

The x-axis shows the number of used prototypes and the y-axis the percentages of the FPRs and TPRs, respectively. It can be seen that the development of the TPR is inverse to the development of the FPR. As the number of prototypes is increased, the FPR decrease from 45.65% to 15.36%. It should be noted that all FPRs are below the value of the standard method (55.80%). Similarly, the true positive rate increases from 63.89% to 89.64% as the number of prototypes is increased. For 110 prototypes, the classifier reaches for the first time a TPR that exceeds the value of the standard method (86.23%). Overall, the classifiers tend to label the test images as unacceptable so that the FPR values are low.

Now, a labelling of the test images is investigated for a classifier using a large number of prototypes. The classification result of the classifier with 257 prototypes, which are determined by the learning rule LVQ1, is shown in Figure 109. The resulting histograms are hardly different from the labelling of the classifier with 103 prototypes. Only the number of incorrectly assigned test images is lower. The classifier reaches the RMSE of 0.52. The recognition accuracy of 88.52% and the TPR of 89.82% are higher than the values of the standard method. Here, also 3.59% more test images with an acceptable quality are considered acceptable. Thus, the classifier would be less costly because fewer vehicles of acceptable quality would be mistakenly sent to rework. The classifier also achieves the FPR of 13.04%, which is well below the FPR of the limit analysis. Thus, 42.76% fewer test images with an unacceptable image quality receive a positive label and would not reach the customers. It is assumed that the classifier is better suited for the given classification task than the standard method.

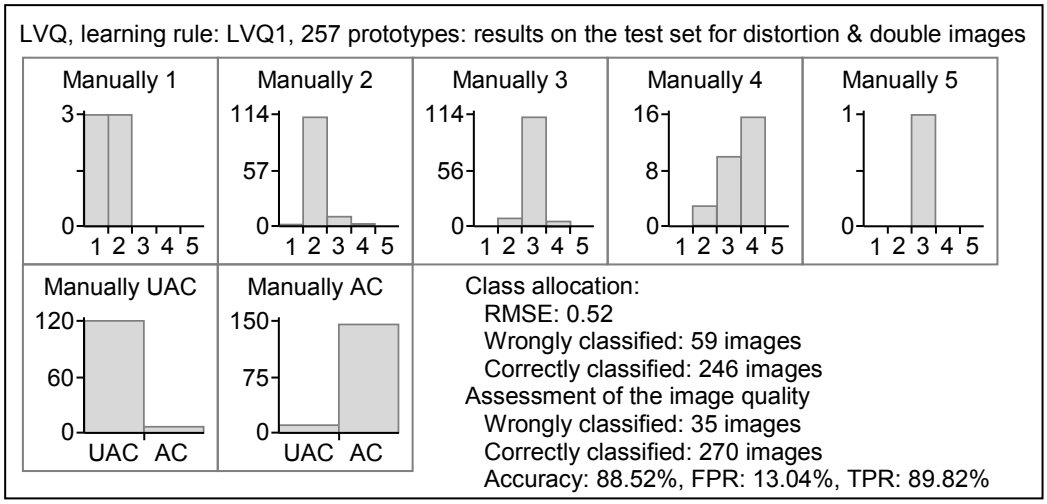


Figure 109: SL, LVQ for the distortion and double image dataset: possible labelling of the test images II

Finally, the mapping rule of the LVQ is again reduced to a 2-class problem. Now the LVQ tries to classify the quality of the test images only as AC or UAC. A test image is classified as acceptable if at least *m* of the *k*-nearest prototype vectors represent a customer suitable image quality. Here, the number of considered nearest prototype vectors is set to $k = 30$, $k = 20$, and $k = 10$. The ROC curve of the classifier is determined by varying the number of required acceptable prototype vectors *m* from 1 to k. The classification is based on 103 prototype vectors that are determined by the learning rule LVQ1. The ROC curve, which results in the maximum AUC value, is shown in Figure 110. The maximum AUC value of 0.72 can be determined if the number of nearest prototype vectors k is set to 10. The diagram shows that the 2-class classifier is not able to achieve a combination of FPR and TPR that is better than the results of the limit value analysis. The choice of the most suitable classifier type is a compromise between a high TPR and low FPR.

The resulting maximum values of the area under the curve are shown. The classification ability of the classifier is better if the AUC value is higher.

	k = 10	k = 20	k = 30
AUC	**0.72**	0.69	0.66
no. of prototypes	**103**	103	103

ROC-curve: this diagram illustrates the compromise between the matches (correctly positive classifications) and the costs (false positive classifications).

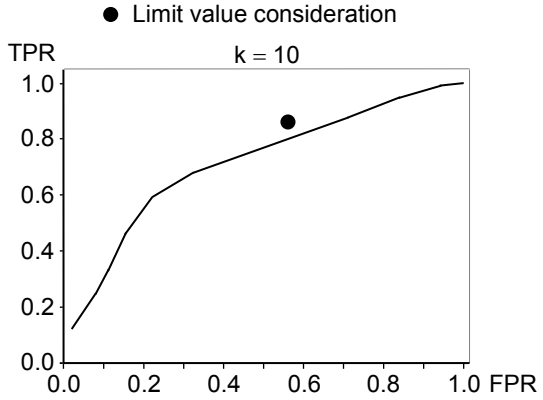


Figure 110: SL, LVQ for the distortion and double image dataset: ROC curve for a 2-class classifier

Summary of the obtained results:

The polynomial classifiers achieve lower FPR values than the standard method. Since the classifiers tend to label the test images as unacceptable, lower TPR values are also obtained. Thus, the classifiers would send fewer vehicles of unacceptable quality to the customers. Unfortunately, the classifiers would also cause more costs because more vehicles of acceptable quality would be mistakenly sent to rework. For example, the 1st order classifier achieves the accuracy of 71.48%, the FPR of 31.88%, and the TPR of 74.25% for 20 principal components. If the allocation rule of the PC is reduced to a 2-class problem, classifiers can be implemented that, depending on the parameter, achieve either higher TPR values or lower FPR values than the standard method.

In contrast, the kNN classifiers appear to be more suitable for the given classification task. Here, FPR values that are lower and TPR values that are higher than the values of the standard method can be determined. For example, the NN classifier for k = 1 and a dimension of 13 principal components achieves the accuracy of 91.48%, the FPR of 11.59%, and the TPR of 94.01%. Even if the mapping rule of the NN classifier is reduced to a 2-class problem, better class assignments than the results of the limit value analysis can be achieved.

By applying the learning rule LVQ1, classifiers can be determined, which achieve lower FPR values than the standard method. If the number of prototypes is chosen large enough, TPR values are also obtained which are higher than the TPR of the limit analysis. For example, the classifier with 257 prototypes achieves the accuracy of 88.52%, the FPR of 13.04%, and the TPR of 89.82%. In contrast, a 2-class LVQ is unable to achieve a combination of FPR and TPR that is better than the reference values of the limit value analysis.

7.5 Semi-supervised learning: impact on the classification results

Until now, it has been investigated, which classification results are possible if all training images are labelled manually. Now tests are conducted to investigate the semi-supervised learning behaviour of selected classifiers. It is analysed whether better classification results can be achieved by combining labelled and unlabelled training data. The principle of semi-supervised learning (SSL), described in chapter *6.2.3*, is an iterative procedure in which the learning process uses its own predictions to teach itself. The classifier is trained based on selected labelled training data and then applied to the unlabelled training data. The images with the most confident predictions are removed from the unlabelled training set and added to the labelled training images. Therefore, the definition of robust criteria for transferring images during the learning phase is essential since an incorrectly labelled image can significantly reduce the performance of the classifier. The following defines selection criteria that evaluate whether the unlabelled image is reliably classified and can be included in the labelled training dataset.

Polynomial classification:

In order to determine whether a training image is reliably classified, 2 selection criteria, shown in Figure 111, are conceivable [SAKIC 12: p. 83]:

- *Threshold* θ_1*:* θ_1 is the limit for the maximum probability that the image belongs to the corresponding rating class. It is checked if the maximum probability that the image belongs the corresponding rating class is greater than the threshold θ_1. If this is the case, the image is clearly classified and the image is transferred to the training dataset with its autonomously generated label, as shown in the upper part of Figure 111.
- *Threshold* θ_2*:* θ_2 is limit for the difference between the largest and the second largest assignment probability. It is analysed whether the difference between the largest and the second largest class assignment probability is greater than the threshold θ_2. If so, the class assignment is unique. The image with its autonomously generated label is transferred in the training dataset, shown in the lower part of Figure 111.

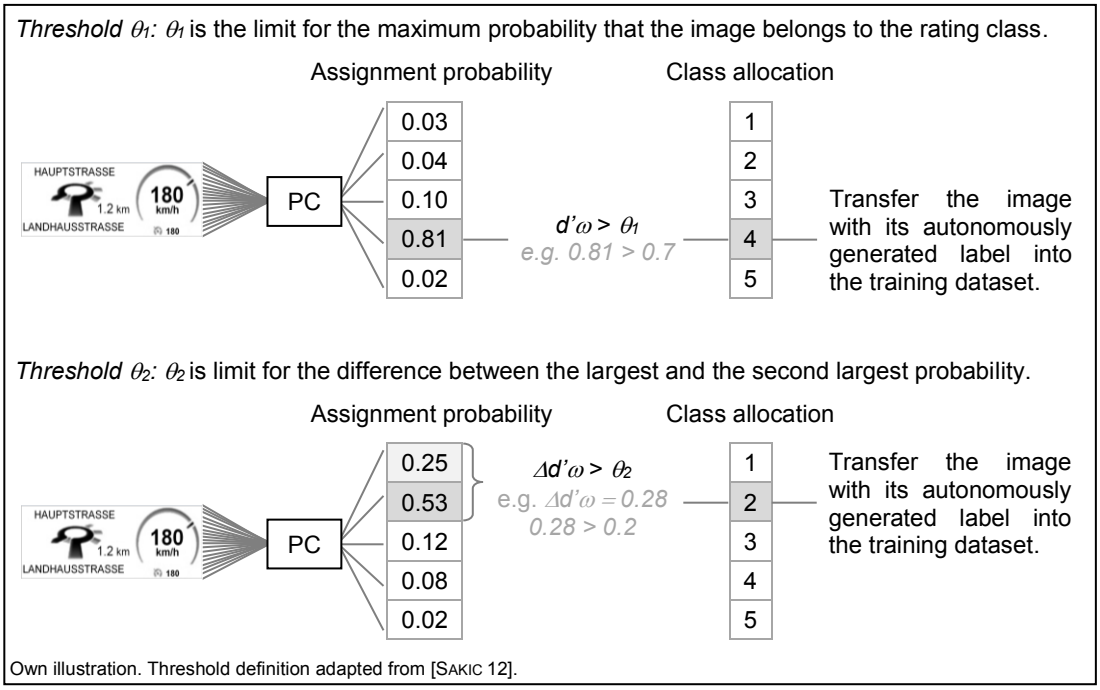


Figure 111: SSL, PC: threshold value definition

K-nearest neighbour classification:

Unlabelled images with a short distance to the next labelled training sample have a high probability of belonging to the same class as the next labelled training image [CEBRON 08: p. 67]. It is assumed that the closer the unlabelled image is to the corresponding labelled training sample, the higher is the probability that the image is classified correctly [WANG et al. 11]. To classify an image uniquely, it is checked whether the image is within a maximum defined distance [SZELISKI 10: p. 201]:

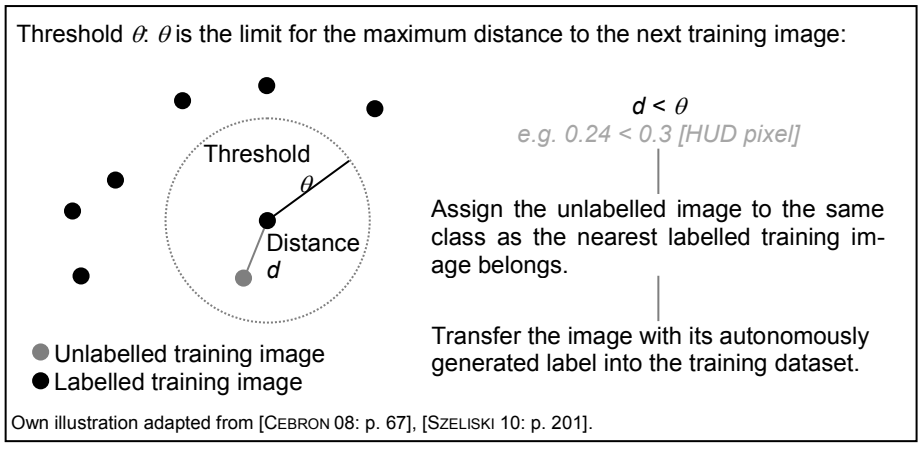


Figure 112: SSL, kNN and LVQ: threshold value definition

- *Threshold θ:* θ is the limit for the maximum distance to the next training image. It is checked if the distance to the next training sample is smaller than the maximum distance given by the threshold θ. If so, the image with its autonomously generated label is transferred into the training set. The transfer condition is shown in Figure 112.

If the threshold distance is too large, too many incorrect labels will be transferred autonomously into the training set. In contrast, if the threshold distance is too small, many correct labels are rejected and not transferred to the training set [SZELISKI 10: p. 201].

Learning vector quantisation:
The threshold definition is similar to that the of the NN classifier. For the LVQ it is also assumed that the shorter the distance between the unlabelled image and the next prototype, the higher the probability that the image is correctly classified [WANG et al. 11]:

- *Threshold* θ: θ is the limit for the maximum distance to the next prototype. It is checked if the distance to the next prototype is smaller than the maximum distance given by the threshold θ. If this criterion is fulfilled, the image with its autonomously generated label is transferred into the training set.

After defining the thresholds for transferring the unlabelled images, the SSL rule is applied to selected classifiers. Classifiers designed for the distortion dataset, the double image dataset, and the distortion and double image dataset are examined.

7.5.1 SSL: assessment algorithms for the distortion dataset

The classifiers are first trained with a manually labelled training set of variable size, consisting of only a few available training images, as shown in Table 32.

Initial number of labelled training images	Classes					Overall
	1	2	3	4	5	
10%	41	55	3	1	1	101
15%	61	83	5	2	1	152
20%	82	110	7	2	1	202
25%	102	138	8	3	1	252
30%	122	165	10	3	2	302
35%	143	193	12	4	2	354
40%	163	220	13	4	2	402
45%	184	248	15	5	2	454
50%	204	275	17	5	3	504
55%	224	303	18	6	3	554
60%	245	330	20	6	3	604
65%	265	358	21	7	3	654
70%	286	385	23	7	4	705

Table 32: SSL for the distortion dataset: initial number of labelled training samples

The selection of the labelled training images is done randomly. The size of the initially labelled images is set to 10%, 15%, …, 70% of all training images from each rating

class. This ensures that that at least 1 image from each rating class is selected. The trained classifiers label the remaining unlabelled training images by themselves. A newly labelled image is transferred into the training set if the image is classified reliably. A new training cycle is initiated if 25 new images have been transferred. This procedure is repeated until all available images that can be reliably classified are added to the training dataset. If no new training image is found that meets the robust selection criterion, the SSL process automatically terminates. In order to check the assessment result, the classifier is applied to the test dataset after each training cycle. The classification results are compared with values of an "ideal" classifier trained with all available manually labelled images.

Polynomial classification:
The semi-supervised learning rule is applied to the PC of 2nd order in the 10-dimensional component space. During supervised learning, this classifier achieves the RMSE of 1.12, the accuracy of 83.89%, the FPR of 39.31% and the TPR of 99.53%.

The outcome of the PC is an assignment probability vector that contains for each rating class the probability that the image belongs to the corresponding rating class. Normally, the image would always be assigned to the class for which the probability is greatest. During the SSL process, in the first step, a new label is accepted and added to the training set if the maximum probability that the image belongs to the rating class is greater than the threshold θ_1. Here, θ_1 is set to 0.6, 0.7, and 0.8. The size of the initially manually labelled training images is set to 10%, 15%, …, 70% of all training images from each rating class, as shown in Table 32.
The classification results after the termination of the SSL process are shown in the appendix *A.7* see Figure 160 and Figure 161. To summarise, the lower the threshold is chosen, the more images are labelled by the SSL process and transferred to the training set. In contrast, the higher the threshold, the less training cycles are possible. By reducing the initial training set size to less than 30%, the PC still yields a high TRP value while the accuracy and the FPR values suffer from wrong-labelled training images. In contrast, if the initial training set size is greater than 30%, the PC will provide good classification results. Depending on the threshold value, higher accuracy values and lower FPR values are achieved, such as those values determined by all manually labelled training images. For a threshold of 0.6 and initial training set sizes larger than 45%, average accuracies slightly exceeding the "ideal" value are obtained. This also applies to a threshold of 0.7 and initial training set sizes greater than 25%, as well as to a threshold of 0.8 and initial training set sizes greater than 30%.

As an example, the learning behaviour of the classifier for an initial training set size of 40% and a threshold of 0.7 is shown in the appendix *A.8* see Figure 166. Here, 7 training cycles are possible and the number of labelled training images increases from 402 to 577. During the SSL process, the average recognition accuracy is increased by around 0.74% with the number of labelled training images. For more than 452 training

images, the “ideal” value is even exceeded. Likewise, the average FPR values fall below the “ideal” value after the 2nd training cycle. In contrast, during the SSL process, the average TPR values are very similar to the corresponding “ideal” value. The average RMSE values increase at the end because too many false labels are added to the training set.

For a single run, the labels of the test images are further investigated, as shown in Figure 113. Overall, the resulting labels of the test images show only minor differences to those labels obtained from the classifier trained with all manually labelled images. Here, the accuracy of 85.00% and the FPR of 36.55% are achieved. Both values are better than the corresponding “ideal” values, indicating that the SSL algorithm uses those training samples that improve the recognition behaviour in an advantageous manner. The TPR of 99.53% can be obtained. This percentage corresponds exactly to the “ideal” value.

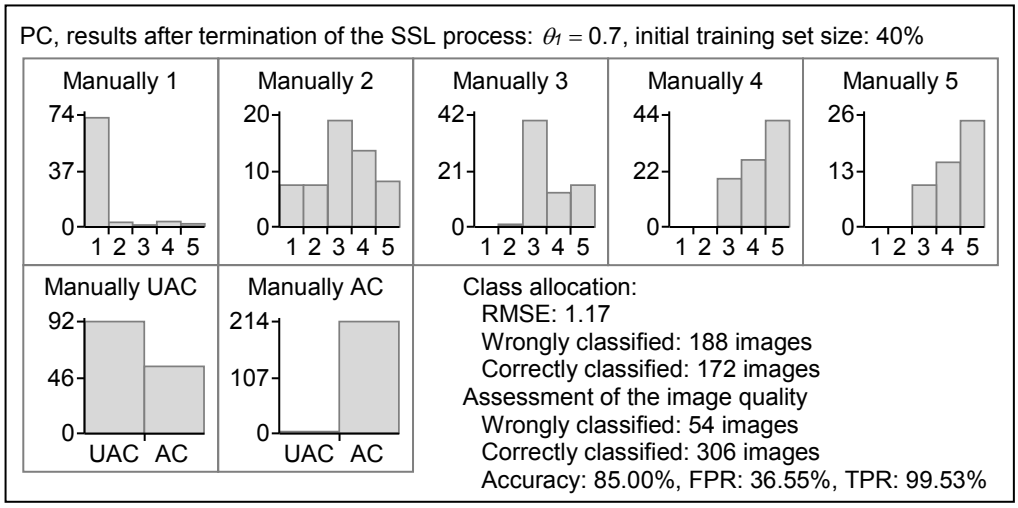


Figure 113: SSL, PC for the distortion dataset: possible labelling of the test images, θ_1

In the second step, an autonomously labelled image is transferred into the training set if the difference between the largest and the second largest class assignment probability is greater than the threshold θ_2, which is set to 0.1, 0.2, and 0.3.

The classification results after the termination of the SSL process are summarised in the appendix *A.7*, see Figure 162 and Figure 163. The smaller the threshold value is chosen, the more training images are autonomously labelled by the SSL process. Even with small training set sizes, high TPR values can be achieved which are above the “ideal” value. Only the average TPR value is lower than the “ideal value” for a threshold of 0.3 and an initial training set size of 45%. In contrast, the FPR values, the recognition accuracies, and the RMSE values become similar to the “ideal” value only for large initial dataset sizes. By applying threshold values of 0.1 and 0.2, the “ideal” accuracy and the “ideal” RMSE value are not achieved on average. The “ideal” accuracy is slightly exceeded only for a threshold value of 0.3 and an initial training set size greater than 60%. The average FPR values will be more similar to the corresponding “ideal” value if the initial training set includes more images. If the initial training set size is too

small, the average FPR values are very high. Consequently, small training sets do not contain sufficient information to transfer them to unknown data. On the other hand, the classifier may not be complex enough to extract the required information from the manually labelled training data. It turns out that there is a conflict between the number of manually labelled training images and the quality of the classification results.

Exemplarily, the learning behaviour of the classifier is investigated for an initial training set size of 40% and a threshold value of 0.3. The resulting learning curves are shown in the appendix *A.8* see Figure 167. The SSL process terminates after 12 training cycles. The number of labelled training images is increased from 402 to 702. During the SSL process, the average RMSE values increase by 0.06 as more training images are autonomously labelled by the classifier. The average recognition accuracy of the classifier increases at the beginning but then decreases when too many false labels are added to the training set. However, for 9 out of 30 runs, the accuracy achieved by the SSL process exceeds the limit of 83.89%. Likewise, the average FPR values decrease during the first 3 training cycles but increase afterwards. The average RMSE, the average accuracy and the average FPR values do not reach the “ideal” values obtained by all available manually labelled training images. In contrast, after each training cycle, the average TPR values increase slightly by 0.38% and exceed the “ideal” value after the termination of the SSL process.
A possible labelling of the test images is shown in Figure 114. It examines 1 of the 9 runs where the “ideal” accuracy is exceeded. It is positive that the TPR value of 100% is achieved. This is because test images that are manually labelled with 3, 4 and 5 rating points are exclusively assigned to those classes. Thus, no vehicle with an acceptable HUD image quality would be mistakenly sent to rework.

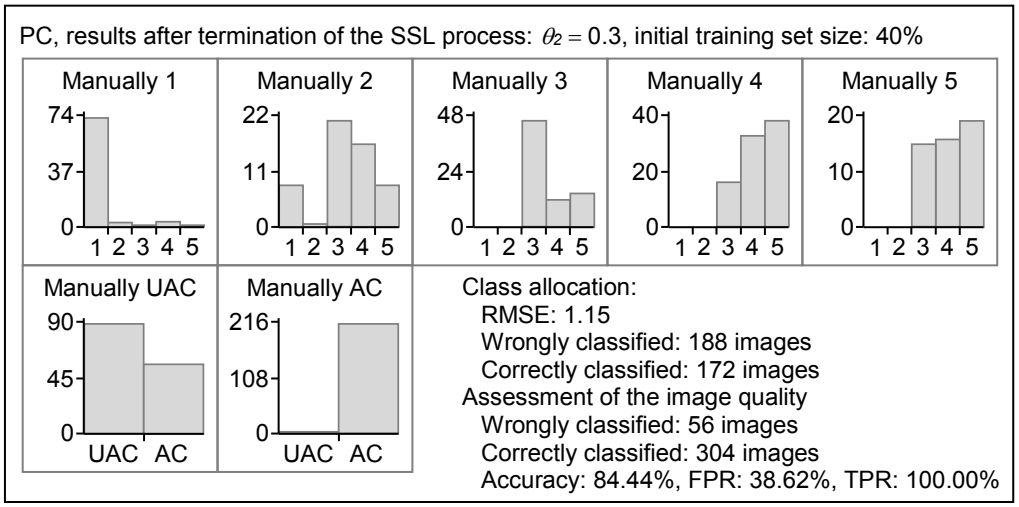


Figure 114: SSL, PC for the distortion dataset: possible labelling of the test images, θ_2

Likewise, test images, which are manually labelled with 1 rating point, are assigned to the corresponding rating class with almost no errors. In contrast, the PC cannot properly label images that are manually assigned to rating class 2. Largely, these images are

incorrectly assigned to rating classes 3, 4 and 5. This explains the FPR value of 38.62%. Nevertheless, this value is slightly below the “ideal” value obtained during the SL process.

In the last step, both selection criteria are combined. An autonomously labelled image is transferred into the training set only if the class assignment probabilities fulfil both threshold conditions simultaneously. The classification results after completing the SSL process are summarised in the appendix *A.7*, see Figure 164 and Figure 165. The smaller the thresholds, the more training images are transferred in the training dataset by the SSL process. By using training sets larger than 25%, good classification results are achieved, which equal to the “ideal” values or even better. It is noteworthy that the TPR of 100% is obtained for all 30 runs when the size of the initially labelled training set is set to 10% or 15% and thresholds of $\theta_1 = 0.6$ and $\theta_2 = 0.1$ are applied. These results are probably obtained due to the absence of poorly labelled images in the training set that affect the classification results. Thus, the manual label effort can be significantly reduced without losing performance and reliability. For small training set sizes, very high TPR values are obtained, but unfortunately, also high FPR values and the recognition accuracies are low. The reason might be that too small initial training sets do not contain sufficient information to generalise to unknown data and thus, many incorrectly labelled images are transferred into the training dataset.
The learning behaviour of the PC is exemplarily investigated for an initially labelled training set size of 40% and threshold parameters of $\theta_1 = 0.7$ and $\theta_2 = 0.2$. The learning curves are shown in Figure 168 in the appendix *A.8*. The SSL process terminates after 7 training cycles during which the number of training images is increased from 402 to 577. After each training cycle, the average recognition accuracy increases and the average FPR value decreases. Here, the accuracy reaches an average increase of 1.07% and the FPR an average decrease of 2.30%. After the 2nd training cycle, the classification accuracy exceeds the “ideal” value. Even just 1 cycle is needed to get FPR values that fall below the “ideal” value. In contrast, the average RMSE values hardly change during the SSL process and do not reach the “ideal” value. During the SSL process, average TPR values of more than 99% are reached after each training cycle.

Finally, the semi-supervised learning rule is applied to a 2-class PC of 2nd order in the 7-dimensional component space. If all manually labelled training images are used for training, a value for the area under the ROC curve of 0.87 can be determined. The learning behaviour of the classifier is investigated for various initially labelled training dataset sizes and a threshold value θ_1 that is set to 0.7. Thus, a new label is only transferred to the training dataset if the maximum probability that the image belongs to the corresponding rating class is greater than 70%. After each training cycle, the ROC curve is determined and the area under the curve is calculated. Therefore, the constant ϕ is varied between 0.1 and 10 and the increase is done in steps of 0.1. Figure 115 shows the average AUC values and the standard deviations over 30 runs after com-

pleting the SSL processes. The x-axis represents the number of initially labelled training images and the y-axis the resulting values for the area under the curve.

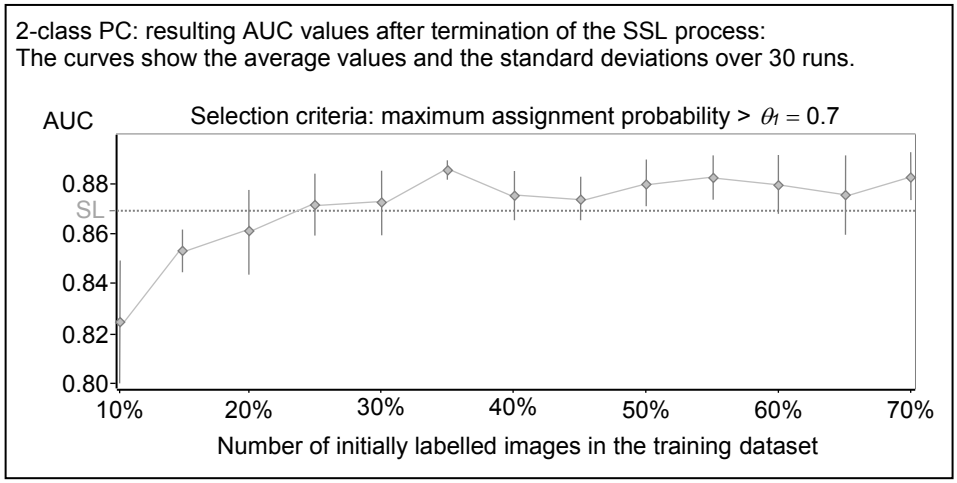


Figure 115: SSL, 2-class PC: results for the distortion dataset
$\theta_1 = 0.7$

If the size of the initial training set is set to 10% of all of all training images from each rating class, an average AUC value of 0.82 is already reached. If at least 25% of the manually labelled images are available, average AUC values greater than the desired value will be determined. In general, the higher the AUC value, the better the classification capability of the classifier. Consequently, it is possible to develop a 2-class PC with a small number of manually labelled images, which might be better suited to evaluate the HUD image quality than the “ideal” classifier. Thus, it is possible to reduce the manual label effort.

K-nearest neighbour classification:

The semi-supervised learning behaviour of the nearest neighbour classifier for $k = 1$ in the 3-dimensional component space is investigated. The “ideal” classifier trained with all available manually labelled training samples achieves the RMSE value of 1.37, the accuracy of 57.22%, the FPR of 6.21% and the TPR of 32.56%.

The NN classifier assigns an unlabelled image to the same class as the nearest reference pattern belongs. During the SSL process, an autonomously labelled image is transferred into the training dataset if the distance to the nearest reference pattern is less than a predetermined threshold distance θ. To start the investigation, the threshold distance must be fixed. Here, a separate threshold distance is determined for each rating class. This results in 5 threshold distances, one for each class. To determine reasonable values for θ, the distances between the test images and the nearest reference patterns are analysed. For this purpose, all manually labelled training images are used as reference patterns. At the beginning, it is checked to which rating class the test image belongs. Then the distance between the test image and the corresponding nearest reference image is determined. This is done for all test images. Finally, the minimal, mean, and maximal assignment distances are calculated for each rating class.

The calculated distances are summarised in Table 33. During the semi-supervised learning process, these distances are used to check if images can be transferred into the training set. Here, the class to which the unlabelled training image belongs is first determined. Thereafter, the class-specific threshold is selected and it is checked whether the distance between the image and the nearest reference pattern is smaller than the threshold. If this is the case, the autonomously generated label is transferred into the training dataset.

Subjective rating class	1	2	3	4	5
$\theta =$ minimal distance [HUD pixel]	0.083	0.075	0.207	0.113	0.038
$\theta =$ mean distance [HUD pixel]	0.371	0.339	0.377	0.362	0.306
$\theta =$ maximal distance [HUD pixel]	0.692	0.805	0.703	0.554	0.599

Table 33: kNN for the distortion dataset: used threshold distances

The obtained classification results after the termination of the SSL process are summarised in the appendix *A.9*, see Figure 169 and Figure 170. The smaller the threshold distances are chosen the more training images are autonomously labelled by the SSL process. By using training sets sizes larger than 30%, classification accuracies and TPR values exceeding the “ideal” values can be achieved. Using the mean or the minimal threshold distances, average TPR values are obtained which are higher than the “ideal” value, regardless of the size of the initially labelled training dataset. The average FPR values will be more similar to the “ideal” value as more training images are manually labelled. Thus, with a few exceptions, the FPR values are higher than the desired value.

The learning behaviour of the classifier is examined in more detail for a training set size of 45% and mean threshold distances. The learning curves are shown in the appendix *A.10* see Figure 171. The SSL process terminates after 11 training cycles, while the number of training images is increased from 454 to 729. During the SSL process, the average RMSE values are similar to the “ideal” value and decrease slightly in the last learning cycle. After each training cycle, the average recognition accuracies show an increase of 3.12%. Likewise, an average increase of 5.31% is obtained for the TPR values. The accuracy values and the TPR values are higher than the corresponding “ideal” value after each training cycle. The average FPR values increase slightly at the beginning but decrease at the end. The “ideal” value is unfortunately not reached.
A possible labelling of the test images is investigated more closely for a single run. The labels of the test images after the termination of the SSL process are shown in Figure 116. Images that are manually labelled with 1 or 2 rating points are mainly assigned to rating classes 1 or 2. This results in the FPR value of 6.21% that corresponds to the “ideal” value. In contrast, images that are manually labelled with 3, 4 or 5 rating points are mainly assigned to rating class 2. Nevertheless, a TPR value of 50.70% is achieved. Overall, the classifier achieves an accuracy of 68.08%. In summary, the SSL process can select training images that contain sufficient information to generalise to

unknown data. The absence of poorly labelled training images leads to better classification results.

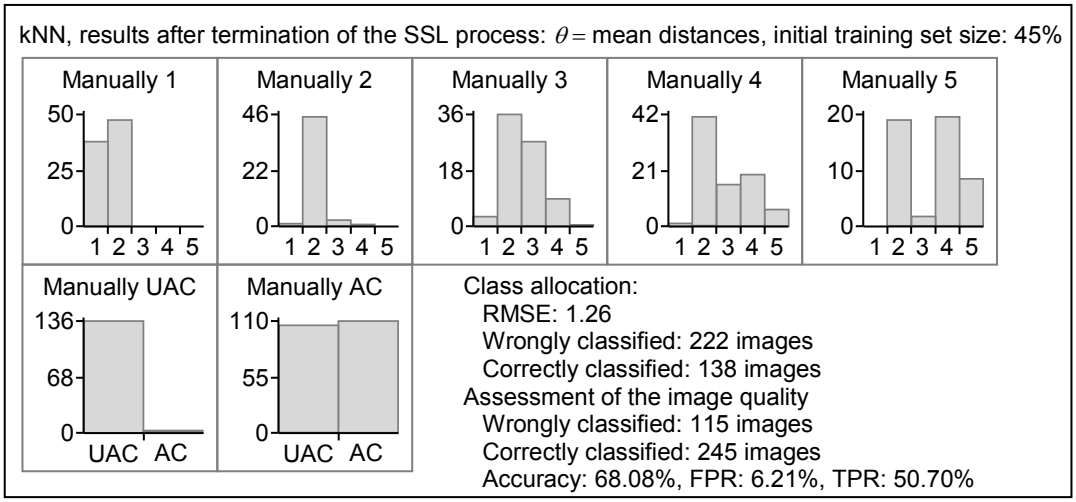


Figure 116: SSL, kNN for the distortion dataset: possible labelling of the test images

Finally, the SSL rule is applied to a 2-class NN in the 7-dimensional component space. The number of nearest prototype vectors k is set to 30. If all manually labelled training images are used for training, an AUC value of 0.88 can be determined for the area under the corresponding ROC curve. During the SSL process, an autonomously labelled image is transferred into the training dataset if the distance to the nearest reference pattern is less than the mean distance θ. After each training cycle, the ROC curve is determined and the area under the curve is calculated. For this, the number of images *m* required to classify the test image as AC is varied between 1 and k. After completing the SSL processes, the average AUC values and the standard deviations over 30 runs are summarised in Figure 117.

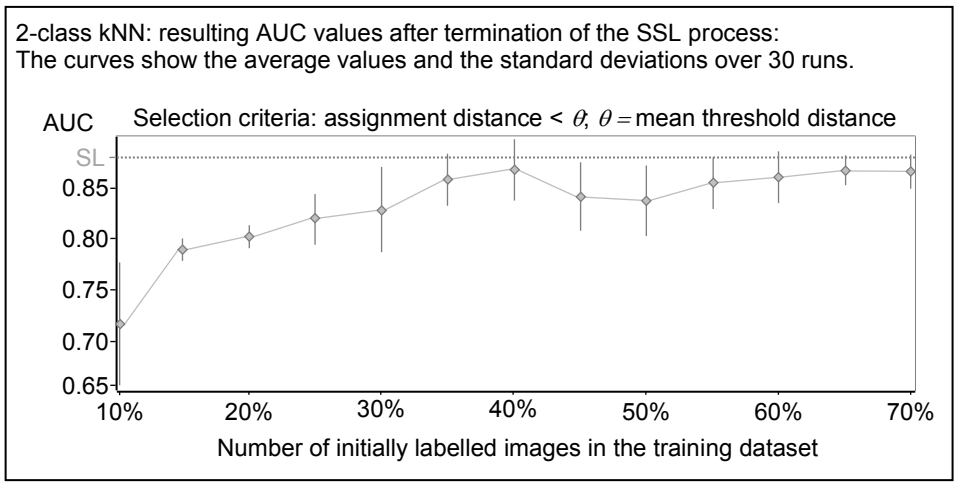


Figure 117: SSL, 2-class kNN: results for the distortion dataset
θ = mean threshold distance

The x-axis represents the number of initially labelled training images and the y-axis the resulting values for the area under the curve. The AUC values increase with the number of manually labelled images in the training set. The AUC values average over 30 runs become very similar to the "ideal" value for large initial training set sizes. However, the value of 0.88 is not exceeded. The largest AUC value of 0.86 is achieved on average for initial training set sizes of 40%. In summary, reducing the manual labelling effort leads to a reduction in the AUC values.

Learning vector quantisation:

The semi-supervised algorithm is applied to the learning vector quantisation. Here, 165 prototypes are used to approximate the underlying distribution of labelled training images. The prototypes are determined by the learning rule LVQ1. It is important to ensure that the same prototype initialisation is used for all investigations. If all manually labelled images are available, the RMSE value of 1.10, the accuracy of 88.06%, the FPR of 29.86% and the TPR of 100.00% are obtained.

The SSL process transfers an autonomously labelled image into the training dataset when the distance to the next prototype is less than a threshold distance θ. The determination of the threshold distances is based on the same principle as for the kNN classifier described in the previous text. For each of the 5 rating classes, a separate threshold value is determined from the available test images. The resulting minimal, mean, and maximal threshold distances are summarised in Table 34.

Subjective rating class	1	2	3	4	5
$\theta =$ minimal distance [HUD pixel]	0.735	1.069	0.499	0.399	0.244
$\theta =$ mean distance [HUD pixel]	1.591	1.890	1.455	1.393	1.188
$\theta =$ maximal distance [HUD pixel]	2.644	3.023	2.864	2.439	2.931

Table 34: LVQ for the distortion dataset: used threshold distances

The classification results after completing the SSL processes are shown in Figure 172 and Figure 173 in the appendix *A.11.* Large initial training set sizes lead to an increase in the average recognition accuracy. The "ideal" accuracy is not achieved for all used threshold distances and initial training set sizes. In contrast, the average RMSE values decrease by increasing the initial training set size. If the size of the initial training set covers more than 25% of all images from each class, the average RMSE values fall below the "ideal" value. For all investigated initial training set sizes and all threshold distances, the average TPR values vary between 99.35% and 99.86%. The average FPR values decrease with the number of initially labelled training images. The "ideal" rate of 29.86% is unfortunately not reached.

The learning behaviour for an initial training set size of 55% and a mean threshold distance is investigated more closely. The resulting learning curves are shown in the appendix *A.12* see Figure 174. In 11 training cycles, the number of labelled training samples increases from 554 to 854. During the SSL process, the average RMSE values

and the recognition accuracies change only slightly. There seems to be no relationship between the number of training images and the resulting prototype vectors. For each training cycle, the average RMSE values and the average accuracy values are below the corresponding “ideal” value. The average FPR values increase after each training cycle. Overall, an increase of 0.35% can be observed. Likewise, the average TPR values show an increase of 0.5% with an increasing number of labelled training images. Despite the low learning success, classification results are obtained which are very similar to the “ideal” results.

The resulting class allocation of the test images is examined for a single run, as shown in Figure 118. With the chosen initially labelled training images, it is possible to obtain the RMSE of 1.06, the accuracy of 88.33, the FPR of 28.47%, and the TPR of 99.54%. These values roughly correspond to the “ideal” values achieved by supervised learning. The investigation shows that images, which are manually labelled with 1 rating point, are mainly assigned to the corresponding rating class. Likewise, images that are manually labelled with 3, 4 or 5 rating points are also mainly assigned to the corresponding class. Only images, which are manually assigned to rating class 2, are sorted into all 5 rating classes by the classifier. This explains the high FPR value of the algorithm. The investigation shows that despite strictly chosen threshold distances, the LVQ algorithm is not able to extract sufficient information from the labelled training images to obtain better classification results than the SL process.

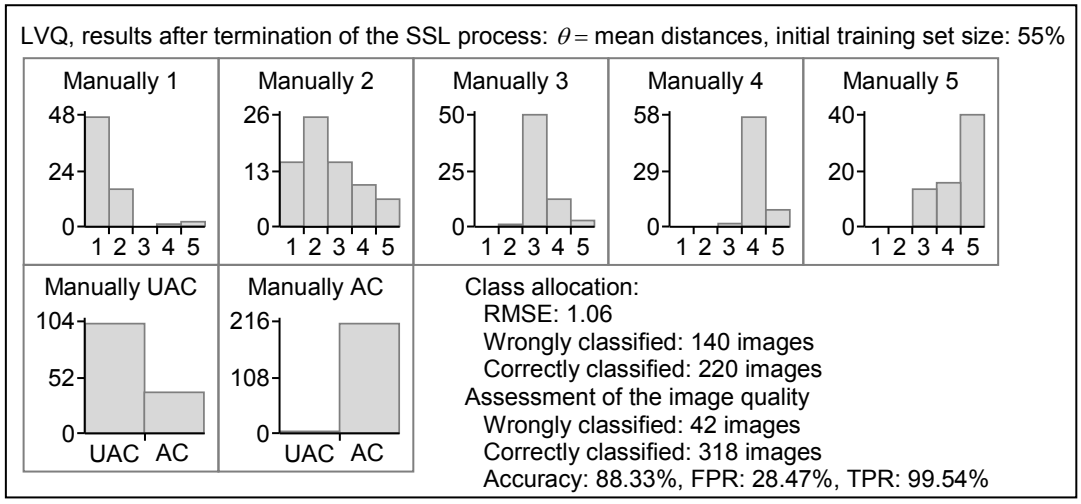


Figure 118: SSL, LVQ for the distortion dataset: possible labelling of the test images

Finally, the semi-supervised learning behaviour of the 2-class LVQ is mentioned. Here, 165 prototype vectors are trained and the number of nearest prototype vectors k is set to 30. If all manually labelled training images are used for training, an area under the ROC curve of 0.86 can be determined. During the SSL process, an autonomously labelled image is transferred into the training dataset if the distance to the nearest prototype vector is less than the mean threshold distance θ. After each training cycle, the AUC value is determined by varying *m*, the number of images that are required to clas-

sify a test image as AC. The average AUC values and the standard deviations over 30 runs after completing the SSL processes are shown in Figure 119.

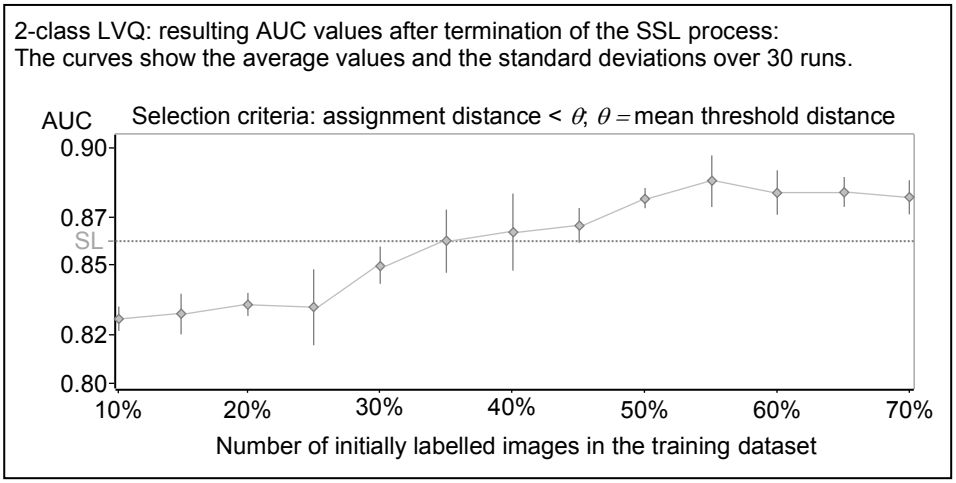


Figure 119: SSL, 2-class LVQ: results for the distortion dataset
θ = mean threshold distance

If the size of the initial training set is set to 10% of all of all training images from each rating class, an average AUC value over 0.82 is already reached. If more than 35% of the manually labelled images are used for the initial training, average AUC values greater than the reference value are determined. In general, the classification ability is better the higher the AUC value. Consequently, with a small number of manually labelled images, it is possible to develop a 2-class LVQ, which might be better suited to assess the HUD image quality than the “ideal” classifier. Thus, it is possible to reduce the manual label effort.

Summary of the obtained results:

For the distortion dataset, it is found that unlabelled data, when used in conjunction with a small amount of labelled data, can produce an improvement in the classification quality. The polynomial classifier yields good classification results depending on the threshold and the selection of manually labelled data. Similarly, the 2-class PC achieves higher AUC values with a small amount of manually labelled training images. After terminating the SSL process, the kNN classifier obtains classification accuracies and TPR values that exceed the “ideal” values. The reason could be that the SSL process attempts to avoid the use of poorly labelled training samples and selects those samples that advantageously improve the recognition behaviour. Unfortunately, for the 2-class kNN classifier, a decrease in the AUC values must be considered when reducing the manual labelling effort. The LVQ algorithm is scarcely able to extract sufficient information from the manually labelled training images to transfer it to unknown data. Only when the SSL rule is applied to a 2-class LVQ, the AUC value can be increased by reducing the manually labelling effort.

7.5.2 SSL: assessment algorithms for the double image dataset

During the SSL process, the classifiers are trained with a manually labelled training set of variable size, comprising a few percent of the available training data, as shown in Table 35. The selection of the labelled images happens randomly and it is ensured that at least 1 image is selected from each rating class. The classifier is retrained if 10 images are autonomously labelled by the SSL algorithm. After each training cycle of the SSL process, the classifier is applied to the test data. The results are compared with the corresponding results of an "ideal" classifier trained with all available manually labelled training samples.

Initial number of labelled training images	Classes					Overall
	1	2	3	4	5	
10%	12	17	3	3	1	36
15%	18	25	5	4	1	53
20%	24	33	7	5	1	70
25%	30	41	8	7	1	87
30%	36	50	10	8	1	105
35%	42	58	12	9	1	122
40%	48	66	13	10	1	138
45%	54	74	15	12	1	156
50%	60	82	17	13	1	173
55%	65	91	18	14	1	189
60%	71	99	20	16	1	207
65%	77	107	21	17	1	223
70%	83	116	23	18	1	241

Table 35: SSL for the double image dataset: initial number of labelled training samples

Polynomial classification:

The SSL procedure is applied to the PC of 3rd order in the 2-principal component space. The "ideal" classifier reaches the RMSE of 0.85, the accuracy of 90.00%, the FPR of 7.48%, and the TPR of 88.26%.

In the first step, a new label is accepted and added to the training set if the maximum probability that the image belongs to the rating class is greater than the threshold θ_1. Again, θ_1 is set to 0.6, 0.7, and 0.8. The classification results after completing the SSL process are shown in the appendix *A.13* see Figure 175 and Figure 176. The more manually labelled images are initially included in the training dataset, the closer the classification result approaches the "ideal" value. For small initial training set sizes, high TPR values are obtained. Unfortunately, the classifiers trained with a small initial training set size achieve high FPR values and high RMSE values. For a threshold of 0.6 and initial training set sizes of 15% and 20% respectively, good classification accuracies can be achieved that exceed the "ideal" value. Likewise, for these constellations, the TPR values are higher and the FPR values lower than the corresponding "ideal"

values. Here, the SSL algorithm selects those training images that advantageously improve the classification quality. Therefore, the learning curves for an initial training set size of 15% are shown by the way of example in the appendix *A.14* see Figure 181. During 23 training cycles, the number of labelled training samples increases from 53 to 283. Here, the learning success starts after 20 training cycles. The success is clearly visible in the last 3 training cycles. Thus, after 20 training cycles, enough information is available to generalise well to unknown data. Upon completion of the SSL process, the recognition accuracy and the true positive rate exceed the “ideal” value and the false positive rate falls below.

The resulting class assignments of the test images for a single run are shown in Figure 120. It can be seen that images, which are manually labelled with 1 rating point, are faultlessly assigned to the appropriate rating class. Images labelled with 4 or 5 rating points are assigned to rating classes 3 and 5. These misclassifications are unproblematic since the rating classes 3, 4 and 5 represent an acceptable image quality. Misclassifications that affect image quality estimation are obtained for test images manually labelled with 2 and 3 rating points. Nevertheless, the misclassifications are so low that still the accuracy of 98.06%, the FPR of 1.36% and the TPR of 97.65% are achieved. Thus, the manual label effort can be significantly reduced without a loss of performance. Here, the SSL process leads to better results. The accuracy and the TPR exceed the corresponding “ideal” value by 7.65% and 9.39%, respectively. Similarly, the obtained FPR is 6.12% below the “ideal” value.

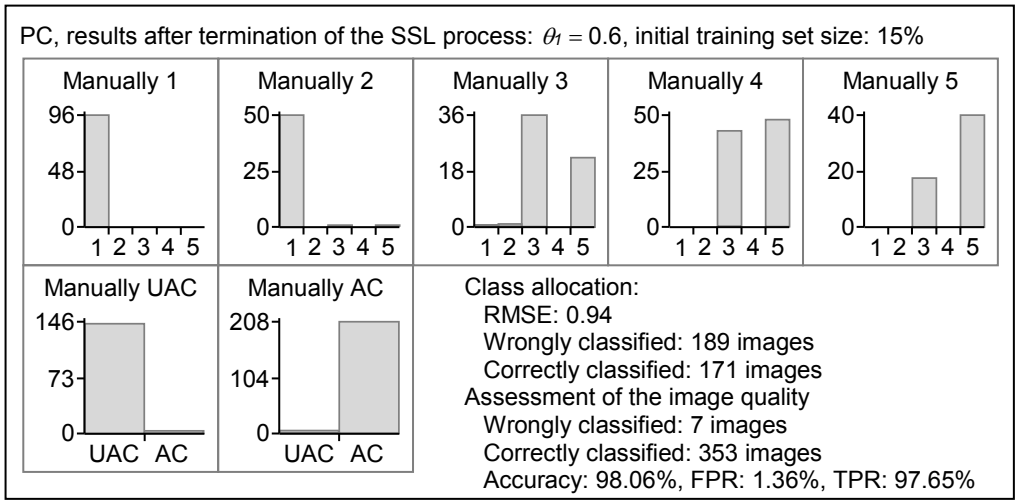


Figure 120: SSL, PC for the double image dataset: possible labelling of the test images, θ_1

In the second step, an autonomously labelled image is transferred into the training set if the difference between the largest and the second largest class probability is greater than the threshold θ_2, which is set to 0.1, 0.2, and 0.3. The classification results after completing the SSL process are shown in the appendix *A.13* see Figure 177 and Figure 178. If the initial training set size is set to more than 10%, recognition accuracies similar to those of the “ideal” value are reached. Smaller initial training set sizes do not

contain enough information to generalise to unknown data. Likewise, the resulting FPR values are similar to the “ideal” value for training set sizes larger than 10%. A threshold of 0.2 also achieves TPR values that approach very well to the “ideal” value.
The learning curve for a threshold of 0.2 and an initial training set size of 20% is exemplified in the appendix *A.14* see Figure 182. After each training cycle, the recognition accuracies increase and the FPR values decrease until the “ideal” values are exceeded or undercut. During the SSL process, the TPR values decrease at the beginning and increase again towards the end. The SSL algorithm tries to avoid the selection of poorly labelled training images. Here, the manual label effort can be considerably reduced without losing performance and quality, as exemplified for a single run in Figure 121.

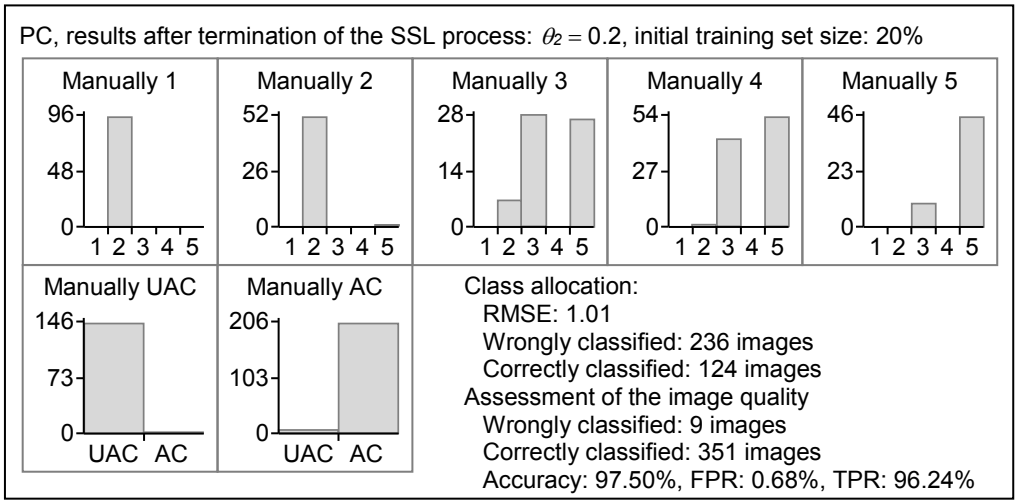


Figure 121: SSL, PC for the double image dataset: possible labelling of the test images, θ_2

With 70 random selected manually labelled training images, it is possible to achieve the accuracy of 97.50%, the FPR of 0.68%, and the TPR of 96.24%, after completing the SSL process. These results are even better than the results obtained for all manually labelled training images. The accuracy and the TPR value are 7.50% and 7.98% higher and the FPR value is 6.80% lower than the corresponding “ideal” values. Test images that are manually labelled with 1 or 2 rating points are mainly assigned to rating class 2. Likewise, test images that are manually labelled with 4 or 5 rating points are mainly assigned to rating classes 3 and 5. The TPR value of 96.24% indicates that 3.76% of the images, which are manually labelled with 3 or 4 rating points, are wrongly assigned to rating class 2. The classifier is able to detect 99.32% of all images that are of unacceptable quality. Thus, only 0.68% of the vehicles with non-custom HUDs would be wrongly sent to the customers.

In the final step of the investigation, the 2 threshold conditions are combined. An autonomously labelled image is transferred into the training dataset if both conditions are fulfilled at the same time. The obtained results on the test set are shown in the appendix *A.13* see Figure 179 and Figure 180. For initial training set sizes greater than 10%, classification results are obtained that are very similar to the “ideal” values. Threshold

parameters of $\theta_1 = 0.8$ and $\theta_2 = 0.3$ and initial training set sizes greater than 20% give TPR values that exceed the desired “ideal” value. Likewise, if the initial training set size is set to 15% or 20%, and threshold parameters of $\theta_1 = 0.6$ and $\theta_2 = 0.1$ are applied, accuracies exceeding the “ideal” value are obtained. The learning curves for an initial training set size of 20% and threshold parameters of $\theta_1 = 0.8$ and $\theta_2 = 0.3$ are shown in the appendix *A.14* see Figure 183. The learning success occurs in the last training cycles. After the SSL process terminates the recognition accuracy and the TPR exceed the “ideal” value and the FPR falls below it. By combining the 2 threshold conditions, good classification results can be achieved, which are very similar to the “ideal” values or even better. Thus, the manual label costs can be reduced.

In the end, the allocation task is again limited to the 2 possibilities acceptable or unacceptable. The investigation is performed with a 2-class PC of 2nd order in the 2-principal component space. If all manually labelled training images are available, the determined ROC curve covers the area of 0.998. The learning behaviour of this 2-class classifier is investigated for various initially labelled training set sizes. An autonomously labelled training image is transferred into the training set if the class assignment probability is larger than 80% ($\theta_1 = 0.8$). Here, the constant ϕ is varied between 0.1 and 10, and the increase is done in steps of 0.1. The resulting AUC values after completing the SSL process are shown in Figure 122 for 30 different runs.

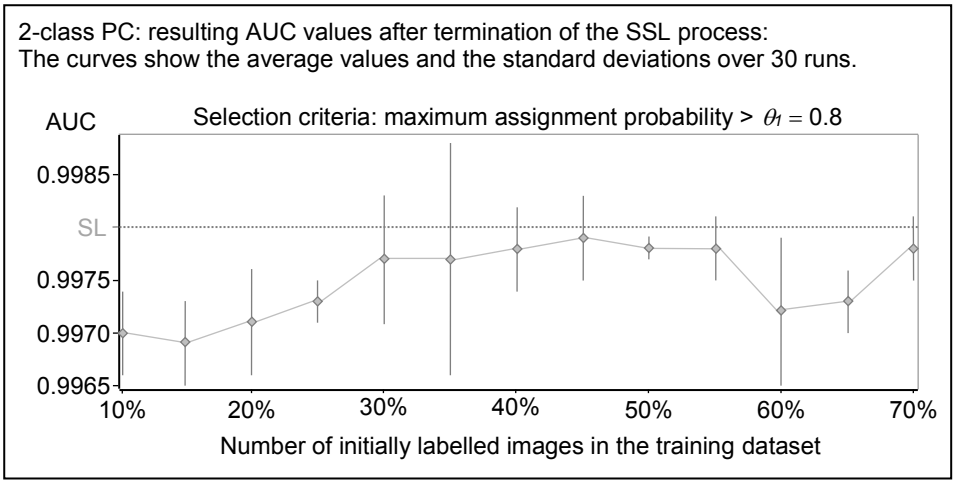


Figure 122: SSL, 2-class PC: results for the double image dataset
$\theta_1 = 0.8$

The investigation shows that AUC values greater than 0.996 are obtained for all initial training set sizes. Thus, the obtained values are very similar to the “ideal” value, but this value is not exceeded on average. Only for individual runs, the 0.998 value is exceeded. Overall, this investigation shows that the manual label effort for training the 2-class PC can be drastically reduced without a significant loss of classification accuracy.

K-nearest neighbour classification:

The SSL procedure is applied to the NN classifier for $k = 3$ in the 7-dimensional component space. If all manually labelled training images are available, the RMSE of 0.65, the accuracy of 95.28%, the FPR of 0.68% and the TPR of 92.49% are obtained.

The implementation of the SSL algorithm corresponds to the procedure described in chapter *7.5.1*. First, the threshold distances for each rating class are determined from the test images. Since this classifier considers 3 nearest neighbours, the threshold distances are determined from the distances between the unlabelled image and the nearest training images of the majority of the 3 nearest neighbours. The used threshold distances are summarised in Table 36.

Subjective rating class	1	2	3	4	5
$\theta =$ minimal distance [HUD pixel]	0.842	0.973	0.164	0.303	0.338
$\theta =$ mean distance [HUD pixel]	1.407	1.780	1.490	1.330	0.636
$\theta =$ maximal distance [HUD pixel]	2.168	2.599	3.042	2.736	2.576

Table 36: kNN for the double image dataset: used threshold distances

During the SSL process, an autonomously labelled training image is transferred into the training set if the distance between the images and the corresponding reference image is less than the threshold distance. The results obtained for different initial training set sizes are shown in the appendix *A.15* see Figure 184 and Figure 185. On the double image dataset, good classification results can be achieved after the SSL process is completed. For initial training set sizes greater than 25%, hardly any difference is noticeable between the classification results and the different threshold distances. The recognition accuracy and the TPR values increase with the number of labelled images in the initial training set. The “ideal” accuracy and the “ideal” TPR are not achieved. The same applies to the average RMSE values, which decrease with the number of manually labelled training images. Likewise, the “ideal” RMSE value is not achieved. The average FPR values are in the lower percentage range and are very similar to the “ideal” value. A particularity is observable for initial training set sizes greater than 20% and the minimum threshold distances. Here, FPR values are obtained that correspond to the “ideal” value for each of the 30 runs.

With an initial training set size of 35% and a maximal threshold distances, it is shown exemplarily that a learning success can hardly be determined during the SSL process. The corresponding learning curves are shown in the appendix *A.16* see Figure 186. The process terminates automatically after 20 cycles. During the SSL process, the average values barely change. Despite well-chosen threshold distances, the “ideal” values are not reached. The reason could be that feature vectors that are incorrectly labelled are transferred into the training set. Thus, the success rate of the classifier cannot improve.

A possible labelling of the test images for a single run after the SSL process completed is shown in Figure 123. Not a single test image is assigned to rating class 5. Images

that are manually labelled with 5 rating points are completely assigned to rating class 4. Likewise, test images that are manually labelled with 1 rating point are assigned to rating classes 1 and 2. However, these false classifications are uncritical. The obtained false positive rate of 0.68% corresponds to the “ideal” value. It results from the fact that some test images labelled manually with 2 rating points are assigned to rating classes 3 and 4. In contrast, some test images, which belong subjectively to rating class 3 and 4, are assigned to rating class 2. This fact reduces the true positive rate to 91.08%. The obtained accuracy of the assigned labels is 91.94%.

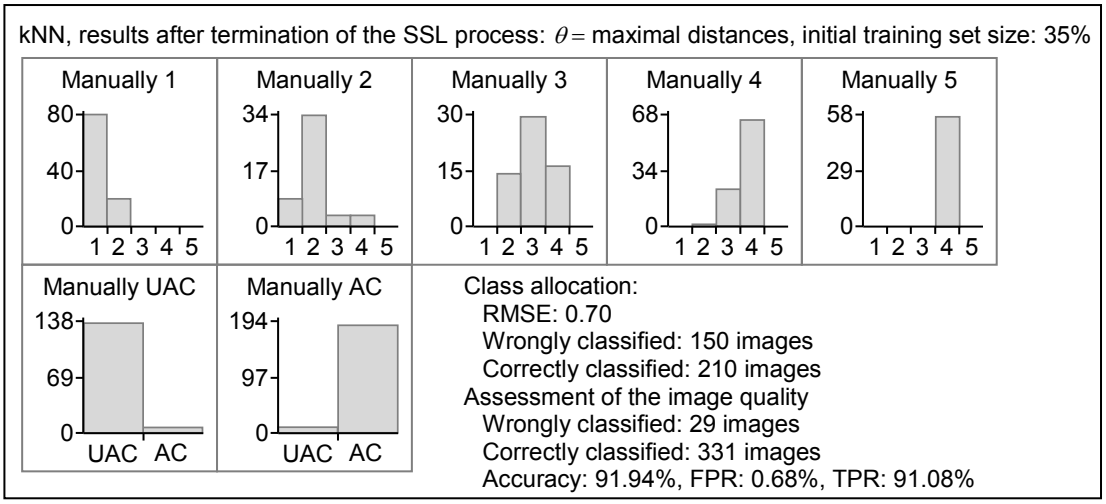


Figure 123: SSL, kNN for the double image dataset: possible labelling of the test images

Finally, the learning behaviour of the 2-class NN in the 7-dimensional component space is investigated. An AUC value of 0.96 can be determined if all manually labelled training images are used for training and the number of nearest prototype vectors k is set to 30. The learning behaviour of the classifier is investigated for different sized initially labelled training dataset. As already announced, an autonomously labelled image is transferred into the training dataset if the distance to the nearest reference pattern is smaller than a predetermined threshold distance θ, which is set to the maximal threshold distances. To determine the ROC curve the number of images *m* required to classify a test image as AC is varied between 1 and k. Figure 124 shows the average AUC value and the standard deviations over 30 runs after the SSL process has ended. The x-axis shows the number of initially labelled training images and the y-axis the resulting AUC values. The more manually labelled training images are available, the greater the resulting AUC value. Unfortunately, the “ideal” value is not reached. The manual labelling effort can only be reduced if a slight decrease in the AUC values can be accepted.

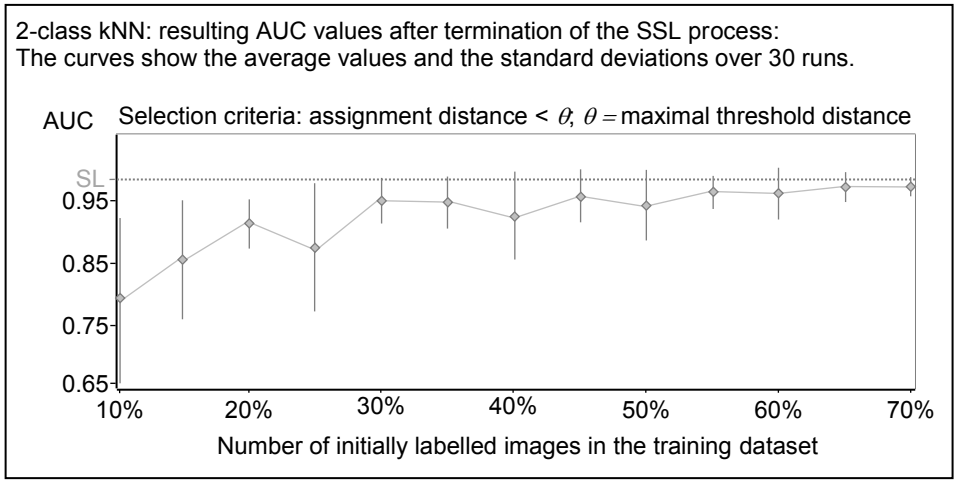


Figure 124: SSL, 2-class kNN: results for the double image dataset
θ = maximal threshold distance

Learning vector quantisation:

The SSL procedure is applied to the LVQ that uses 69 prototypes, which are trained by learning rule LVQ2.1, to approximate the underlying distribution of labelled training images. For all subsequent investigations, the same primarily initialisation of the prototypes is used, which reaches for all manually labelled training images the RMSE of 1.16, the accuracy of 83.06%, the FPR of 33.33% and the TPR of 94.37%.

Since the LVQ assigns an unlabelled image to the same class as the nearest prototype, the implementation of the SSL algorithm follows the known scheme. The threshold distances used, which are determined from test images, are shown in Table 37.

Subjective rating class	1	2	3	4	5
θ = minimal distance [HUD pixel]	0.160	0.564	0.397	0.147	0.000
θ = mean distance [HUD pixel]	0.370	0.869	0.837	0.521	0.172
θ = maximal distance [HUD pixel]	0.851	1.276	1.179	1.061	0.963

Table 37: LVQ for the double image dataset: used threshold distances

The shown distances are successively used as thresholds during the SSL process. The classification results on the test images are shown in the appendix *A.17* see Figure 187 and Figure 188. Upon completion of the SSL process, the resulting recognition accuracies are similar to the “ideal” value. For occasional runs, the value of 83.06% is even exceeded. In addition, no direct relationship can be determined between the number of initially labelled training images and the classification accuracy. Likewise, the average RMSE values are very similar to the “ideal” value, which is sometimes undercut. In contrast, the average TPR values, which vary between 94.60% and 97.42%, are higher than the “ideal” value. At the same time, the classifiers achieve FPR values higher than 33.33%.

The learning behaviour of the classifier is analysed exemplarily more precisely for an initial training set size of 25%. Here, a newly labelled training image is only accepted if the allocation distance is less than the minimum threshold distance, as shown in Figure

189 in the appendix *A.18*. During the SSL process, the number of labelled training images is increased from 87 to 167. The average TPR values and average FPR values barely change in the 8 training cycles. In contrast, the average RMSE values decrease by increasing the number of labelled training images. The average accuracy values show that the maximum gain achieved is 0.26%. For 7 of the 30 runs, the accuracy obtained after completing the SSL process exceeds the “ideal” value.
The investigation is completed by a detailed analysis of 1 of the 7 runs for which the “ideal” accuracy is exceeded. A possible labelling of the test images is shown in Figure 125. The high FPR is due to the positive labelling of many test images that represent an unacceptable quality. Here, test images that are manually labelled with 1 or 2 rating points are assigned mainly to rating class 4. In contrast, test images manually labelled with 3, 4 or 5 rating points are primarily assigned to the rating classes representing an acceptable quality. Overall, the recognition accuracy and the TPR are 1.94% and 4.22% higher than the corresponding “ideal” values. Therefore, the SSL algorithm tries to select labelled training images that beneficially affect the classification performance. Here, the manual label effort can be significantly reduced without a loss of classification accuracy.

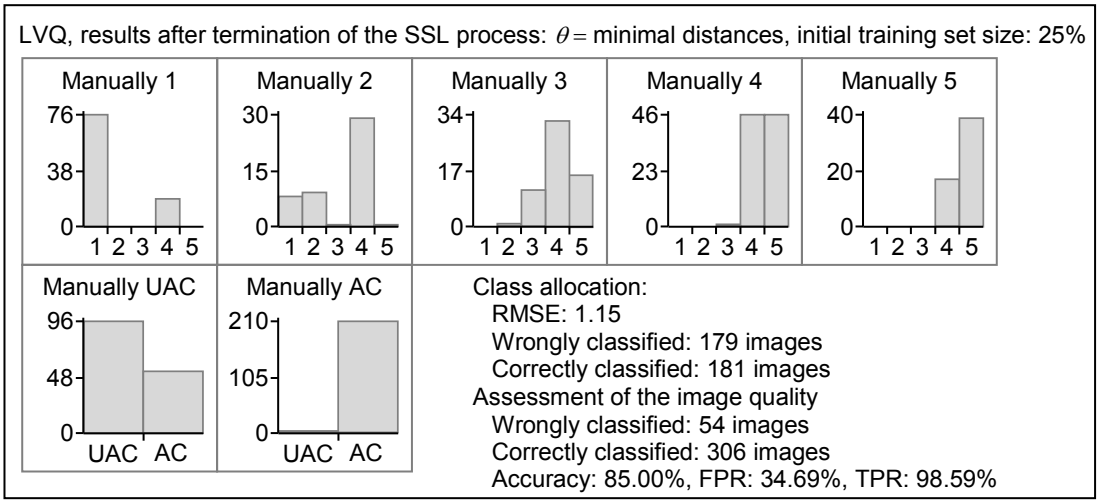


Figure 125: SSL, LVQ for the double image dataset: possible labelling of the test images

Finally, the SSL learning behaviour of a 2-class LVQ is investigated. The 2-class classifier uses 69 prototype vectors, which are determined from the manually and autonomously labelled training images. The minimal threshold distances θ are used to check whether the autonomously labelled image can be used for training. After each training cycle, the AUC value is determined by varying the number of images *m* needed to classify a test image as AC. After the SSL processes have finished, the average AUC values and the standard deviations over 30 runs are shown in Figure 126. If all manually labelled training images are used for training, an area under the ROC curve of 0.99 can be determined. The investigation reveals that AUC values greater than 0.70 are obtained for all training set sizes. Thus, the values obtained are well below the “ideal” value. If the manual labelling effort is reduced, the classification quality decreases.

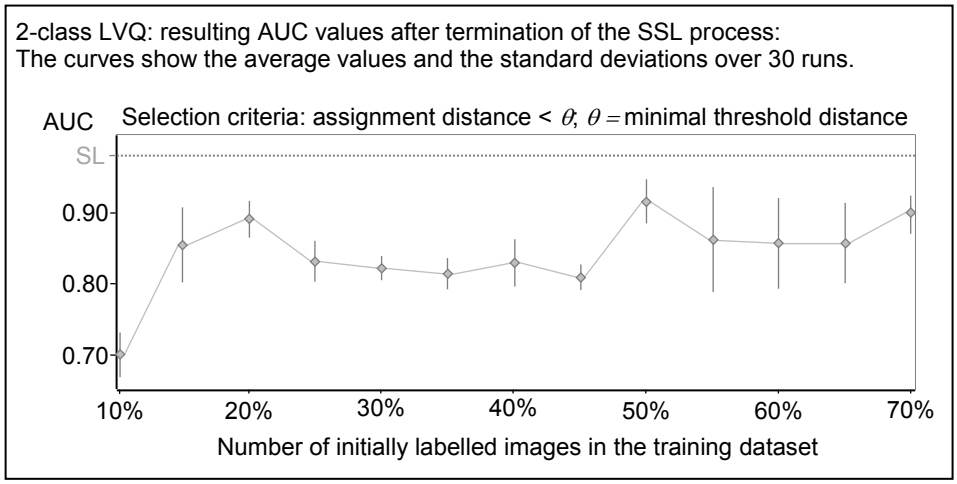


Figure 126: SSL, 2-class LVQ: results for the double image dataset
θ = minimal threshold distance

Summary of the obtained results:

The experimental evaluation shows that for a small number of manually labelled training images some classifiers are able to increase autonomously the recognition performance to a level that exceeds the “ideal” performance. Using the LVQ classifier, the manual label effort can be reduced significantly, without any loss of classification quality. Depending on the threshold and the initial training set size, the PC yields a very high accuracy of more than 98%. During the semi-supervised learning process, the algorithms select those samples that improve the classification performance in an advantageous manner. Unfortunately, the NN classifier shows hardly any learning success during the SSL process. Nevertheless, this classifier type still obtains good classification results after the SSL process is finished. If the allocation task is restricted to acceptable or unacceptable, the 2-class classifiers are able to obtain AUC values that are below the “ideal” value. Here, the manual labelling effort can only be reduced if a decrease in the AUC values can be accepted.

7.5.3 SSL: assessment algorithms for the distortion and double image dataset

During the SSL process, the classifiers are trained with a manually labelled training set of variable size. This training set contains at least 1 training image from each rating class, as shown in Table 38. A new training cycle is initiated if 9 images are autonomously labelled by the SSL algorithm and inserted into the training dataset. After each training cycle, the classification results on the test set are compared with the results of an “ideal” classifier trained with all available manually labelled training samples.

Initial number of labelled training images	Classes					Overall
	1	2	3	4	5	
10%	1	13	14	4	1	33
15%	1	20	20	5	1	47
20%	1	26	27	7	1	62
25%	2	33	34	8	1	78
30%	2	40	40	10	1	93
35%	2	46	47	11	1	107
40%	2	53	54	13	1	123
45%	3	59	60	14	1	137
50%	3	66	67	16	1	153
55%	3	73	74	18	1	169
60%	4	79	80	19	1	183
65%	4	86	87	21	1	199
70%	4	92	94	22	1	213

Table 38: SSL for the distortion and double image dataset: initial number of labelled training samples

Polynomial classification:

The SSL algorithm is applied to a PC of 1st order in the 20-dimensional component space. If all manually labelled training images are available, the RMSE of 0.67, the accuracy of 71.48%, the FPR of 31.88%, and the TPR of 74.25% are obtained.

Subsequently, the autonomously labelled images have to fulfil different conditions to be transmitted to the training dataset. First, the maximum probability that the image belongs to the rating class must be greater than a threshold θ_1, which is set to 0.6, 0.7, and 0.8. Afterwards, the difference between the largest and the second largest class probability must be greater than the threshold θ_2, which is set to 0.1, 0.2, and 0.3. Finally, both conditions are combined so that both conditions must be met simultaneously. The classification results obtained after the completion of the SSL process are summarised in the appendix *A.19* see Figure 190 to Figure 195. Reducing the number of manually labelled training images also reduces the recognition accuracies after the termination of the SSL process. The “ideal” accuracy of 71.48% is not achieved when fewer manually labelled images are used. Similarly, the TPR depends on the number of manually labelled training images. With a few exceptions, fewer manually labelled images lead to decreasing values. If the initially labelled training set size is set to 55% and the threshold θ_1 is set to 0.6 or θ_2 is set to 0.1, an average TPR exceeding the “ideal” value is reached. The same applies to an initial training set size of 25% in combination with $\theta_2 = 0.2$ or $\theta_1 = 0.6$ and $\theta_2 = 0.1$. The RMSE and the FPR values decrease by increasing the number of initially labelled training images. The average RMSE values fall not below the “ideal” value. In contrast, the “ideal” FPR value is on average undercut by certain combinations of initially manually labelled training set sizes and threshold values. This is the case for an initial training set size of 45% in combination with a threshold value $\theta_1 = 0.7$ or $\theta_1 = 0.7$ and $\theta_2 = 0.2$, and an initial training set size of

55% in combination with a threshold value $\theta_1 = 0.8$ or $\theta_1 = 0.8$ and $\theta_2 = 0.3$. To summarise, if the manual labelling effort is reduced, a loss of accuracy must be accepted. Less manually labelled images do not contain enough information to generalise to unknown data.

The learning behaviour of the PC is exemplarily investigated for an initial training set size of 35%, 15%, and 30% in combination with threshold values $\theta_1 = 0.7$, $\theta_2 = 0.3$, and $\theta_1 = 0.8$ and $\theta_2 = 0.3$. The learning curves are shown in the appendix *A.20* see Figure 196 to Figure 198. During the semi-supervised learning process, the recognition accuracies, as well as the TPR values, may decrease if too many false-labelled images are added to the training dataset. Similarly, the RMSE values and the FPR values may increase, indicating that the 1st order classifier is not complex enough to extract sufficient information to generalise to unknown data. The learning success is achieved only at the end of the SSL process for an initial training set size of 30% and a threshold combination of $\theta_1 = 0.8$ and $\theta_2 = 0.3$. The same applies to an initial training set size of 35% and a threshold value $\theta_1 = 0.7$. The experimental evaluation shows that reducing the manual label effort results in a reduction in the recognition accuracy. Dependent on the selected threshold, it is either possible to get either a high TPR or a low FPR.
A possible labelling of the test images is exemplified for a single run and an initial training set size of 35% and a threshold value $\theta_1 = 0.7$, as shown in Figure 127.

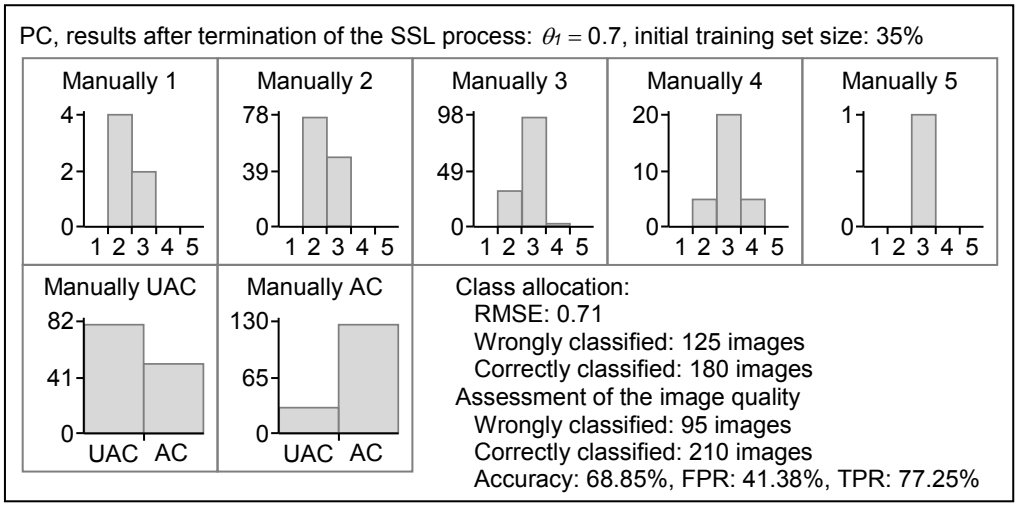


Figure 127: SSL, PC for the distortion and double image dataset: possible labelling of the test images, θ_1

The learning process ends after 12 training cycles in which the number of training samples is increased from 107 to 215. Images that are manually labelled with 3 or 4 rating points are assigned to the associated class, but also to rating class 2. The classifier obtains a TPR of 77.25%, which is 3% higher than the “ideal” value. Therefore, the classifier trained by the SSL procedure would incur fewer costs, since fewer vehicles would be mistakenly sent to rework. In contrast, the accuracy of 68.85%, the FPR of 41.38%, and the RMSE value of 0.71 do not achieve the corresponding “ideal” values.

The reason is that images that are manually labelled with 1 or 2 rating points are assigned to rating classes 2 and 3 by the classifier.

Finally, the learning behaviour of the 2nd order 2-class classifier in the 2 principal component space is investigated exemplarily. Here, the 5 class problem is reduced to a 2-class problem where only the assessment of the image quality as acceptable or unacceptable is considered. If all manually labelled images are available, it is possible to obtain the ROC curve that corresponds to an AUC value of 0.68. The resulting AUC values after the completion of the SSL process, which uses a threshold of $\theta_1 = 0.7$ and various manually labelled training set sizes, are shown in Figure 128.

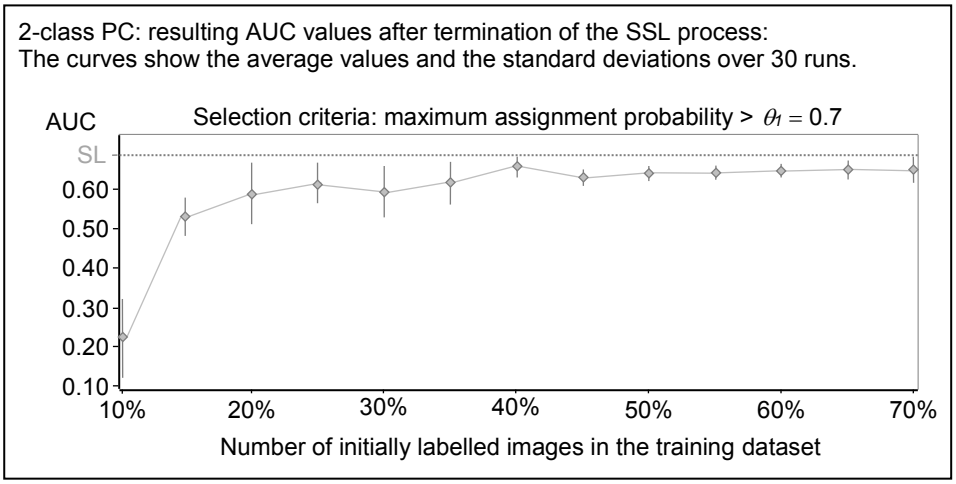


Figure 128: SSL, 2-class PC: results for the distortion and double image dataset, $\theta_1 = 0.7$

If the size of the initial training set is set to 10%, an average AUC value of 0.22 is obtained. This low value indicates that the test images are mainly wrongly classified. Perhaps the classifier approximates random noise rather than the correlation between the feature vectors and the manual labels. The AUC values increase with the number of manually labelled images in the training set. For an initial training set size of 15%, an average AUC value of 0.53 is already obtained. This value refers to a random process where the FPR is equal to the TPR. The AUC values average over 30 runs become very similar to the “ideal” value for initial training set sizes greater than 35%. However, the value of 0.68 is not exceeded. The largest AUC value of 0.66 is achieved on average for initial training set sizes of 40%. The increase in the classification performance during the SSL process is rather low.

K-nearest neighbour classification:

The SSL approach is applied to the NN classier. In the 13-dimensional component space, the unlabelled feature vector is assigned to the same class as the nearest labelled training image belongs. If all manually labelled training images are used for training, the RMSE of 0.43, the accuracy of 91.48%, the FPR of 11.59%, and the TPR of 94.01% are obtained. The used threshold distances for accepting or rejecting the au-

tonomously generated class label are summarised in Table 39. The determination of the thresholds is based on the same principle as described in the previous text.

Subjective rating class	1	2	3	4	5
$\theta =$ minimal distance [HUD pixel]	0.565	0.264	0.189	0.394	0.716
$\theta =$ mean distance [HUD pixel]	1.144	0.756	0.821	0.817	1.096
$\theta =$ maximal distance [HUD pixel]	1.463	1.952	2.153	1.708	1.331

Table 39: kNN for the distortion and double image dataset: used threshold distances

The corresponding classification results after completion of the SSL processes are shown in the appendix *A.21* see Figure 199 and Figure 200. The proposed semi-supervised learning approach yields accuracies and TPR values that increase with the number of manually labelled training images. In contrast, the obtained RMSE values and the FPR values decrease on average by increasing the size of the initially labelled training set. Unfortunately, the “ideal” values are not reached.

The learning behaviour of the semi-supervised learning scheme is exemplified in the appendix *A.22* for an initial training set size of 35% and a mean threshold distances, shown in Figure 201. During the SSL process, the size of the training set increases from 107 to 179 labelled images. It can be seen that the recognition accuracy hardly changes in the first 7 training cycle until the accuracy in the last cycle increases slightly. The gain in accuracy is low and is only 0.82%. Likewise, the TPR increases by 4.65% in the last training cycle. In contrast, the FPR increases slightly in each training cycle, indicating to wrongly labelled training images. A possible labelling of the test images is shown for a single run in Figure 129.

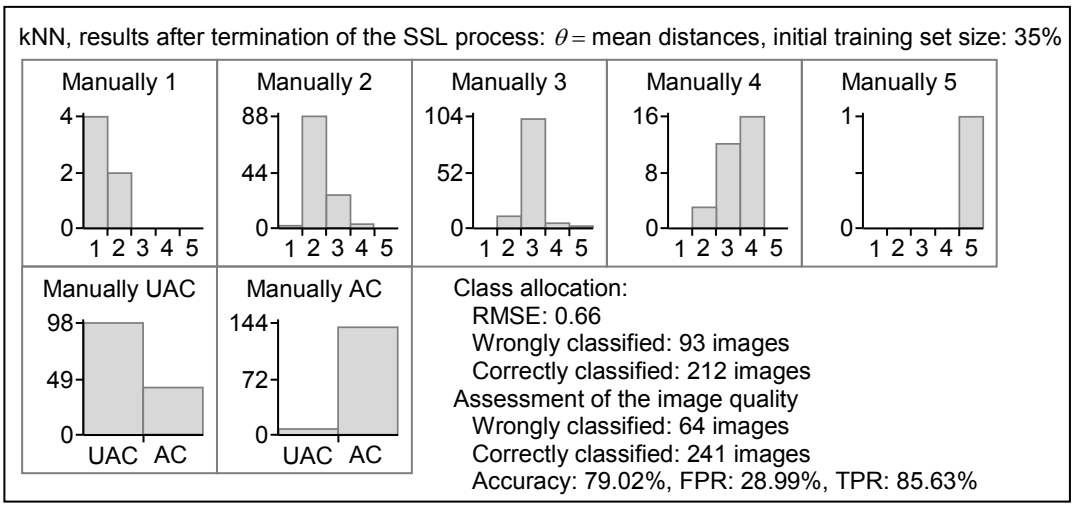


Figure 129: SSL, kNN for the distortion and double image dataset: possible labelling of the test images

The proposed SSL approach yields an RMSE that is 0.23 higher than that of a classifier trained with all manually labelled training images. In addition, the accuracy, the FPR and the TPR are far from the “ideal” value. A few test images that are manually labelled with 2 rating points are incorrectly assigned to rating classes 3 and 4, resulting in the high FPR of 28.99%. In contrast, a few test images manually labelled with 3 or 4 points are mistakenly assigned to rating class 2. Thus, the TPR drops to 85.63%. Since the “ideal” values are not exceeded on average, the few manually labelled training images do not contain enough information to generalise well to unknown data. Choosing the most appropriate training set is always a compromise between the manual labelling effort and the classification performance.

Finally, the mapping rule of the NN is reduced to a 2-class problem. The learning behaviour is investigated for various initially labelled training dataset sizes in the 13-dimensional component space. If all manually labelled training images are available, the classifier achieves an AUC value of 0.86. For this purpose, the number of nearest prototype vectors k is set to 10. An autonomously labelled image is transferred into the training dataset if the distance to the nearest reference pattern is less than the mean assignment distance θ. To determine the ROC curve the number of images *m* required to classify a test image as AC is varied between 1 and k. Figure 130 shows the average AUC values and the standard deviations achieved over 30 runs.

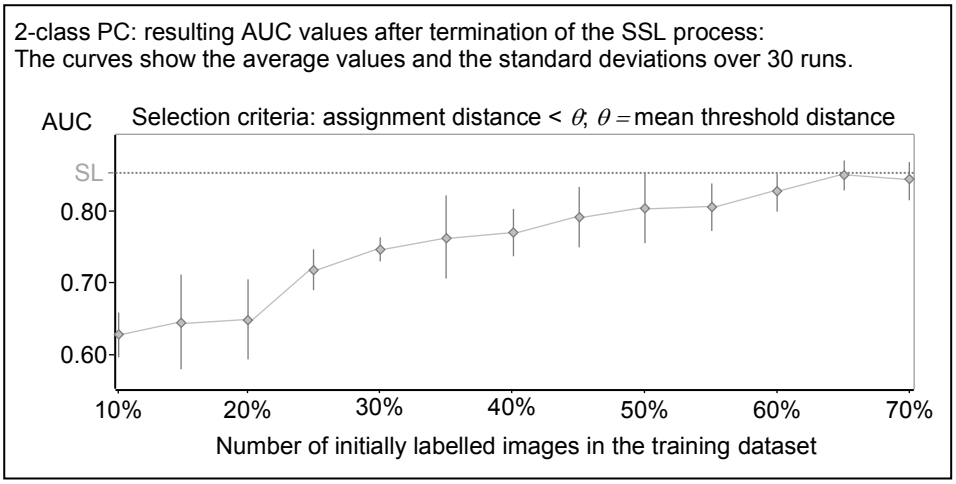


Figure 130: SSL, 2-class kNN: results for the distortion and double image dataset, θ = mean threshold distance

The diagram shows that the classification quality depends on the number of manually labelled training images. The less manually labelled training images are available, the lower the resulting AUC value. The manual labelling effort can only be reduced if a decrease in the AUC value can be accepted.

Learning vector quantisation:

The learning behaviour of the LVQ is investigated. Exemplary, 103 prototypes determined by the LVQ1 learning rule are used to approximate the underlying distribution of

labelled training images. For all investigations, the same prototype initialisation is used. If all manually labelled training images are available, the RMSE of 0.59, the accuracy of 85.90%, the FPR of 13.77%, and the TPR of 85.63% are obtained.
The determination of the threshold distances for transferring a labelled image into the training set is based on the same principle as described in the previous text. The threshold distances used are summarised in Table 40. For rating class 5, no threshold distances can be determined. Since rating class 5 is not assigned to a test pattern, no threshold distance can be determined from the non-existing distances between the test images and the nearest reference pattern.

Subjective rating class	1	2	3	4	5
$\theta =$ minimal distance [HUD pixel]	0.982	0.324	0.401	0.562	---
$\theta =$ mean distance [HUD pixel]	1.695	1.209	1.143	1.248	---
$\theta =$ maximal distance [HUD pixel]	3.381	2.785	2.476	2.436	---

Table 40: LVQ for the distortion and double image dataset: used threshold
used threshold distances

The results for various initial training set sizes after the termination of the SSL process are summarised in the appendix *A.23* see Figure 202 and Figure 203. It turns out, that the classification performance increases with the number of initially labelled training images. The average recognition accuracies and the average TPR values initially become very similar to the corresponding “ideal” values and then exceed these values by increasing the number of training images. The larger the threshold distances are chosen, the less manually labelled training images are needed to exceed the “ideal” values. Similarly, the average FPR values decrease as the number of manually labelled images increases, but the “ideal” value is not reached on average. Apparently, the SSL algorithm is able to avoid the selection of poorly labelled training images and selects those images that beneficially affect the classification performance.

The learning behaviour of the classifier is investigated as an example for an initial training set size of 35% and mean threshold distances. The learning curves are shown in the appendix *A.24* see Figure 204. In 17 training cycles, the number of labelled training images is increased from 107 to 260. The average RMSE values decrease on average about 0.06. For 8 of 30 runs the average RMSE values fall below the “ideal” value. Likewise, the average FPR values decrease on average by 3.48%. In 4 out of 30 runs, the average FPR values are below 13.77%. In contrast, the average accuracy values increase with each training cycle where the average gain amounts 3.44%. After completing the SSL process, the accuracy values exceed 85.90% for each run. For each training cycle, the average TPR values are higher the “ideal” value and the average gain amounts 5.33%.
For a single run, the resulting class assignments of the test images are exemplified in Figure 131. The SSL approach results in a labelling of the test images, which is very similar to the labelling of the classifier with all manually assigned class labels. The test

images are assigned mainly to the corresponding rating class. According to a few wrongly assigned test images, the accuracy of 86.56% and the TPR of 89.22% are obtained, which are 0.66% and 3.59% higher than the “ideal” values. This indicates that fewer test images that are manually assigned to rating classes 3 or 4 are incorrectly assigned to rating class 2 and 1. In contrast, more images that are manually assigned to rating class 2 are incorrectly assigned to rating class 3 and 4. Thus, the FPR of 16.67% and the RMSE value of 0.61 are obtained.

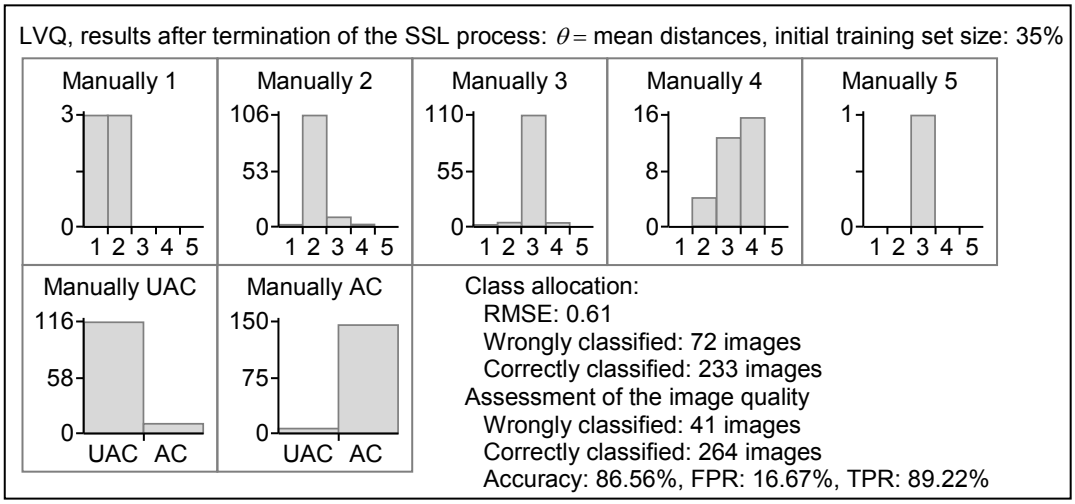


Figure 131: SSL, LVQ for the distortion and double image dataset: possible labelling of the test images

This evaluation shows that the SSL algorithm can achieve higher recognition accuracies with less manually labelled test images than the “ideal” classifier. The manual label effort can be reduced without any loss of classification performance.

At the end, the semi-supervised learning behaviour of the 2-class LVQ is investigated. The classification is based on 103 prototypes, which are determined by the LVQ1 learning rule. The mean threshold distances θ are used to check if the autonomously labelled images can be used for training. After each training cycle, the FPR and TPR value are calculated by varying the number *m* of the *k*-nearest prototype vectors that are required to classify a test image as AC. After completing the SSL process, the resulting AUC value is determined. The number of considered nearest prototype vectors is set to k = 10. The development of the AUC values is shown in Figure 132. If all manually labelled training images are used for training, an AUC value of 0.71 can be determined. Even with only 10% manually labelled training images, average AUC values are determined which are higher than the desired value. Consequently, it is possible to develop a 2-class LVQ with a small number of manually labelled images, which might be better suited to assess the HUD image quality than the “ideal” neural net. The manual label effort can be reduced since the SSL process tries to avoid the use of poorly labelled training samples and selects those samples that improve the recognition behaviour.

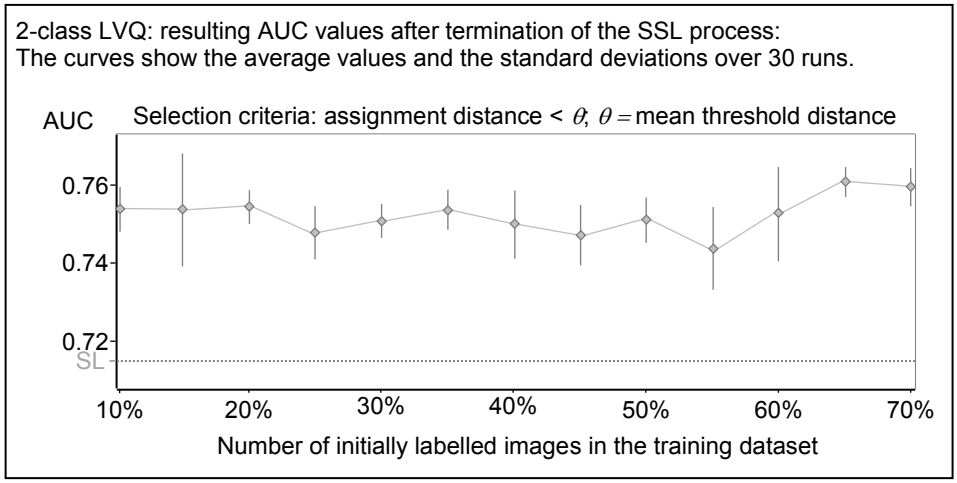


Figure 132: SSL, 2-class LVQ: results for the distortion and double image dataset, $\theta =$ mean threshold distance

Summary of the obtained results:

For the distortion and double image dataset, the SSL procedure may result in the LVQ classifier that obtains an accuracy that exceeds the “ideal” value. This shows that the proposed algorithm selects those labelled training images that improve the recognition behaviour. For example, upon completion of the SSL process, the accuracy of 87.54% is obtained using an initial training set size of 35%. Even if classification problem is reduced to a 2-class assignment task, the LVQ algorithm trained with a small number of manually labelled images is able to assess the HUD image quality better than the “ideal” classifier. In contrast, for the polynomial classifier and the nearest neighbour classifier, a loss of the classifier accuracy must be accepted if the manual labelling effort is reduced. The reason could be that fewer manually labelled images do not contain sufficient information to generalise to unknown data. If the mapping rule of the PC and the kNN is reduced to a 2-class problem, the experimental evaluation shows that the manual label effort can be reduced only if minimal deviations in the ROC curve can be accepted.

7.6 Active learning: impact on the classification results

The previous chapter shows that the number of labelled training samples can be reduced by using the semi-supervised learning approach. Therefore, it is now checked whether this can also be achieved by active learning (AL). Active learning differs from supervised learning by the fact that the learning algorithm can select the data from which it learns, see chapter *6.2.4*. The algorithm learns from data that are most informative or for which the classifier predicts the wrong label with a high probability. These images are manually labelled by the Oracle and transferred into the training dataset [PERSELLO & BRUZZONE 12], [YEH & GALLAGHER 08]. The assessments of the test persons serve as Oracle labels. The active learning rule is applied to selected classifiers that are designed for the distortion dataset, the double image dataset, and the distortion and double image dataset. To select the most informative training images,

the following selection criteria are defined for the polynomial classifier, the nearest neighbour classifier and learning vector quantisation:

Polynomial classification:

Similar to the SSL process, the following 2 selection criteria, shown in Figure 133, are defined that determine whether the training image needs to be labelled by the Oracle:

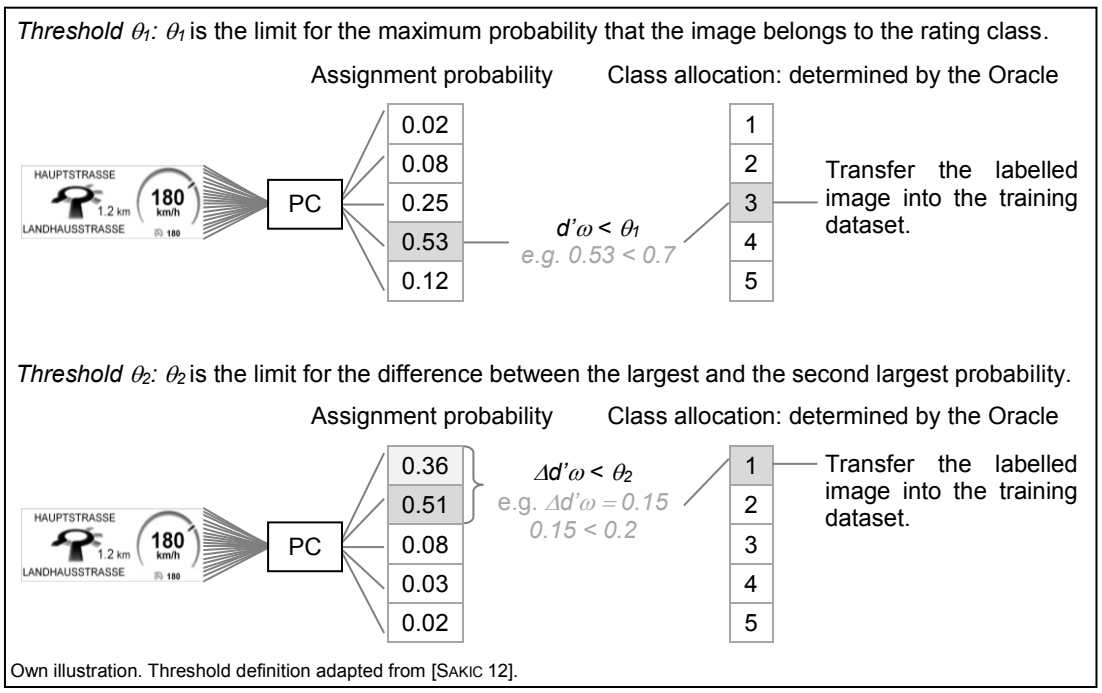


Figure 133: AL, PC: threshold value definition

- *Threshold* θ_1: θ_1 is the limit for the maximum probability that the image belongs to the corresponding rating class. If the maximum assignment probability is less than the threshold, the image is labelled manually by the Oracle and transferred into the training dataset.

- *Threshold* θ_2: θ_2 is the limit for the difference between the largest and the second largest class probability. If this probability difference is less than the threshold, the image needs to be labelled by the Oracle to be included in the training set.

K-nearest neighbour classification:

The largest gain of information is expected from images located in the region of the largest uncertainty [CEBRON 08: p. 67]. It is assumed that the further away the unlabelled image is to the corresponding labelled training sample, the higher the probability of uncertainty [WANG et al. 11]. Therefore, the following selection criterion is defined for the classification with the kNN algorithm:

- *Threshold* θ: θ is the limit for the maximum distance to the next training image. If the distance to the next training sample is greater than the maximum distance

given by the threshold θ, the class assignment is uncertain and the image needs to be labelled by the Oracle, shown in Figure 134.

Learning vector quantisation:
Unlabelled images between 2 prototypes provide the largest gain of information [CEBRON 08: p. 67]. Therefore, the following criterion is defined for querying the Oracle:

- *Threshold θ:* θ is the limit for the maximum distance to the next prototype. If the distance to the next prototype is greater than the maximum distance given by the threshold θ, the image is labelled by the Oracle, shown in Figure 134.

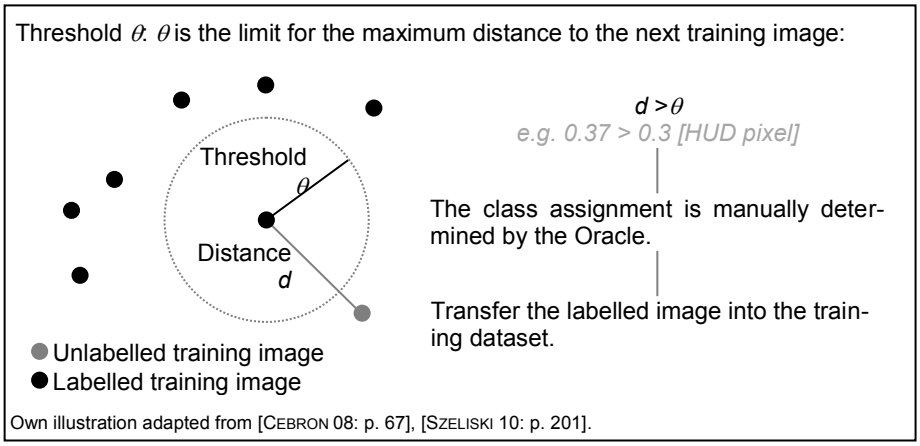


Figure 134: AL, kNN and LVQ: threshold value definition

7.6.1 AL: assessment algorithms for the distortion dataset

During the AL procedure, the classifiers are trained with a manually labelled training set of variable size. Here, the size of the initial training set is to 0.5%, 2%, and 5% of all training images from each rating class. Again, it is ensured that at least 1 image is selected from each rating class. The numbers of initially labelled images are summarised in Table 41.

Initial number of labelled training images	Classes					Overall
	1	2	3	4	5	
0.5%	2	3	1	1	1	8
2%	8	11	1	1	1	22
5%	20	28	2	1	1	52

Table 41: AL for the distortion dataset: initial number of labelled training samples

After the classifiers are trained with the initially labelled images, the classifiers check for each remaining training image if the class allocation is predicted wrongly with a high probability. If so, the Oracle is asked for the right label and the image is added to the labelled training dataset. Here, a new training cycle is initiated if 25 images are manually labelled by the Oracle. After each training cycle, the classifiers are applied to label

the test images. Thus, the learning successes of the classifiers can be monitored. This is repeated until all informative training samples are labelled by the Oracle and then the process ends automatically.

Polynomial classification:
The active learning scenario is applied to a PC of 2nd order in the 10-dimensional component space. If the classifier is trained with all manually labelled images the RMSE of 1.12, the accuracy of 83.89%, the FPR of 39.31%, and the TPR of 99.53% is achieved. First, a training image is labelled by the Oracle if the maximum probability that the image belongs to the corresponding rating class is less than the threshold θ_1. Here, θ_1 is set to 0.6, 0.7, and 0.8. The classification results after the termination of the AL process are summarised in the appendix *A.25* see Figure 205 and Figure 206. The experimental evaluation shows that the larger the threshold value is chosen, the more images need to be labelled manually. If a threshold value of 0.8 is applied, the active learning procedure results in a better classification performance than that of a classifier trained with all manually assigned class labels. The same applies to a threshold of 0.7 and initial training set sizes of 2% and 5%, or a threshold value of 0.6 and an initial training set size of 5%. Thus, the AL algorithms select those training images that improve the classification performance. If the size of the initial training set is set to 2%, a TPR of 100% is reached after each run.

As an example, the learning curves of the classifier for the initial training set size of 2% and a threshold value of 0.7 are shown in the appendix *A.26* see Figure 211. The AL process results in 21 training cycles and increases the number of labelled images from 22 to 547. In the course of the AL process, the accuracies and the TPR values increase. Likewise, the RMSE values and the FPR values are decreasing and are below the corresponding “ideal” values after 16 training cycles. Here, the average accuracy values exceed the “ideal” value after 16 training cycles and the average gain in accuracy is 26.67%. The average TPR values exceed the “ideal” value after 5 training cycles and give the TPR of 100% after each run.
For a single run, the resulting labels of the test images are shown in Figure 135. As expected, the TPR is 100%, indicating that all images that are manually labelled with 3, 4 or 5 rating points are assigned to these classes. An accuracy of 86.11% is achieved that is 2.22% higher than that of a classifier using all manually assigned class labels. Test images that are manually labelled with 2 rating points are also assigned to rating classes 3, 4 and 5 giving the FPR of 34.48%. It becomes clear that not all labelled images are needed to achieve a better classification performance. This can be explained by the selection of the most informative training samples.

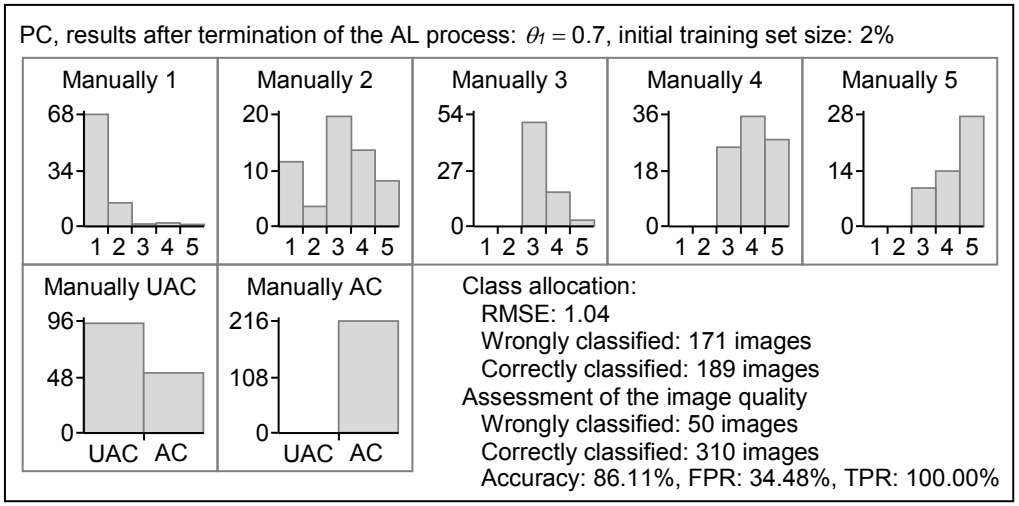


Figure 135: AL, PC for the distortion dataset: possible labelling of the test images, $\theta_1 = 0.7$

In the second step, the Oracle is asked for the correct label if the difference between the largest and the second largest class probability is less than a threshold θ_2, which is set to 0.1, 0.2, and 0.3. The classification results obtained after the termination of the AL processes are shown in the appendix *A.25* see Figure 207 and Figure 208. The larger the threshold value is chosen, the more images need to be labelled manually. Regardless of the threshold and the size of the initial training set, the proposed AL approach yields average recognition accuracies that exceed the “ideal” value. Similarly, the average RMSE and FPR values fall below the corresponding “ideal” values for each investigated threshold and initial training set size. The average TPR values are greater than 99%, and for a threshold of 0.3, the “ideal” value is exceeded. Likewise, the active learning approach yields the TPR larger than 99.53% for an initial training set size of 0.5% and a threshold value of 0.2.

The learning behaviour of the classifier is shown by the way of example for an initial training set size of 5% and a threshold of 0.1. The learning curves are shown in the appendix *A.26* see Figure 212. Overall, the AL process demands the labelling of 425 images in 17 training cycles. Thus, 477 labelled images are needed to train the classifier, which achieves the average accuracy of 85.00%. The average gain in accuracy is 8.73% and the average values exceed the “ideal” value after 6 training cycles. The average FPR values and the average RMSE values fall below the corresponding “ideal” values. Similarly, the average TPR values increase after each training cycle.
A possible labelling of the test images for a single run is further investigated, as shown in Figure 136. Upon completion of the AL process, the classifier achieves an accuracy of 86.39%, which is 2.50% higher than the “ideal” value. The obtained TPR value of 99.53% corresponds exactly to the value obtained with all manually labelled training images. The archived FPR is 6.21% lower than 39.31%. Thus, fewer test images, which are manually labelled with 1 or 2 rating points, receive incorrectly a positive label.

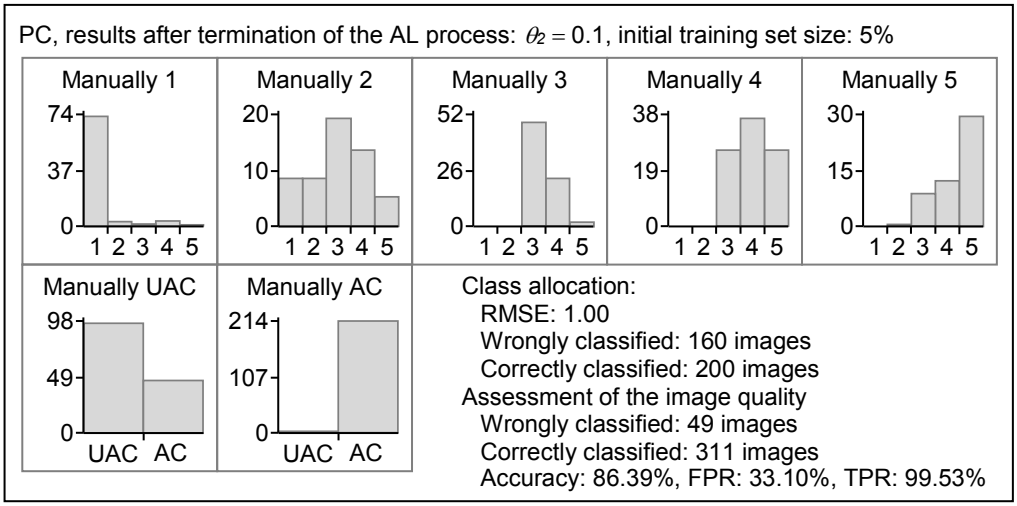


Figure 136: AL, PC for the distortion dataset: possible labelling of the test images, $\theta_2 = 0.1$

Now the 2 selection criteria are combined. A training image needs to be labelled manually if the maximum probability that the image belongs to the corresponding rating class is less than the threshold θ_1 and if the difference between the largest and the second largest class probability is less than a threshold θ_2. The classification results obtained for various initial training set sizes after the termination of the AL process are summarised in the appendix *A.25* see Figure 209 and Figure 210. Large threshold values lead to an increasing label effort by the Oracle. On average over 30 runs, the active learning approach yields recognition accuracies higher than 83.89%, TPR values higher than 99.53% and FPR values lower than 39.31%. If the threshold combination $\theta_1 = 0.6$ and $\theta_2 = 0.1$ is applied, the average RMSE values are below 1.12.

The learning behaviour of the classifier, for a threshold combination of $\theta_1 = 0.6$ and $\theta_2 = 0.1$ and an initial training set size of 5%, is shown exemplarily in the appendix *A.26* see Figure 213. The more training images are labelled the better the classification result. The average accuracies and the average TPR values increase with the number of labelled training images and exceed the corresponding “ideal” values after 14 and 16 training cycles, respectively. Similarly, the average RMSE values and the average FPR values decrease after each training cycle and fall below the corresponding “ideal” values after 16 and 6 training cycles, respectively. After completing the AL process, 677 training images are manually labelled.

A possible labelling of the test images is shown in Figure 137 for a single run. With 677 labelled training images, the classifier obtains the accuracy of 86.11%, the TPR of 99.53%, the FPR of 33.79%, and the RMSE of 1.01. The TPR is equal to the value obtained by all manually labelled training images, the accuracy is 2.22% higher, and the FPR is 5.52% lower than the corresponding “ideal” values. Depending on the used threshold values, the experimental evaluation shows that the active learning approaches can achieve higher recognition accuracies than the classifier trained with all manually labelled training images. The AL process selects those images that beneficially affect the recognition behaviour. Good classification results are obtained due to the absence

of many uninformative training images. Thus, the AL process tries to reduce the labelling costs.

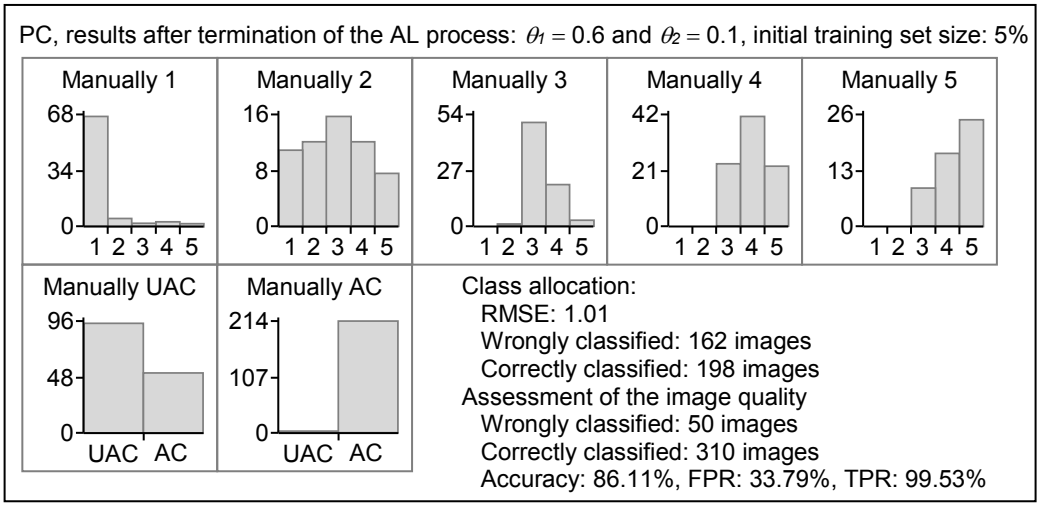


Figure 137: AL, PC for the distortion dataset: possible labelling of the test images, θ_1 and θ_2

Finally, the active learning approach is applied to a 2-class PC of 2nd order in the 7-dimensional component space. An informative training image is detected if the difference between the largest and the second largest class probability is less than a threshold θ_2, which is set to 0.1. If all manually labelled training images are used for training, the area under the ROC curve is 0.87. The experimental evaluation shows that the proposed active learning approach yields AUC values between 0.72 and 0.75, which are lower than the value of the "ideal" classifier, as shown in Figure 138. Thus, the applied AL process is only less suitable for identifying the most informative training samples and extracting enough information to generalise to unknown data. The manual label effort can only be reduced if a decline in the recognition accuracy is acceptable.

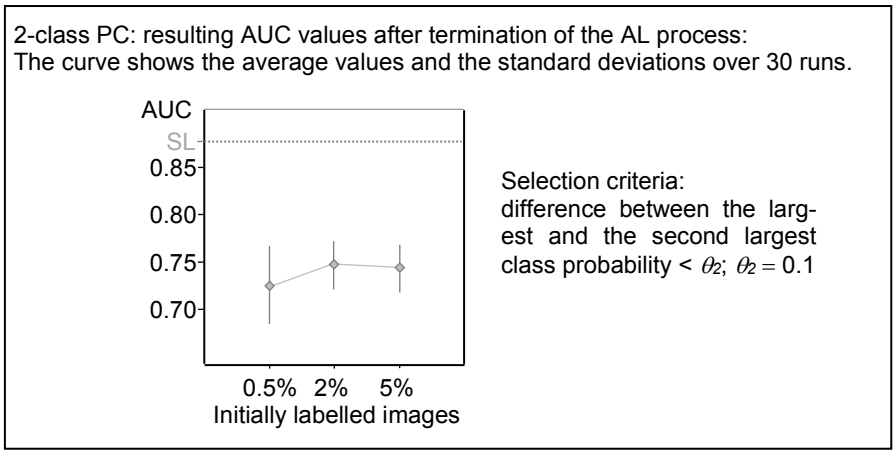


Figure 138: AL, 2-class PC: results for the distortion dataset $\theta_2 = 0.1$

K-nearest neighbour classification:

The AL scenario is applied to the NN classifier for $k = 1$ in the 3-dimensional component space. Using all manually labelled training samples the RMSE of 1.37, the accuracy of 57.22%, the FPR of 6.21% and the TPR of 32.56% are obtained. The AL algorithm is able to query interactively for the correct label if the distance to the next training sample is greater than a threshold distance. The threshold distances are defined already in chapter *7.5.1*; see Table 33.

The results of the test images obtained after the termination of the AL process are summarised in the appendix *A.27* see Figure 214 and Figure 215. The smaller the threshold distance, the more images need to be labelled manually and the average results will be very similar to the corresponding "ideal" values. If the maximal threshold distances are applied as criteria for ignorance or querying the Oracle, the average accuracy values exceed 57.22%. Likewise, the average TPR values exceed 32.56%. If the mean threshold distances are applied, average FPR values are obtained that fall below 6.21%. The nearest neighbour classifier appears to be unable to cope with all manually labelled test images. The amount of uninformative images negatively affects the classification accuracy. Reducing the manual label effort to a few informative images, better classification results can be achieved.

The learning curves of the classifier are shown by the way of example for an initial training set size of 2% and a maximal threshold distance, see Figure 216 in the appendix *A.28.* Based on 22 labelled training images, the Oracle is interactively queried 10 times to label another 250 training images. The increase in performance during active learning is low. The average gain in accuracy is 2.05%. The obtained recognition accuracies are above 57.22% after each training cycle. During the active learning process, the average RMSE values and the average FPR values decrease with increasing the number of manually labelled training images. For a single run, the labelling of the test images is exemplarily shown in Figure 139.

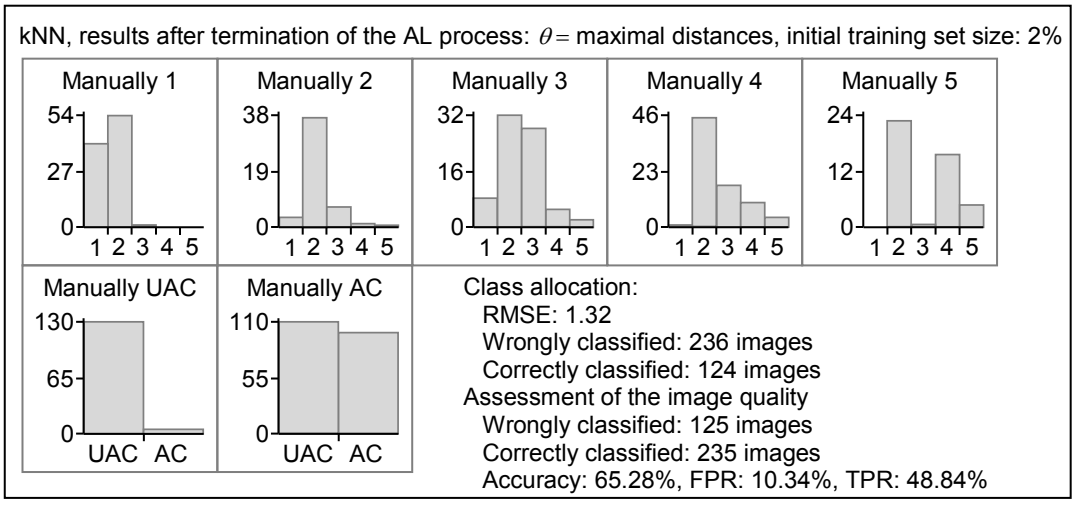


Figure 139: AL, kNN for the distortion dataset: possible labelling of the test images, θ = maximal threshold distance

The proposed active learning process yields the accuracy of 65.28% and the TPR of 48.84%. The values are 8.06% and 16.28% higher than the corresponding “ideal” values. However, many test images that are manually labelled with 3, 4 or 5 rating points are incorrectly assigned to rating class 2. Therefore, the TPR obtained is still less than 50%. Unfortunately, the FPR amounts 10.34%, which indicates that 4.13% more test images wrongly receive a positive label. Applying the active learning procedure yields good classification accuracies that are higher than the values of a classifier trained with all manually assigned class labels. Nevertheless, the learning success during the AL procedure is low.

Finally, the active learning behaviour of a 2-class NN classifier in the 7-dimensional component space is investigated. An area under the ROC curve of 0.87 can be determined if all manually labelled training images are used for training and the number of nearest prototype vectors k is set to 10. An informative training image is detected if the distance to the nearest reference pattern is greater than the mean threshold distance θ. To determine the ROC curve the number of images *m* required to classify a test image as AC is varied between 1 and k. Figure 140 shows the average AUC values and the standard deviations over 30 runs after the AL process has ended. The x-axis shows the number of initially labelled training images and the y-axis the resulting AUC values. The obtained AUC values are between 0.77 and 0.82. Noticeable is the large standard deviation of the AUC values for an initially labelled training set size of 2%. Here, the ability of the classifier to predict the correct labels depends strongly on the choice of the initially labelled training images.

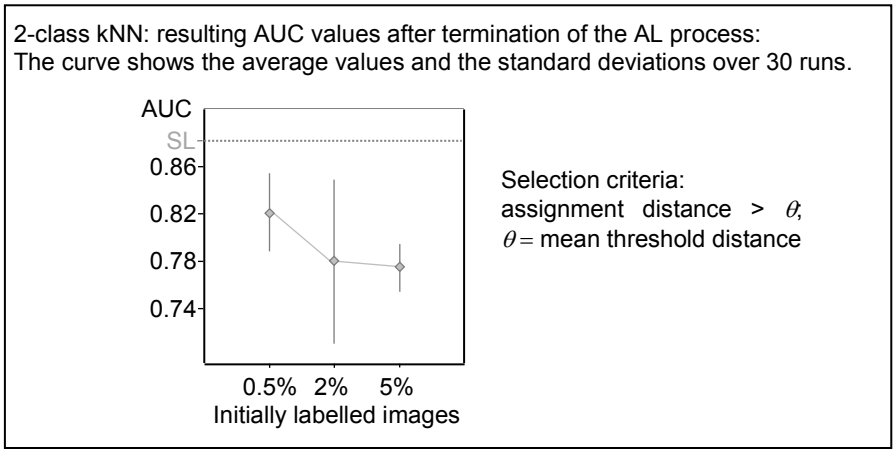


Figure 140: AL, 2-class kNN: results for the distortion dataset
θ = mean threshold distance

The experimental evaluation shows that the proposed active learning approach yields AUC values that are below the “ideal” value. Thus, the active learning procedure does not seem to be able to select the most informative training samples. If the manual labelling effort is reduced, the 2-class classifier is less suitable to extract enough information to generalise to unknown data.

Learning vector quantisation:

Now, the classifier behind the AL approach is the LVQ. Here, 165 prototypes are applied to approximate the relationship between the training images and the manually obtained class labels. The prototypes are determined by learning rule LVQ1. If all manually labelled training images are used for training, the RMSE of 1.10, the accuracy of 88.06%, the FPR of 29.86% and the TPR of 100.00% are obtained. For the following investigations, the same prototype initialisation is applied as before. The threshold distances, which decide if the training image has to be labelled by the Oracle, are introduced in chapter *7.5.1* see Table 34.

The results obtained on the test images after terminating the AL process are summarised in the appendix *A.29* see Figure 217 and Figure 218. The smaller the threshold distances are chosen the more training images need to be labelled manually. If the minimal or the mean threshold distances are applied, good accuracy values are yield that are very similar to the “ideal” value. In contrast, if the maximal threshold distances are used, only a few training images are labelled manually, resulting in accuracy values below 88.06%. The lowest RMSE values and the lowest FPR values are obtained if the mean threshold distances are applied. The TPR of 100.00% is not reached on average.

The learning behaviour of the classifier is examined as an example for an initial training set size of 0.5%. The resulting learning curves are shown in the appendix *A.30* see Figure 219. If the mean threshold distances are applied, 583 training images are manually labelled during 23 training cycles. The manual labelling effort is roughly halved. The average FPR values and the average RMSE values decrease after each training cycle. The values fall below the corresponding “ideal” values after 11 and 5 training cycles. The average accuracy exceeds the “ideal” value. The average TPR values increase with the number of manually labelled training images, but 100.00% is not reached. A possible labelling of the test image for a single run is shown in Figure 141.

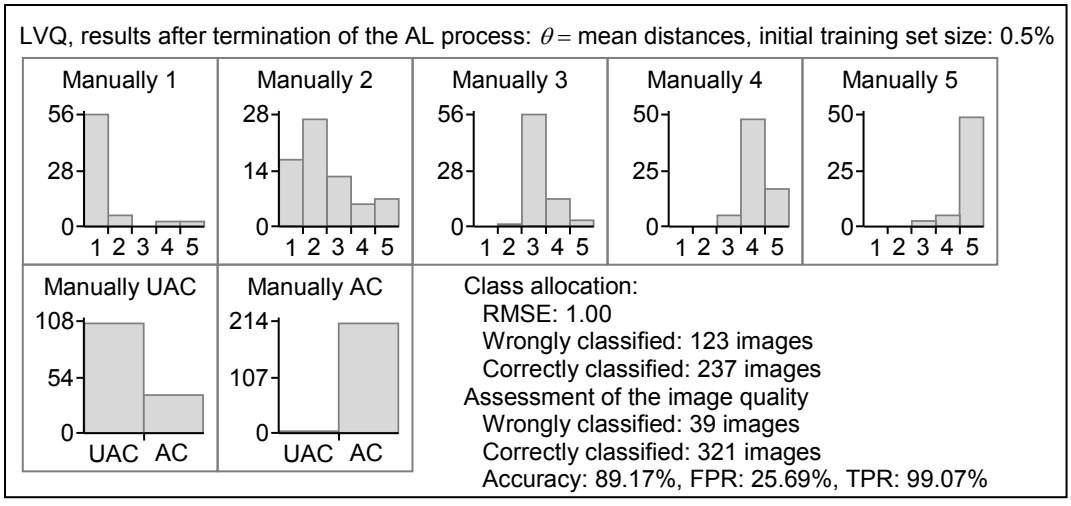


Figure 141: AL, LVQ for the distortion dataset: possible labelling of the test images, θ = mean threshold distance

This active learning approach yields the accuracy that is 1.11% higher, and the FPR that is 4.17% lower than that of a classifier using all manually assigned class labels. Here, less manually labelled test images with 1 or 2 rating points are mistakenly assigned to rating classes 3, 4 or 5. Test images that are manually labelled with 4 or 5 rating points are assigned to rating classes 3, 4 and 5 that represent an acceptable image quality. The AL approach yields the TPR of 99.07%, indicating that an insignificant number of test images, manually labelled with 3 rating points, are incorrectly assigned to rating class 2.

Finally, the active learning procedure is applied to a 2-class LVQ. Here, 165 prototype vectors are determined and the number of nearest prototype vectors k is set to 30. If all manually labelled training images are used for training, an area under the ROC curve of 0.86 can be determined. During the AL process, an image is manually labelled and transferred into the training dataset if the distance to the nearest prototype vector is greater than the mean threshold distance θ. After each training cycle, the AUC value is determined by varying the number of images *m* required to classify a test image as AC. Upon completion of the AL processes, the average AUC values and the standard deviations over 30 runs are shown in Figure 142.

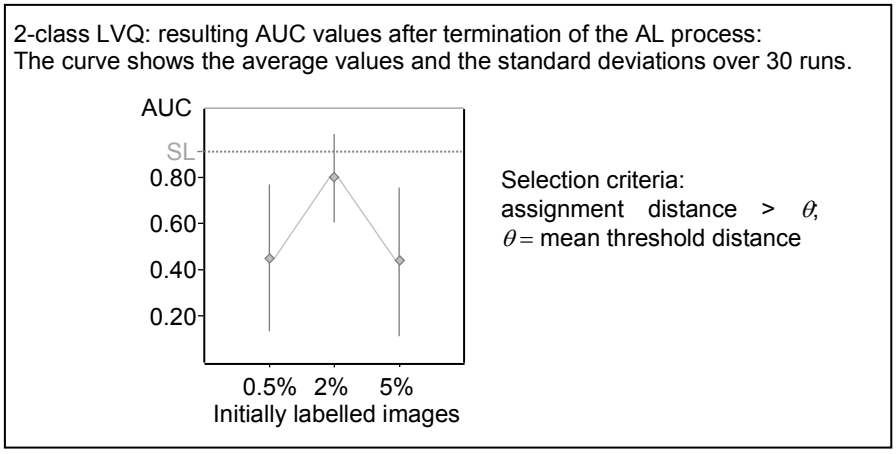


Figure 142: AL, 2-class LVQ: results for the distortion dataset
θ = mean threshold distance

The diagram shows that averaged AUC values of 0.81 can be achieved if the AL process starts with 2% of all manually labelled training images. Here, depending on the initially labelled training images, an AUC value exceeding 0.86 can be achieved. In contrast, for other starting conditions the AL process yields AUC values lower than that of a classifier using all manually assigned class labels. Depending on the selection of the initially labelled training images, AUC values less than 0.20 are achieved. Since the most test images get a wrong class label, these classifiers are not able to make decisions. Thus, the applied AL process is only less suitable for identifying the most informative training samples and extracting enough information to generalise to unknown data.

Summary of the obtained results:

The experimental evaluation shows that the classifiers appear to be overcharged with the labels of the uninformative training images. Many uninformative images negatively affect the classification accuracy. The proposed AL approaches yield higher recognition accuracies than classifiers trained with all training images labelled in advance. Depending on the criteria for querying the Oracle, the PC yields the TPR of 100% and the recognition accuracy greater than 85%. The NN classifier yields the accuracy and the TPR that are 8.06% and 16.28% higher than the corresponding "ideal" values. Similarly, the FPR on the test images is reduced by 4.17% if the LVQ algorithm is allowed to select the most informative training images. If the classification problem is reduced to a 2-class assignment task, the 3 classifiers do not seem to be able to identify the most informative training images. On average, the "ideal" values are not achieved and the labelling effort can only be reduced if a decline in the recognition accuracy is acceptable.

7.6.2 AL: assessment algorithms for the double image dataset

Again, the size of the initially labelled training set is set to 0.5%, 2% and 5% of all training images from each rating class, as shown in Table 42. It is ensured that at least 1 training image is selected for each rating class. A new training cycle is intended when 10 new informative training images are interactively labelled by the Oracle and inserted into the training set. This is repeated until all informative training samples are identified by the Oracle. After each training cycle, the classifiers are used to label the test images.

Initial number of labelled training images	Classes					Overall
	1	2	3	4	5	
0.5%	1	1	1	1	1	5
2%	2	3	1	1	1	8
5%	6	8	2	1	1	18

Table 42: AL for the double image dataset: initial number of labelled training samples

Polynomial classification:

The active learning behaviour of a 3rd order classifier in the 2-principal component space is investigated. Using the entire labelled training set, the classifier yields the RMSE of 0.85, the accuracy of 90.00%, the FPR of 7.48%, and the TPR of 88.26%.

Several criteria for ignorance or querying the Oracle for the right class label are applied. First, a training image is identified as informative if the maximum probability that the image belongs to the corresponding rating class is less than the threshold θ_1, which is set to 0.6, 0.7 and 0.8. In the second step, an image needs to be labelled by the Oracle if the difference between the largest and the second largest class probability is less than the threshold θ_2, which is set to 0.1, 0.2 and 0.3. Finally, the 2 threshold criteria are combined. The size of the initially labelled training set is set to 0.5%, 2%, and 5%.

The results obtained after the termination of the AL process are shown in the appendix *A.31* see Figure 220 to Figure 225.

The experimental evaluation shows that the proposed active learning approach yields average accuracy values less than 90.00%. The accuracy values average over 30 runs become very similar to the "ideal" value the more labelled training images are available. In 4 out of 30 runs, the accuracy exceeds 90.00% after completion of active learning ($\theta_1 = 0.7$ and $\theta_2 = 0.2$, initial training set size of 0.5%). Likewise, the average RMSE values are above 0.85 and are similar to this value, as more labelled training images are requested. The same applies to the average FPR values, which are above 7.48%. The average TPR values do not reach 88.26%, except for a threshold value $\theta_2 = 0.2$ and an initial training set size of 0.5%.
Some learning curves of the classifier are shown in the appendix *A.32*. The learning curves shown in Figure 226, Figure 227 and Figure 228, make clear that there is hardly any learning success for the average TPR values. After each training cycle, the obtained TPR values are nearly the same. The development of the average RMSE values and the average FPR values are almost equal. The more manually labelled training images are available, the lower the RMSE and FPR values. The average FPR values decrease by 61.77% during the AL process if the threshold values $\theta_1 = 0.7$ and $\theta_2 = 0.2$ are applied and the initial training set size is set to 0.5%. The average accuracies increase with the number of labelled training images, but the "ideal" value is not reached on average. Therefore, if the labelling effort is reduced, a decrease in the classification accuracy must be accepted. The most informative training images do not contain sufficient information to obtain the desired outputs for the test images. Depending on the threshold parameters, results are obtained that are very similar to those of a classifier using all manually assigned class labels. A possible labelling of the test images for a single run is shown in Figure 143.

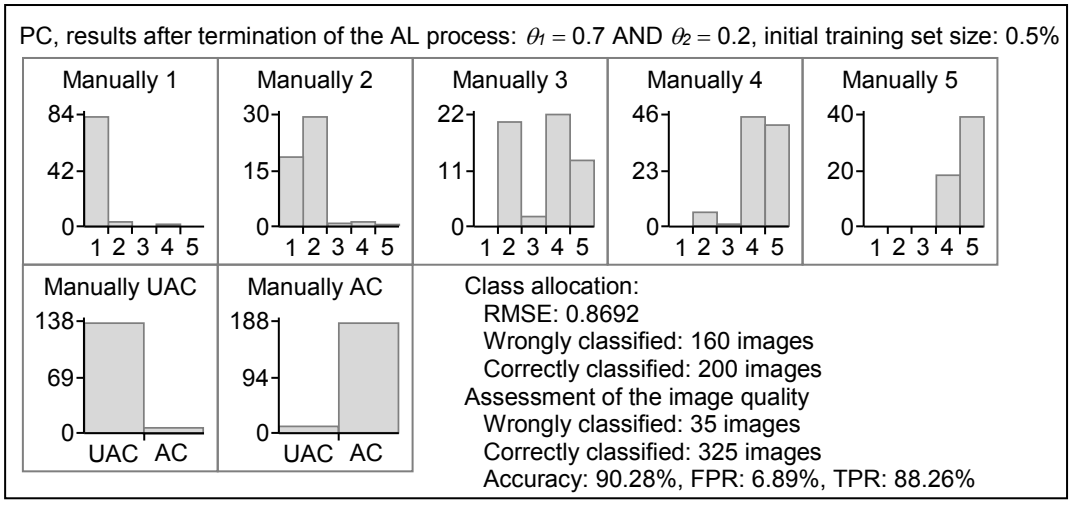


Figure 143: AL, PC for the double image dataset: possible labelling of the test images, θ_1 and θ_2

In 23 training cycles, the number of manually labelled images is increased from 5 to 235. Upon completion of the AL procedure, an accuracy of 90.28% is obtained. This value is 0.28% higher than the “ideal” value. Many test images that are manually labelled with 3 rating points are incorrectly assigned to rating class 2. Thus, a TPR of 88.26% is obtained, indicating that 11.74% of the test images representing an acceptable quality receive wrongly a negative label. The TPR value corresponds exactly to the “ideal” value. The PC yields the FPR that is 0.59% lower than the “ideal” value. Thus, fewer vehicles of unacceptable quality would receive a positive label.

Finally, the active learning rule is applied to a 2-class PC of 2nd order in the 2-principal component space. If all manually labelled training images are available, the area under the ROC curve is 0.998. To select the most informative training images, the threshold value θ_1 is used, which is set to 0.8. The resulting ROC curves are hardly different from the “ideal” curve, and the curves give averaged AUC values greater than 0.990, as shown in Figure 144. The standard deviations over 30 runs are negligible. This indicates that the resulting ROC curves are nearly independent on the selection of initially labelled training images.

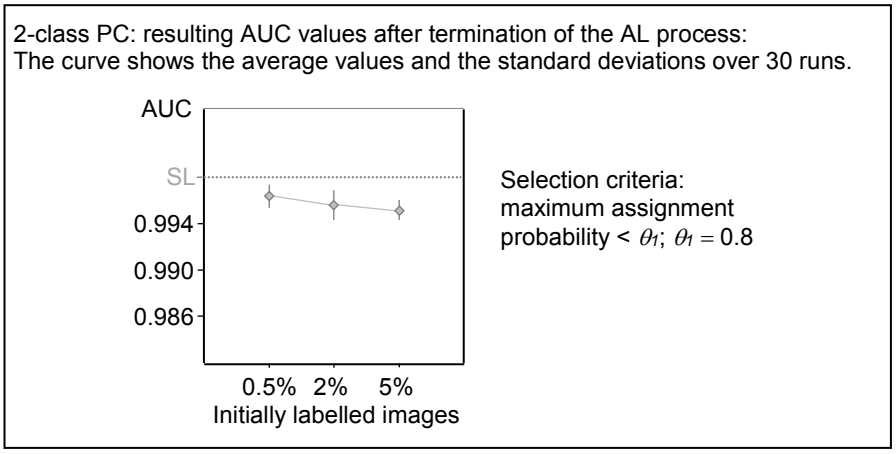


Figure 144: AL, 2-class PC: results for the double image dataset
$\theta_1 = 0.8$

The experimental evaluation shows that the proposed active learning approach yields AUC values similar to those of a classifier using all manually labelled images. Here, the manual label effort can be reduced without a significant loss of recognition accuracy. This is possible because the classifier selects the most informative training images.

K-nearest neighbour classification:

The active learning behaviour of the nearest neighbour classifier is investigated for $k = 3$ in the 7-principal component space. If all manually labelled training images are used for training, the classifier yields the RMSE of 0.65, the accuracy of 95.28%, the FPR of 0.68% and the TPR of 92.49%. A training image is labelled by the Oracle if the distance to the next training sample is greater than the threshold. The threshold distances are already determined in chapter *7.5.2*; see Table 36.

The results after completion of the active learning process are summarised in the appendix *A.33* see Figure 229 and Figure 230. The experimental evaluation shows that the larger the threshold distances are chosen, the less frequent the Oracle is queried for the right class label. The more labelled training images are available, the less the difference between the results obtained and the "ideal" result. For all cases, the average FPR is below 5.10%. If the minimal threshold distances are used, the average FPR is exactly the "ideal" value. Likewise, the average values of the TPR, the accuracy and the RMSE are very similar to the corresponding "ideal" values. However, on average the "ideal" values are not reached. In contrast, if the maximal threshold distances are used, the labels of test images are very different from the subjective perception.

The learning behaviour of the classifier is shown by the way of example in the appendix *A.34* see Figure 231. The curves are determined for an initial training set size of 5% and mean threshold distances. 158 training images are labelled in 14 training cycles. Average over 30 runs, the accuracy values increase with the number of labelled training samples and the achieved gain is 32.23%. For 1 out of 30 runs, the accuracy achieved by active learning is higher than 95.28%, and for 1 run, the accuracy is exactly 95.28%. Likewise, the average TPR values increase after each training cycle and the average gain is 34.24%. In contrast, the average RMSE values and the average FPR values decrease with the number of labelled training samples. A possible labelling of the test images for a single run is shown in Figure 145. Here only 158 labelled images are used for training.

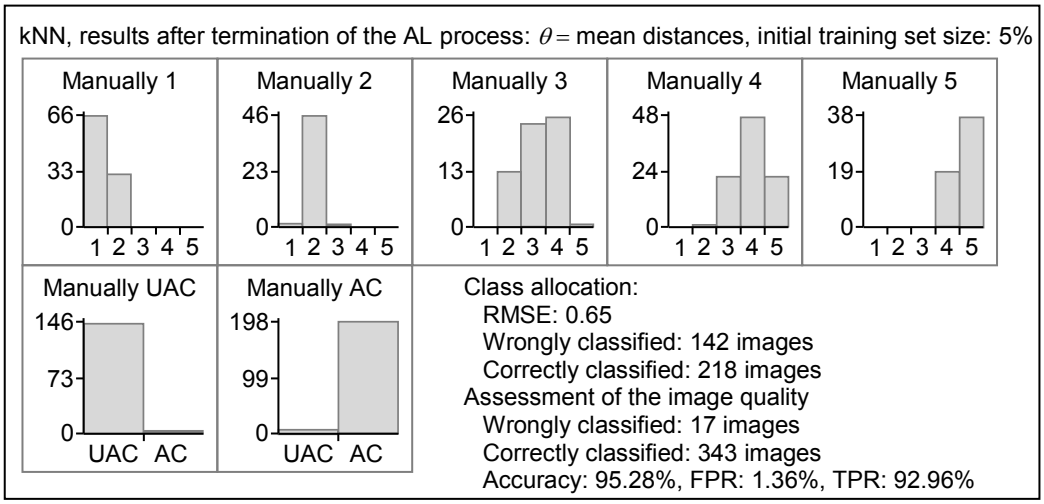


Figure 145: AL, kNN for the double image dataset: possible labelling of the test images, θ = mean threshold distances

Upon completion of the active learning scenario, a TPR exceeding the "ideal" value is obtained. The RMSE and the accuracy correspond to the associated "ideal" values. Unfortunately, the FPR is higher than the "ideal" value. Thus, 0.68% more vehicles would mistakenly receive a positive label. Test images that are manually labelled with 1 or 2 rating points are mainly assigned to rating classes 1 and 2. Likewise, test images that are manually labelled with 4 or 5 rating points are assigned to rating classes 3, 4

and 5. Difficult is the assignment of test images, which are manually labelled with 3 rating points. These images are assigned to rating classes 2, 3, 4 and 5.

The experimental evaluation shows that the AL algorithm is able to select the most informative training images. If the most informative images are used as reference patterns, similar results are obtained on the test images as with a classifier using all manually labelled training images. Since the “ideal” values are not reached on average, the unconsidered training images contain information that cannot be mapped by an active learning algorithm. By reducing the manual labelling effort, low losses in the recognition accuracy must be accepted.

Finally, the active learning behaviour of the 2-class NN in the 7-dimensional component space is investigated. If all manually labelled training images are available, an area under the ROC curve of 0.964 is achieved. A training image needs to be labelled manually if the distance to the nearest reference image is greater than the maximal assignment distance θ. The average AUC values and the standard deviations over 30 runs are shown in Figure 146. The experimental evaluation shows that the proposed active learning approach gives AUC values between 0.830 and 0.870. These values are lower than the values of a classifier using all manually assigned class labels. The AUC values obtained mainly depend on the selection of the initially labelled training images. This is indicated by the large standard deviation of the resulting AUC values. The most informative training images are not sufficient to train an algorithm that performs better than the “ideal” classifier. The manual labelling effort can only be reduced if a decrease in the AUC values can be accepted.

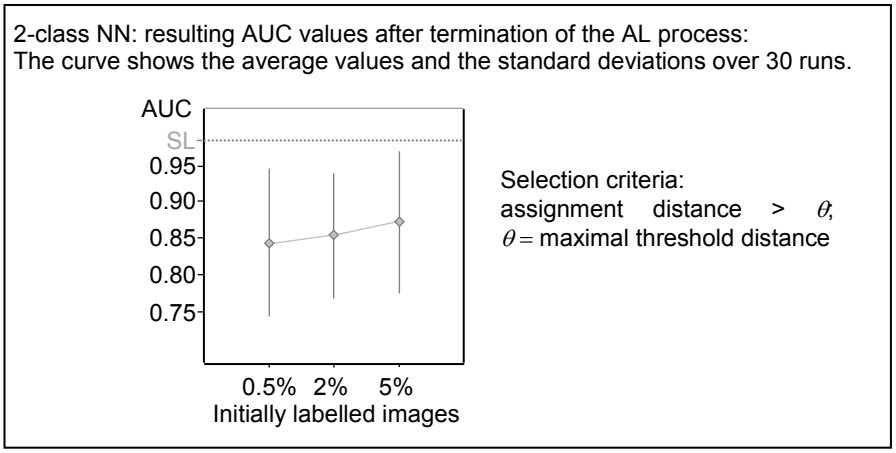


Figure 146: AL, 2-class kNN: results for the double image dataset
θ = maximal threshold distance

Learning vector quantisation:

The classifier for the active learning approach is the LVQ. Here, the labelled training images are mapped to 69 prototypes, which are determined by the LVQ2.1 learning rule. The following investigations are based on the same prototype initialisation. If all training images are labelled in advance, the RMSE of 1.16, the accuracy of 83.06%,

the FPR of 33.33%, and the TPR of 94.37% are obtained. The threshold distances for detecting the most informative images are defined in chapter *7.5.2*; see Table 37.
The results obtained on the test images after completion of the AL process are summarised in the appendix *A.35* see Figure 232 and Figure 233. The shorter the threshold distances, the more training images need to be labelled manually by the Oracle. If the training images which are beyond the maximal threshold distances are labelled by the Oracle, the worst result on the test images is obtained. Here, the labelled training set consists of less than 80 images. The relatively small training set does not contain sufficient information to generalise to the unknown test data. Results similar to the "ideal" values are obtained if the mean threshold distances are used to detect the informative training samples. The accuracy average over 30 runs is similar to the "ideal" accuracy. The average TPR values exceed 94.37% for initial training set sizes of 2% and 5%. If the minimal threshold distances are applied, the average accuracy values and TPR values are obtained that exceed the corresponding "ideal" values. Similarly, the proposed active learning approach yields RMSE and FPR values, which are very similar to the "ideal" values. A possible labelling of the test images is shown in Figure 147. The most informative training images are selected by the minimum threshold distance.

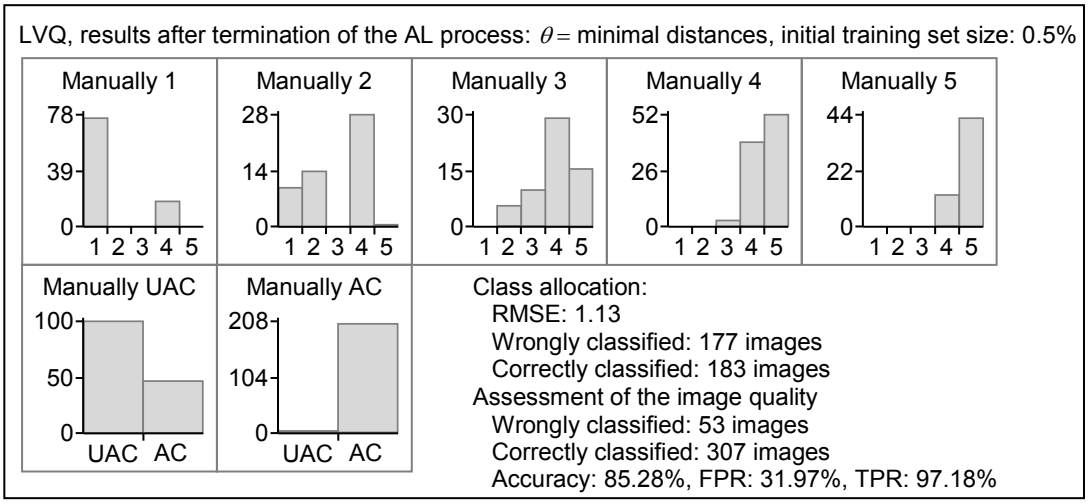


Figure 147: AL, LVQ for the double image dataset: possible labelling of the test images, $\theta =$ minimal threshold distances

Due to the absence of uninformative training images, the classifier yields the RMSE of 1.13, which is slightly below the "ideal" value. The obtained FPR of 31.97% indicates that 1.36% fewer test images receive wrongly a positive label. Likewise, the obtained TPR is 2.81% higher than 94.37% and indicates that 97.18% of all test images showing an acceptable quality receive a positive label. Overall, the classifier yields an accuracy of 85.28%, which exceeds the "ideal" value.

The learning behaviour of the classifier is investigated as an example for an initial training set size of 0.5%. If the minimal threshold distances are applied, the active learning scenario demands the labelling of 245 training images. The corresponding learning curves are displayed in the appendix *A.36* see Figure 234. After each training cycle,

the average accuracy values increase, while the average RMSE values and the FPR values decrease. The gain in accuracy is low and is only 2.09%. After 19 training cycles, the average recognition accuracies exceed the “ideal” value. For 24 training cycles, the average TPR values are above the “ideal” value. The experimental evaluation shows that the active learning approach gives recognition accuracies higher than that of a classifier using all manually assigned class labels. Thus, the manual label effort can be reduced and the classifier is not overcharged with too many uninformative training images.

At the end of this chapter, the active learning behaviour of a 2-class LVQ is investigated. The classifier uses 69 prototype vectors. The minimal threshold distances θ are applied to check if an informative training image should be labelled by the Oracle. After each training cycle, the area under the ROC curve is determined by varying the number of images *m* required to classify a test image as AC. The average AUC values and the standard deviations over 30 runs are shown in Figure 148. The averaged AUC values between 0.64 and 0.83 are below the “ideal” value of 0.98. Depending on the selection of the initially labelled training images, AUC values less than 0.50 are obtained. In such cases, more test images receive a wrong label than a correct one. On the contrary, if the most informative training images are found, the AUC values are close to the “ideal” value.

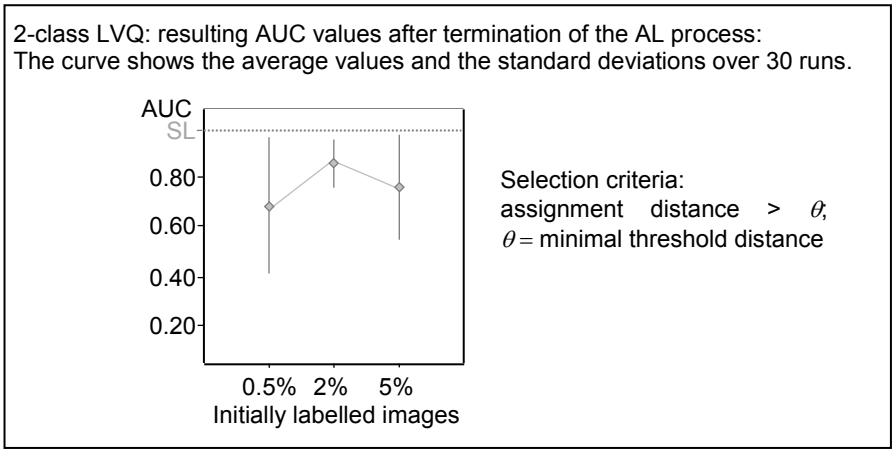


Figure 148: AL, 2-class LVQ: results for the double image dataset
θ = minimal threshold distance

Summary of the obtained results:

The experimental evaluation shows that the different classifiers have a different learning behaviour. In the course of the active learning process, the average recognition accuracies of the PC increase, but the “ideal” value is not reached. If the most informative images are used as reference patterns, the nearest neighbour classifier yields similar results on the test images as a classifier using all manually labelled training images. Thus, these classifiers are only partially able to extract sufficient information to generalise to unknown data. Depending on the threshold parameters, the LVQ algorithm yields a recognition accuracy that is higher than the “ideal” value. By the selecting the most

informative training images, the classifier is not overcharged with too many uninformative training images. If classification problem is limited to an acceptable or unacceptable quality, the 2-class PC yields AUC values similar to those of a classifier using all manually assigned class labels. Here, the manual label effort can be reduced without a significant loss of recognition accuracy. In contrast, the 2-class kNN and the 2-class LVQ are not able to extract sufficient information to generalise to the unknown test data. Here, the averaged AUC values are below the corresponding "ideal" values. In addition, the AUC values mainly depend on the selection of the initially labelled training images.

7.6.3 AL: assessment algorithms for the distortion and double image dataset

The classifier asks the Oracle to label the most informative training images and the classifier is retrained with the enlarged training dataset. After each training cycle, the classifier is applied to label the test images. This is repeated until all informative training images are found. A new training cycle is intended if 9 informative manually labelled training images are transferred into the training dataset. The size of the initially labelled training set is set to 0.5%, 2% and 5% of all training images from each rating class, as shown in Table 43. It is ensured, that at least 1 training image from each rating class is selected.

Initial number of labelled training images	Classes					Overall
	1	2	3	4	5	
0.5%	1	1	1	1	1	5
2%	1	3	3	1	1	9
5%	1	7	7	2	1	18

Table 43: AL for the distortion and double image dataset: initial number of labelled training samples

Polynomial classification:

The active learning scenario is investigated for a polynomial classifier of 1st order in the 20-dimensional component space. If the labels of all training images are known in advance, the RMSE of 0.67, the accuracy of 71.48%, the FPR of 31.88%, and the TPR of 74.25% are obtained.

Several criteria are used to select the most informative training samples. The results obtained after completion of the active learning process are summarised in the appendix *A.37*. First, the Oracle is asked for the correct label if the maximum probability that the image belongs to the corresponding rating class is lower than the threshold θ_1, which is set to 0.6, 0.7 and 0.8. The results are shown in Figure 235 and Figure 236. Depending on the selection of the initially labelled training images, the active learning scenario yields recognition accuracies that exceed the "ideal" accuracy for threshold values of 0.7 and initial training set sizes of 2% and 5%. The average RMSE values

and the average FPR values do not fall below the corresponding “ideal” values. If θ_1 is set to 0.7, the average TPR values are slight above 74.25%.
During the second step, an informative image needs to be labelled by the Oracle if the difference between the largest and the second largest class probability is less than a threshold θ_2, which is set to 0.1, 0.2 and 0.3. Here, the obtained classification results depend on the used threshold and the number of initially labelled training images as shown in Figure 237 and Figure 238. With an initial training set size of 5% and a threshold of 0.2, the AL procedure results in better recognition accuracies and higher TPR values than a classifier using all manually assigned class labels. The average FPR values fall below the “ideal” value for initial training set sizes of 2% or 5% and a threshold value of 0.1. The same applies to a threshold value of 0.2 and an initial training set size of 0.5%.
If both selection criteria are combined, no improvements in the classification results can be observed, as shown in Figure 239 and Figure 240. Average over 30 runs the “ideal” accuracy and the “ideal” TPR are not reached. The average RMSE values are above 0.67. Depending on the used threshold parameters and the size of the initially labelled training set, FPR values are achieved, which fall below 31.88%.

The learning behaviour of the classifier is investigated exemplarily for an initial training set size of 2% and different threshold criteria as shown in the appendix *A.38* see Figure 226, Figure 227 and Figure 228. The learning curves are very similar for the different thresholds. The average RMSE values and the average FPR values decrease after each training cycle. Likewise, the average accuracy values increase with the number of labelled training images. If the threshold value θ_1 is set to 0.7, the average gain in accuracy is 17.71%. If the thresholds θ_1 and θ_2 are set to 0.6 and 0.1, the FPR values fall by 41.38% during the active learning procedure. The resulting average FPR value is below the “ideal” value. Similarly, the average RMSE values fall by 0.80 when the threshold value θ_2 is set to 0.2. In contrast, no direct learning success can be noted for the average TPR values. During the active learning process, the TPR values decrease at the beginning and then increase again. Depending on the threshold criteria and the initially labelled training images, the “ideal” value is even exceeded.
A possible labelling of the test images is shown in Figure 149. Shown is 1 run for which the “ideal” accuracy is exceeded. The obtained labels for the test images are very similar to the labels of the classifier trained with all manually labelled training images. The AL procedure gives the RMSE of 0.68, the accuracy of 72.13%, the FPR of 28.99% and the TPR of 73.05%. In 20 training cycles, the number of labelled training images is increased from 9 to 189. Since the FPR value is below the “ideal” value, 2.89% fewer test images are falsely given a positive label. As mentioned above, the active learning process does not result in any improvement in the true positive rate. Thus, the TPR value is below 74.25%. Consequently, this classifier would cause further costs since fewer test images of acceptable quality will receive correctly a positive label.

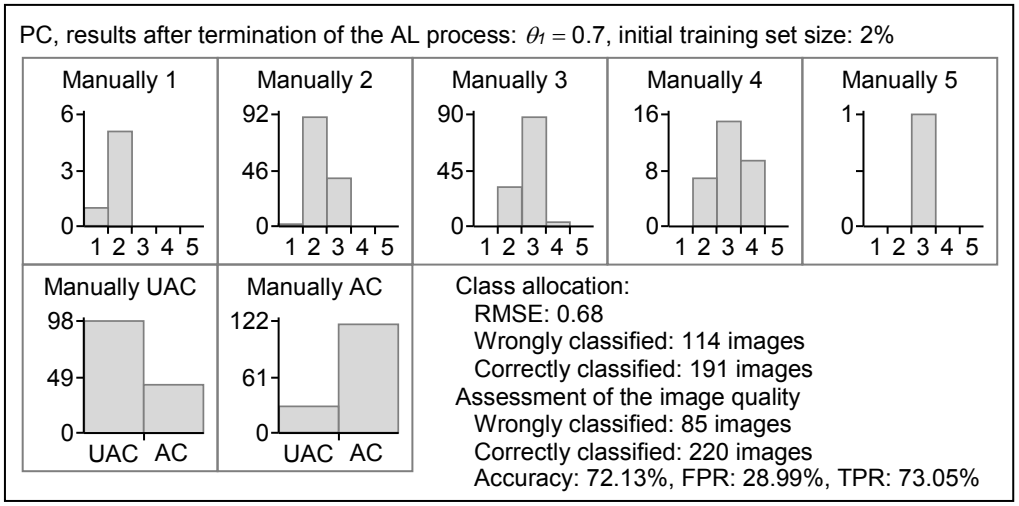


Figure 149: AL, PC for the distortion and double image dataset: possible labelling of the test images, $\theta_1 = 0.7$

The experimental evaluation shows that the more labelled training images are available, the more similar the results obtained to the "ideal" results. Depending on the thresholds, the AL procedures may even lead to accuracies that slightly exceed the "ideal" accuracy. This indicates that the proposed algorithm selects those training samples, which improve the recognition behaviour in an advantageous manner.

Finally, the active learning behaviour of a 2-class 2nd order classifier in the 2-principal component space is investigated. If all training images are labelled in advance, the area under the ROC curve is 0.69. The results obtained for the test images are summarised in Figure 150. Here, the threshold θ_2 is set to 0.2. Thus, an informative training sample is detected if the difference between the largest and the second largest class probability is less 20%.

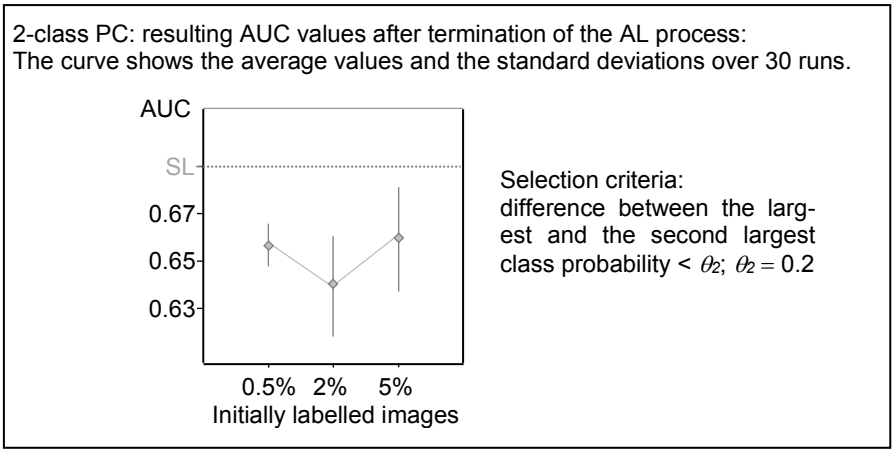


Figure 150: AL, 2-class PC: results for the distortion and double image dataset $\theta_2 = 0.2$

The experimental evaluation shows that the proposed active learning approach gives an AUC value that is on average slightly lower than that of a classifier using all manual-

ly assigned class labels. If the manual labelling effort is reduced to some informative training images, the classifier is still able to generalise well to the unknown test images.

K-nearest neighbour classification:

The active learning behaviour of the nearest neighbour classifier is investigated for $k = 1$ in the 13-principal component space. If all labelled training images are available in advance, the classifier achieves the RMSE of 0.43, the accuracy of 91.48%, the FPR of 11.59% and the TPR of 94.01%. A training image needs to be labelled by the Oracle if the distance to the next training sample is greater than the threshold distance. The threshold distances are already determined in *chapter 7.5.3*; see Table 39.

The results after completion of the active learning process are summarised in the appendix *A.39*; see Figure 244 and Figure 245. If the minimal threshold distances are applied as selection criteria for the active learning process, almost 300 images must be labelled manually. Therefore, classification results are achieved that are very similar to the result of a classifier that uses all manually assigned class labels. If the threshold distances are extended, fewer training images need to be labelled by the Oracle. If the mean threshold distances are applied, the kNN still provides good accuracy values. In contrast, if the maximal threshold distances are used, too few labelled training images are available to achieve good classification results.

By the way of example, the learning behaviour of the classifier is investigated for an initial training set size of 5% and mean threshold distances, as shown in Figure 246 in the appendix *A.40*. In 22 training cycles, the number of labelled training images is increased from 18 to 216. The learning curves show after each training cycle that the average RMSE and the average FPR values decrease and the average accuracy and the average TPR values increase. The average gain in accuracy is 29.16%. In 3 out of 30 runs, the accuracy obtained by active learning exceeds the “ideal” value. The class assignments of the test images for a single run are shown in Figure 151.

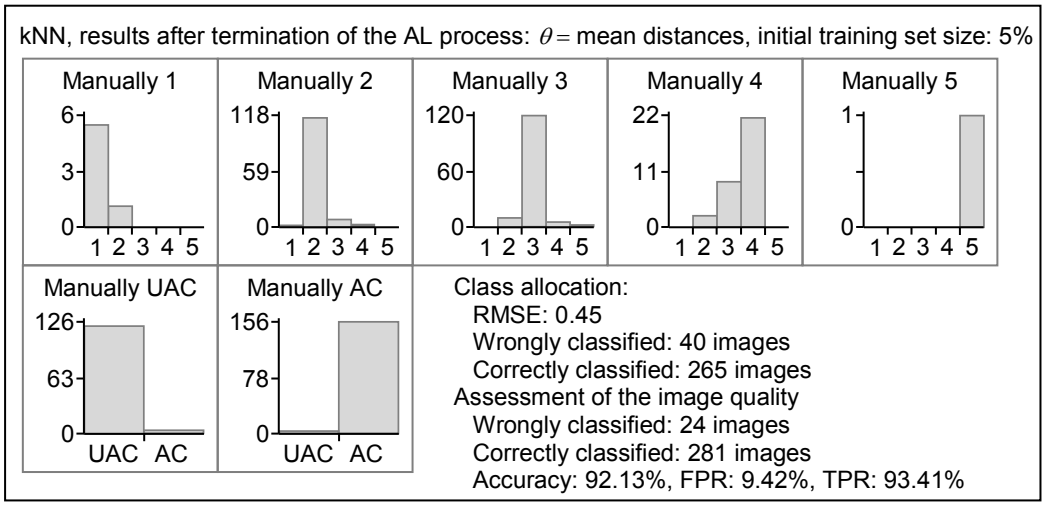


Figure 151: AL, kNN for the distortion and double image dataset: possible labelling of the test images, θ = mean threshold distances

Upon completion of the active learning process, the classifier gives the accuracy of 92.13%, the FPR of 9.43%, the TPR of 93.41% and the RMSE of 0.45. The accuracy is 0.65% higher, the FPR is 2.17% lower, and the TPR on the test images is equal to the corresponding “ideal” values. The assignment of the test images is very similar to the assignment of the classifier trained with all manually labelled training images. The test images are mainly assigned to the desired rating class.
The experimental evaluation shows that the classifier trained with the active learning process performs well on unseen data. Since for individual runs higher accuracy values are obtained, the active learning process selects the most informative training samples that advantageously improve the recognition behaviour. The manual labelling effort can be reduced because the classifier is not overcharged with too many uninformative training images.

Finally, the mapping rule of the NN is reduced to a 2-class problem. The learning behaviour of the 2-class classifier is investigated for different labelled training dataset sizes in the 13-dimensional component space. If all manually labelled training images are available, the classifier reaches an area under the ROC curve of 0.86. An informative training image is manually labelled and transferred into the training dataset if the distance to the nearest reference pattern is less than the mean assignment distance θ.
Figure 152 shows the average AUC values and the standard deviations over 30 runs.
The diagram shows that the average AUC values are greater than 0.88. Regardless of the number of training images initially labelled, the AUC values obtained for each run are above the "ideal" value. AUC values even above 0.9 can be obtained for certain runs. Thus, the active learning algorithm is able to select the most informative training images and to generalise to unknown data.

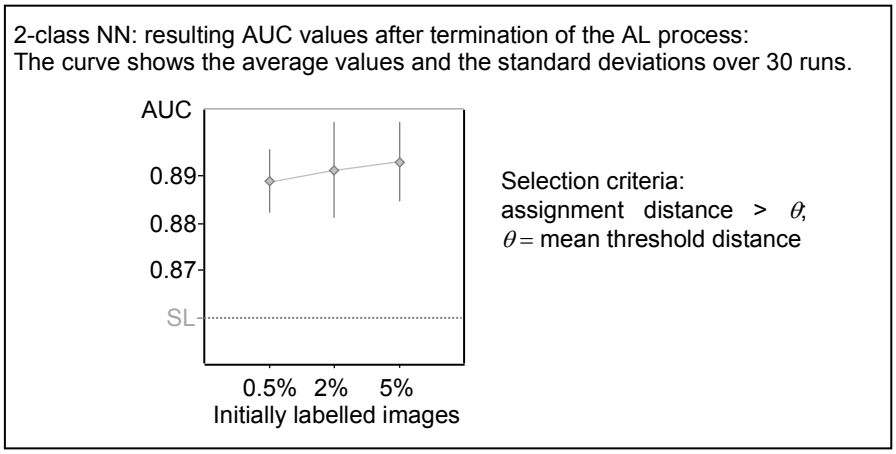


Figure 152: AL, 2-class kNN: results for the distortion and double image dataset, θ = mean threshold distance

Learning vector quantisation:
The active learning behaviour of the LVQ is investigated. The LVQ uses 103 prototypes, which are determined by the LVQ1 learning rule. If all manually labelled training images are used to determine the prototypes the RMSE of 0.56, the accuracy of 85.90%, the

FPR of 13.77%, and the TPR of 85.63% are obtained. A training image is identified as informative if the distance to the next prototype is greater than a threshold distance. The threshold distances are already determined in chapter *7.5.3*; see Table 40. It is important to note that all investigations are all based on the same prototype initialisation.
The obtained results after completion of the AL process are summarised in the appendix *A.41*, see Figure 247 and Figure 248. When applying the minimal or the mean threshold distances, the active learning process results in accuracies and TPR values, which are higher than the corresponding "ideal" values. Likewise, the RMSE values are lower than 0.56. If the minimum threshold distance is applied and the size of the initial training set is set to 5%, the classifier achieves the average FPR that is below 13.77%. If the maximal threshold distances are applied, the results obtained are far away from the "ideal" results that could be explained by too few labelled training images.

The learning behaviour of the classifier is exemplarily investigated for an initial training set size of 2% and maximal threshold distances. The learning curves are shown in the appendix *A.42*; see Figure 240. In 17 training cycles, the number of labelled training images is increased from 9 to 162. The average RMSE and FPR values decrease while the average accuracy values and the average TPR values increase after each training cycle. After 12 training cycles, the accuracy obtained by active learning exceeds the "ideal" value. The average achieved gain in accuracy amounts 10.26%. Likewise, after 5 training cycles, the average TPR exceed 85.63% and the average RMSE values fall below 0.56 after the 12th training cycle. The average FPR values are not below the "ideal" value. A possible labelling of the test images for a single run is shown in Figure 153.

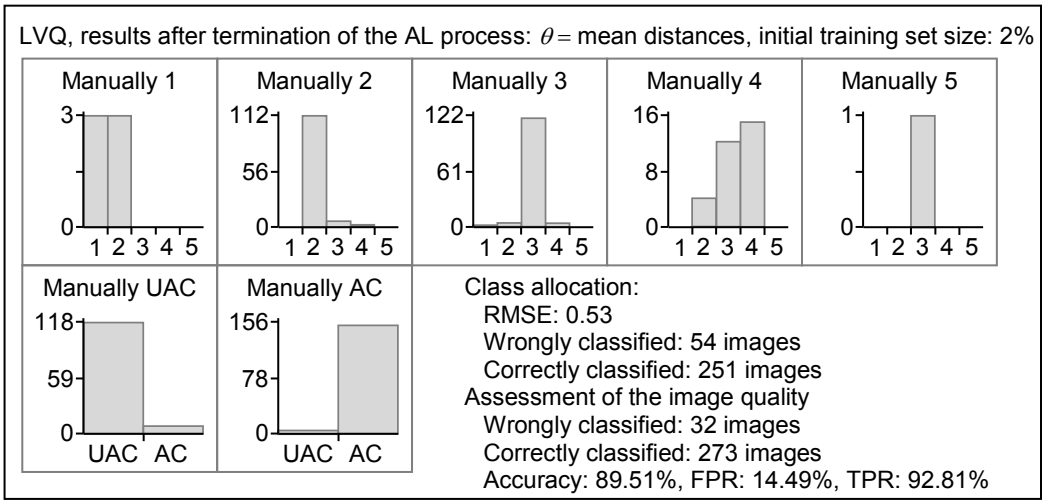


Figure 153: AL, LVQ for the distortion and double image dataset: possible labelling of the test images, $\theta =$ mean threshold distances

The accuracy of the test images is 3.61% higher than that of a classifier using all manually assigned class labels. Likewise, the TPR is 7.18% higher than the "ideal" value. The active learning process tends to reduce costs since fewer vehicles with an ac-

ceptable image quality would be mistakenly sent to rework. In contrast, the FPR is 0.72% higher than the “ideal” value and more vehicles with an unacceptable image quality would wrongly reach the customers. Overall, the labelling of the test images is very similar to the results obtained with all labelled training images.
In general, the experimental evaluation shows that the classifier trained by an active learning process is able to perform well on images that are not presented during training. Selecting the most informative training images may even lead to classifier accuracies that exceed the “ideal” value. Thus, the number of examples for learning the relationship between the labels and the objective features can be less than the number required for passive supervised learning.

Finally, the AL behaviour of the 2-class LVQ is investigated. The 2-class LVQ is based on 103 prototypes, which are determined by the LVQ1 learning rule. The mean threshold distances θ are applied to check if the training image needs to be labelled manually. After each training cycle, the FPR and TPR values are calculated by varying the number *m* of the *k*-nearest prototype vectors required to classify a test image as AC. The average AUC values and the standard deviations over 30 runs are shown in Figure 154. For this purpose, the number of considered nearest prototype vectors is set to k = 10.

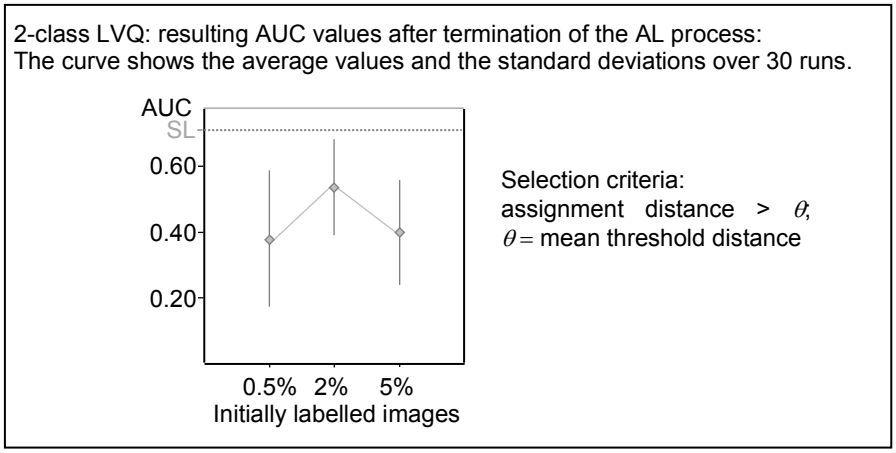


Figure 154: AL, 2-class LVQ: results for the distortion and double image dataset, θ = mean threshold distance

In summary, the diagram shows that the “ideal” AUC value of 0.72 cannot be reached. The resulting areas under the ROC curves are in some cases well below the “ideal” value. Depending on the initially labelled training images, AUC values less than 0.30 are achieved. This low value indicates that the test images are mainly wrongly classified. Perhaps the classifier approximates random noise rather than the correlation between the feature vectors and the manual labels.

Summary of the obtained results:
The experimental evaluation shows that the training set of the distortion and double image dataset contains some uninformative images, which adversely affect the classification accuracy. Depending on the thresholds and the initially labelled training images,

the PC and the kNN achieve accuracy values that slightly exceed the corresponding “ideal” values. Likewise, the manual labelling effort can be reduced if the active learning process is applied to the LVQ. Thus, higher accuracy values can be obtained by using the most informative images for training. If classification problem is reduced to a 2-class assignment task, the polynomial classifier can still generalise well to images not presented during training. In addition, the 2-class kNN can select the most informative training images and apply the acquired knowledge to unlabelled test images. In contrast, the “ideal” area under the ROC curve is not achieved if the active learning approach is applied to the 2-class LVQ.

7.7 Realisation possibilities in the production line

The previous chapters show that classification methods are better suited to assess combined aberration types than the standard method. Therefore, the aim is to implement the classification methods in the production line. This chapter identifies common issues that should be considered before the classification methods can be implemented in the production plant.

First, it has to be clarified whether a passive or a learning algorithm is preferable. The passive algorithm is initially trained with all available labelled training images. During the use in the production line, the algorithm would not learn something new. The HUD image quality of each vehicle would be assessed regardless of the reliability of the rating. Since there is no possibility to identify unreliable assessments, the passive algorithm is not recommended. There are two possibilities are conceivable for implementing a learning algorithm, namely semi-supervised or active.
The semi-supervised algorithm uses its own predictions to teach itself. In the production plant, the image quality evaluation would proceed as follows. A classifier trained on the assessment task is introduced to the production line. If the classifier detects a vehicle for which the probability of a correct label is high, the labelled image is transferred to the training set and a new training cycle is initiated. On the other hand, if the algorithm recognises a vehicle for which the classifier is not sure how to assess the image quality, the vehicle does not receive a label. In addition, the vehicle would be excluded from the production cycle as long as the classifier cannot reliably capture the image quality. Since vehicles could be excluded from the production cycle, this procedure is not recommended.
In contrast, the implementation of the active learning algorithm is recommended. The active learning algorithm can interactively query for the right label. Again, a trained classifier would be transferred to the production line. If the classifier detects a vehicle for which the image quality is most likely be misjudged, the algorithm asks the development engineer for the right label. Based on the reported label, the classifier transfers the image into the training dataset and intends a new training cycle. Thus, the algorithm would be able to adjust itself to a changing classification task. If the algorithm is very confident that the assigned label is correct, the algorithm does not ask for the right label and no new training cycle is intended. Thus, the classifier would also be able to adapt to new classification tasks that may occur, for example, when starting a new model series.

In addition, the selection of the most appropriate classifier type depends on the required accuracy and the maximum complexity. In general, the computational complexity of the LVQ and the PC remains the same regardless of the number of labelled training images. The complexity of the polynomial classifier is characterised by the polyno-

mial order. Since the polynomial order is not influenced by the number of labelled training images, the complexity of the PC remains the same. Likewise, the calculation effort of the LVQ is determined by the number of prototypes. Since the number of prototypes is not changed during training, the complexity of the LVQ is independent of the training set size. In contrast, the kNN becomes computationally complex as the number of training samples increases. Since the classifier uses all training samples as a reference, the calculation effort increases with the number of labelled training images. This can become a problem since in the length of a production-cycle is fixed and the time for each cycle is limited. Therefore, the implementation of the kNN classifier is not recommended.

In general, a classifier type that which minimises the false positive rate (FPR) is recommended for the HUD image assessment of high-end vehicles in the luxury segment. Therefore, a classifier type should be chosen that is able to detect as many vehicles as possible that are not suitable for the customer. The secondary objective is to reduce the rework costs as much as possible. These costs increase when customer suitable vehicles are mistakenly sent to rework because they are incorrectly given a negative label by the classifier. Thus, a high true positive rate (TPR) of the classifier is expected. Unfortunately, the choice of the most suitable classifier type is often a compromise between a low FPR and a high TPR.

Distorted HUD images can be easily corrected by image warping. This indicates that a lower TPR can be accepted since the rework costs are low. In contrast, distortions are perceived before double image are recognised. This leads to the demand for a low FPR to ensure that only images without distortions reach the customer. Therefore, a classifier should be chosen that has a lower FPR rate than the limit analysis. Since the limit value analysis already has a very low FPR of 2.67%, only the polynomial classifier should be considered. The 2-class PC (4^{th} order, 3 principal components, $\phi = 0.1$) achieves the FPR of 1.38% for the distortion dataset. Here, only 1.38% of the images that are manually labelled as unacceptable would wrongly receive a positive label. Likewise, the TPR of 46.98% and the accuracy of 67.78% are achieved. Both values are higher than the values obtained by the limit value analysis. Nevertheless, 53.02% of all costumer suitable vehicles would be mistakenly sent to rework. However, this could be accepted since the effort in the rework process is low.

Double images cannot be corrected. If occurring double images are perceived as annoying, the windscreen needs to be replaced. This procedure is very time consuming and costly. Therefore, it must be avoided in any case that the windscreen is mistakenly replaced if the vehicle does not show double images. Thus, the classifier has to achieve a very high TPR value. Even if the customers perceive double images, they will not accept it and complain about a headache and eyestrain. Therefore, only vehicles without double images should reach the customer and the classifier must receive a

very low FPR. This conflict between a low FPR and a high TPR is best solved by the 2-class polynomial classifier. For the double image dataset, the PC (2^{nd} order, 2 principal components, $\phi = 0.6$) achieves the FPR of 1.36%, and the TPR of 97.18%. The percentage of correspondence between the received labels and the reality is 97.78%. This classifier could easily replace the limit analysis.

If the combination of acceptable distortions and acceptable double images is perceived as not customer suitable, a separate check is made for each vehicle to decide what to do. Either the windscreen is changed or the image is corrected by warping. Again, this is very time consuming and expensive. To minimise these costs, vehicles with an acceptable image quality should not be mistakenly sent to rework. The classifier must achieve a very high TPR value. Furthermore, again a low FPR value must be obtained, because no vehicles with an unacceptable image quality should reach the customers. For the distortion and double image dataset, it is difficult to implement a classifier, which satisfies both requirements. The best compromise is achieved by the LVQ (257 prototypes trained by the LVQ1 learning rule). Here, the FPR of 13.04%, the TPR of 89.82%, and the accuracy of 88.52% is obtained. This result is not ideal but obviously exceeds the result of the limit analysis.

Chapter

8 Summary and conclusion

This thesis describes a new method to evaluate the perceived quality of HUD images. The HUD supports as HMI component the primary driving task. HMI components must meet the highest quality standards in order not to interfere with the driving task. The idea of the head-up display is to have all relevant information directly in the field of vision while driving. The HUD consists of a picture generation unit and complex mirror optics. The generated light rays are reflected by the mirrors and the windscreen. Due to the optical perception, the virtual image seems to float just above the hood. The quality of the virtual images is affected by assembly tolerances of the head-up display in the dashboard, the windscreen in the vehicle and by a poor production of the components involved. In this thesis, various aberration types such as distortions and double images are considered. Distortions describe the difference between the actual and the desired image geometry. Double images arise because the light is reflected at the inner and outer sides of the windscreen.

The assessment of the perceived image quality is done gradually. First, the perception of distortions and double images is assessed separately, since the different aberration types can be corrected by different measures. Only when the separate analysis of distortions and double images is successful, the combination of both aberration errors is assessed. The implementation of the assessment algorithm is based on 136305 images (71895 distorted images, 64410 images with different double images), which are specially generated for this work in the existing laboratory setup of the Daimler AG. The images show an already established test pattern, consisting of 9 x 21 measuring marks, which can be detected by accurate and robust image processing routines. The results of the image processing routines are transferred into 21 objective features, which numerically capture the occurring aberration errors. 13 objective features are needed to describe occurring distortions. The remaining 8 objective features capture appearing double images. The main task of the assessment algorithm is to approximate the underlying relationship between the objective features and the subjective impression.
The subjective impression is recorded by the double-stimulus impairment scale method. This method is usually used to assess the subjective perception of television images. For this work, the method is now adapted to quantify the perceived quality of HUD images. The overall impression is rated according to the 5-grade impairment scale. For

this thesis, 12 test persons, who drive daily several kilometres and are enthusiastic about technology, label representative images.

The standard procedure for determining the customer suitability of HUD images is the limit analysis. The aim is to define ergonomic limits for the objective features. Compliance with these limits ensures an impairment-free reading of the information in the virtual image. The standard method has a big weakness. Even simple combinations of various aberration types cannot be assessed. Therefore, classification methods are applied which no longer show this weak point. Classification tasks are divided into the training and the testing phase. In the training phase, the relationship between the subjective labels and the objective feature vectors is detected. Therefore, a training set consisting of representative labelled images is necessary. In the testing phase, the trained algorithm is used to label unlabelled test data. This determines how well the classifier performs on unseen data. Here, the kNN (k-nearest neighbour classifier), the LVQ (learning vector quantisation) and the PC (polynomial classifier) are implemented as assessment algorithms.
The kNN classifier uses reference patterns to classify an unlabelled image. The complete training data are used as reference. The class assignment of an unknown image is determined by the class labels of the nearest reference images. The LVQ uses prototype vectors to classify unlabelled images. After training, the prototype vectors approximate the underlying distribution of training samples. According to the competitive winner-takes-it-all strategy, an unlabelled input sample is assigned to the same class to which the nearest prototype vector belongs. The PC adapts the relationship between the subjective labels and the objective features to decision equations. An unlabelled image is assigned to the class which the highest probability.

The selection of the representative training images is done by clustering methods. During clustering, similar images are grouped. The aim is that each group contains only images with the same subjective assessment. In addition, the labels of a single image should correspond to the labels of all other images in the same group.
Some preliminary investigations show that only 5 objective features are needed to describe the perceived difference in distorted HUD images. In contrast, all 8 characteristic features are needed to describe the subjectively perceived difference between double images. The images are sorted correctly if the maximum difference of the relevant feature values within the clusters is less than 1 HUD pixel. If a suitable cluster solution is found, it is assumed that the representative images are close to the cluster centres. Thus, the distortion dataset and the double image dataset include 1006 and 345 images, respectively. In addition, each dataset includes 360 test images that are randomly selected from the available images. The images of the datasets are manually labelled by 12 test persons. Then the dataset for the combined aberration types is determined, comprising of 303 training images and 305 test images, which are also labelled by the 12 test persons.

Based on the labelled training images, the simple limit analysis is implemented first. The distortion dataset achieves the accuracy of 56.11%, the FPR of 2.76% and the TPR of 28.37%. The accuracy of 89.44%, the FPR of 8.16% and the TPR of 87.79% are obtained for the double image dataset. The distortion and double image dataset achieves the accuracy of 67.21%, the FPR of 55.80% and the TPR of 86.23%. In the second step, supervised learning methods are applied and checked whether these methods are more suitable for estimating the subjectively perceived image quality.

For the distortion dataset, classifiers can be implemented that achieve a higher accuracy than the standard method. These classifiers are, for example, the 2nd order PC in the 10-dimensional feature space, the kNN classifier for $k = 1$ and 3 principal components, and the LVQ for 165 prototypes trained by the LVQ1 learning rule. The highest accuracy of 88.06% is achieved by the LVQ, followed by the PC with the accuracy of 83.89%, and the accuracy of the kNN classifier is 57.22%. Thus, the supervised learning algorithms, especially the LVQ and the PC are able to assess combinations of various distortion types. In addition, the proposed PC and LVQ algorithms achieve TPR values greater than 99%. Unfortunately, both classifiers give FPR values of more than 29%. In contrast, the proposed kNN classifier gives the FPR of 6.21%, but also the low TPR of 32.56%. Here, the choice of the appropriate assessment algorithm is a compromise between a high TPR and a low FPR.

Similar results are also obtained for the double image dataset. The highest accuracy of 95.28% is achieved by the nearest neighbour classifier in the 7-dimensional feature space, taking into account 3 nearest neighbours. This classifier also achieves the FPR of 0.68% and the TPR of 92.49%. Thus, this classifier yields better results for the given classification problem than the limit analysis. Likewise, the 3rd order PC in the 2-dimensional feature space achieves the accuracy of 90.00%. These classifier types are able to assess the combination of various double image types very well. The accuracy obtained for the LVQ algorithm is 83.06%. This value is below the accuracy of the limit analysis. Thus, the LVQ is less suitable for the given classification problem.

For the distortion and double image dataset, the kNN classifier achieves the accuracy of 91.48%, the FPR of 11.59%, and the TPR of 94.01% for $k = 1$ and 13 principal components. Thus, this classifier is able to reliably assess the combination of various aberration types and is thus better suited for the given classification problem than the limit analysis. The 1st order PC yields the accuracy of 71.48% in the 20 principal components space. The corresponding FPR and the TPR are lower than the values obtained from the standard method. By implementing the LVQ, classifiers can be determined that achieve lower FPR values than the standard method. If the number of prototypes is chosen to be large enough, TPR values are also obtained that are higher than the TPR of the limit analysis.

As a general result, the experimental evaluation shows that classification methods are well suited to assess the subjectively perceived quality of HUD images. Depending on the classifier type, higher accuracy values are obtained on the test images than for the limit value consideration. Thus, the supervised learning methods appear to be able to assess combinations of various aberration types.

In the next step, the classification rule is reduced to a 2-class problem, as is the case with the limit analysis. The customer suitability of the HUD images is only assessed as acceptable (rating classes 3, 4 and 5) or unacceptable (rating classes 1 and 2). By varying the classification parameters, the receiver operator characteristic, the so-called ROC curve, of the classifier is determined. The ROC curve is interpolated by plotting the TPR against the FPR. The closer this curve is to the upper left corner, the larger the area under the curve (AUC), and the better the performance of the classifier.
The obtained ROC curves of the classifiers show that the choice of the most suitable classifier is a compromise between a high TPR and low FPR. For the distortion dataset, the 2-class classifiers achieve AUC values greater than 0.86. Depending on the classification parameters, these classifiers reach better classification results than the standard limit consideration. Almost perfect ROC curves are obtained for the double image dataset. Here, the 2-class classifiers reach AUC values greater than 0.96. Depending on the selected parameters, the classifiers work more accurately than the limit analysis. In contrast, for the distortion and double image dataset, only the 2-class LVQ can achieve better results than the limit analysis. For this classifier type, an area under the ROC curve of 0.72 is obtained.

The classifiers used require a large number of labelled training samples in order to obtain a comprehensive and generalising recognition behaviour. However, the manual labelling of large training datasets is costly and time-consuming. Therefore, the semi-supervised learning (SSL) rule is applied to selected classifiers. It is investigated if the manual labelling effort can be reduced by combining labelled and unlabelled training images. SSL is an iterative procedure in which the learning process uses its own predictions to teach itself. Here, different criteria are applied for the rejection or the acceptance of the autonomously generated labels for unlabelled images. All accuracies on the test set are compared to the accuracy of an “ideal” classifier trained with all available manually labelled training samples.
The experimental evaluation shows that unlabelled data, when used in conjunction with a small amount of labelled data, can lead to an improvement in the classification quality. The reason could be that the SSL process attempt to avoid the use of poorly labelled training samples. Those samples are selected that improve the recognition behaviour in an advantageous manner. Thus, the manual label effort can be significantly reduced without a loss of classification quality. Depending on the dataset, the classifiers have different learning behaviours.
The distortion dataset implies that the PC and the kNN classifiers yield accuracy values that exceed the corresponding values of an “ideal” classifier trained with all available labelled training samples. For example, if the size of the initially labelled training set is set to 40%, the PC reaches achieves the accuracy of 85.28% and the kNN an accuracy of 63.06%, depending on the criteria applied. Similar results are obtained for the double image dataset. The PC shows a good learning behaviour. Depending on the applied criteria, accuracy values exceeding the “ideal” value are achieved. If the size of the initial training set is set to 15%, the accuracy of the PC is 98.06%. For the distortion

and double image dataset, the LVQ classifier shows a good learning behaviour. Here, the SSL process results in accuracy values that exceed the “ideal” value, depending on the applied criteria. If the initial training set size is set to 35%, the LVQ reaches an accuracy of 87.54% after completing the SSL process.
For the distortion dataset, only the 2-class PC achieves higher AUC values with a small amount of manually labelled training images. For the 2-class kNN and the 2-class LVQ, a reduction in the area under the ROC curves must be considered when reducing the manual labelling effort. If the allocation task of the double image dataset is limited to acceptable or unacceptable, the 2-class classifiers achieve AUC values that are below the “ideal” values. Here, the manual labelling effort can only be reduced if a decrease in the AUC values can be accepted. For the distortion and double image dataset, only the 2-class LVQ can better evaluate the HUD image quality with a small number of manually labelled images than the “ideal” classifier

Finally, it is checked if the labelling effort can also be reduced by an active learning (AL) procedure. An active learning algorithm can select the most informative training images that are manually labelled by an Oracle and transferred to the training dataset. In this thesis, different criteria, for ignorance or query for the right class label, are applied.
The experimental evaluation shows that the AL process can lead to an improvement in the classification quality when using the most informative training samples. During the AL process, the manual labelling effort can be reduced to some informative training images. The training sets probably include many uninformative images. Since the AL process gives higher recognition accuracies than classifiers trained with all labelled training images, the uninformative images adversely affect the classification accuracy. The classifiers seem to be overcharged with the uninformative images.
For the distortion dataset, the PC provides the TPR of 100%, and the accuracy of 86.11%, depending on the criteria for the query for the right label. If the kNN and the LVQ are allowed to select the most informative training images, accuracy values of 65.28% and 89.17% are also obtained on the test images. For the double image dataset, the classifiers show different learning successes. In the course of the AL process, the recognition accuracies of the PC and the kNN increase. However, the “ideal” values are not or only just reached. Therefore, the 2 classifiers are barely able to extract sufficient information to generalise to unknown data. In contrast, the LVQ algorithm provides the recognition accuracy of 85.28%, which is higher than the “ideal” value, depending on the applied criteria. Similarly, the distortion and double image dataset includes many uninformative images. Depending on the active learning criteria used, the PC gives the accuracy of 72.79%, the kNN the accuracy of 92.13%, and the LVQ the accuracy of 89.51%. Thus, the accuracy values of these classifiers exceed the corresponding “ideal” values.
When classification problem is reduced to a 2-class assignment task, the classifiers for the distortion dataset and the double image dataset do not seem to be able to identify the most informative training images. On average, the “ideal” areas under the ROC curves are not achieved and the labelling effort can only be reduced if a decrease in

the recognition accuracy can be accepted. In contrast, the 2-class PC and the 2-class kNN for the distortion and double image dataset are still able to generalise well to images not presented during training.

Since the classification methods are well suited to assess combinations of different aberration types, the implementation in the production line seems to make sense. It is recommended that an active learning algorithm should be implemented in order to be able to adapt the algorithm to changing classification tasks. The choice of the most suited classifier type depends on the required accuracy and the maximum complexity.
The computational complexity of the LVQ and the PC remains the same regardless of the number of labelled training images. In contrast, as the number of training samples increases, the kNN becomes more and more complex. Since the length of a production cycle is fixed and the time for each cycle is limited, the implementation of the kNN classifier does not seem sensible.
To assess the HUD image quality of vehicles in the luxury segment, it is necessary to implement classifiers that minimize the FPR value. This is because the number of vehicles with an unacceptable image quality that reach the customers should be as low as possible. For the distortion dataset, the 2-class PC seems to be best suited. Here, the classifier achieves the FPR of 1.38% and the TPR of 46.98%. The low TPR could be accepted because the effort in the rework process is low. For the other datasets, classifiers are intended that achieve both a high TPR and a low FPR. The requirement for the high TPR is based on the time consuming and costly rework process. The conflict between a low FPR and a high TPR is best resolved for the double image dataset by the 2-class PC. The 2-class PC achieves the FPR of 1.36% and the TPR of 97.18%. The best compromise for the distortion and double image dataset is achieved by the LVQ. Here the FPR of 13.04% and the TPR of 89.82% are obtained.
Thus, an assessment system could be implemented that supports the production process and contributes to the business objective, which requires that only high-quality vehicles should leave the production line.

Future work could be the development of an “ensemble” of 2 or 3 different classifiers, with the aim of further improving the classification accuracy. In addition, the learning behaviour of the active learning algorithms should be improved by a special adaptation of the threshold criteria. Finally, the evaluation algorithm should be extended to unconsidered aberrations such as astigmatism, dynamic variance, and binocular misalignment.
Since it is shown that classification methods are better suited to assess the perceived image quality than the standard method, the ultimate goal is the implementation of the classification methods in the production line.

Chapter

IV Bibliography

[ATTESLANDER 03] ..... ATTESLANDER, Peter. *Methoden der empirischen Sozialforschung*, *10. Auflage*. Walter de Gruyter, 2003.

[BACKHAUS et al. 00]..... BACKHAUS, Klaus; ERICHSON, Bernd; PLINKE, Wulff; WEIBER, Rolf. *Multivariante Analysemethoden, 9. Auflage*. Springer Verlag, 2000.

[BILLE & SCHLEGEL 05] ..... BILLE, Josef; SCHLEGEL Wolfgang. *Medizinische Physik 3*. Springer Verlag, 2005.

[BLUME et al. 13] ..... BLUME, Jochen, et al. *Head-up display next generation with augmented reality*. *ATZelektronik worldwide*, 2013, 8. Jg., Nr. 4, S. 4-7.

[BORTZ & SCHUSTER 10]..... BORTZ, Jürgen; SCHUSTER, Christof. *Statistik für Human-und Sozialwissenschaftler, 7. Auflage, Springer-Lehrbuch*. Springer, Verlag, 2010.

[BROWNLEE 14]..... BROWNLEE, Jason. *An introduction to feature selection.* http:// machinelearningmastery.com/ an-introduction-to-feature-selection/, 2014.

[CEBRON 08]..... CEBRON, Nicolas. *Aktives Lernen zur Klassifikation großer Datenmengen mittels Exploration und Spezialisierung*. Universität Konstanz, 2008.

[CHENG 95]..... CHENG, Yizong. *Mean shift, mode seeking, and clustering*. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 1995, 17. Jg., Nr. 8, S. 790-799.

[CONTI 14] ..... CONTINENTAL. H*ead-up displays – Eying safety and comfort at all times*, www.conti-online.com, 2014.

[CONTI MICRO 14] ..... CONTINENTAL. *Microsite*, http://continental-head-up-display.de, 2014.

[CUNNINGHAM & DELANY 07] .. CUNNINGHAM, Padraig; DELANY, Sarah Jane. *k-Nearest neighbour classifiers*. In: Multiple Classifier Systems. S. 1-17, 2007.

[CURCIO & ALLEN 90] ..... CURCIO, Christine A.; ALLEN, Kimberly A. *Topography of ganglion cells in human retina*. Journal of Comparative Neurology, 1990, 300. Jg., Nr. 1, S. 5-25.

[DAHM 06]..... DAHM, Markus. *Grundlagen der Mensch-computer-interaktion*. New York: Pearson Studium, 2006.

[DAIMLER AG] ..... DAIMLER AG. *Bildmaterial* http://media.daimler.com, 2012.

[DEMANT et al. 10] ..... DEMANT et al. *Industrielle Bildverarbeitung*, *3. Auflage*. Springer, Berlin-Heidelberg, 2010.

[DIAZ 05] ..... DÍAZ, Luis Sampedro. *Optical aberrations in Head-Up Displays*. Universidad Pontificia Comillas Madrid, 2005.

[DIEKMANN 95] ..... DIEKMANN, Andreas. *Empirische Sozialforschung - Grundlagen, Methoden, Anwendungen*. Rowohlts Enzyklopädie, 1995.

[DIVAKARAN et al. 15] ..... DIVAKARAN, D. M., SU, L., LIAU, Y. S., & THING, V. L.. SLIC: Self-Learning Intelligent Classifier for ne2rk traffic. *Computer Ne2rks*, 2015, 91. Jg., S. 283-297.

[EICHHORN & ZINK 12] ..... EICHHORN, Norbert; ZINK, Oliver. *Auto Head-up displays: "View-Through" for Drivers.* GefaSoft Vision and More, 2012.

[FELDER 11] ..... FELDER, Helmut. *Autoelektrik Grundlagen- und Fachwissen*. 2011.

[GABLER 14] ..... GABLER Wirtschaftslexikon. *Mittlerer quadratischer Vorhersagefehler.* http://wirtschaftslexikon.gabler.de/Archiv/88972/mittlerer-quadratischer-vorhersagefehler-v8.html, Springer Gabler Verlag, 2014.

[GENUIT 10] ..... GENUIT, Klaus. *Sound-Engineering im Automobilbereich: Methoden zur Messung und Auswertung von Geräuschen und Schwingungen*. Springer Verlag, 2010.

[GISH & STAPLIN 95] ..... GISH, Kenneth W.; STAPLIN, Loren. *Human factors aspects of using head up displays in automobiles: A review of the literature*. 1995.

[GÖTZE & BENGLER 15] ..... GÖTZE, Martin; BENGLER, Klaus. Urban Driving: Where to Present What Types of Information–Comparison of Head-Down and Head-Up Displays. In: *International Conference on Human Interface and the Management of Information*. Springer International Publishing, 2015. S. 190-200.

[GREHN 08] ..... GREHN, Franz. *Augenheilkunde, 31. Auflage*. Springer Medizin Verlag Heidelberg, 2008.

[HÄDER 10] ..... HÄDER, Michael. *Empirische Sozialforschung, 2. Auflage*. VS Verlag für Sozialwissenschaften/GWV Fachverlage GmbH, Wiesbaden, 2010.

[HADY & FAROUK 11] ..... HADY, Abdel; FAROUK, Mohamed. *Semi-supervised learning with committees: exploiting unlabeled data using ensemble learning algorithms.* 2011. Doktorarbeit. Universität Ulm.

[HAN 05] ..... HAN, Dongil. Real-Time digital image warping for display distortion correction. In: *Image Analysis and Recognition*. Springer Verlag, 2005. S. 1258-1265.

[HAQUE et al. 13] ..... HAQUE, M. M., HOLDER, L. B., SKINNER, M. K., & COOK, D. J. Generalized query-based active learning to identify differentially methylated regions in DNA. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2013, 10. Jg., Nr. 3, S. 632-644.

[HARTEN 74] ..... HARTEN, Hans-Ulrich. *Physik für Mediziner, 13. Auflage*. Springer Verlag, 1974.

[HELLMANN 12] ..... HELLMANN, Markus. *Gütemaß für die Bewertung der Abbildung von Head-up Displays*. Technische Universität Dortmund, 2012.

[HERING et al. 09] ..... HERING, Ekbert, et al. *Taschenbuch der Mathematik und Physik*. Berlin: Springer Verlag, 2009.

[ITU-R BT 12]..... RECOMMENDATION, ITU-R BT.500-13. *Methodology for the subjective assessment of the quality of television pictures*. Standardization Sector of ITU, 2012.

[JIANG et al. 06] ..... JIANG Xiaoyi, et al. *Klassifikation mit Distanzfunktionen*. Vorlesungsskript Mustererkennung of the University Münster, Winter semester 2006.

[JORDAN 13] ..... JORDAN, Markus. *Details zum Head-up Display im S65 AMG – Auflösung von 480x240 Pixel mit Leuchtdichte von 10.000 Cd/m²*. 2013.

[JORDAN 14] ..... JORDAN, Markus. *Mercedes-Benz Accessories mit Head Up Display als Nachrüstlösung ab 2. Quartal 2014*. 2014.

[KAUFMANN 04] ..... KAUFMANN, Joachim. *Head-up Display: Neue Technik für mehr Verkehrssicherheit*. 2004.

[KEM 06] ..... KEM, Konstrukion, Entwicklung, Management. *Head-up Displays.* http://www.kem.de, 2006.

[KOHONEN 01]..... KOHONEN, Teuvo. *Self-organizing maps, 3. Auflage.* Springer, Berlin Heidelberg New York, 2001.

[KÖPPL et al. 16]..... KÖPPL, Sonja, et a. *Evaluation of the individually perceived quality from head-up display images relating to distortions*. In: Studies in classification, Data analysis and knowledge organisation, Status: Reviewed and accepted, 2016

[KRISTIAN et al. 11] ..... KRISTIAN, K., et al. *Statistische Informationstechnik Signal – und Mustererkennung, Parameter- und Signalschätzung*. Springer Verlag, 2011.

[KRUSE 14] ..... KRUSE, Jochen. *Das Head-up Display projiziert Informationen direkt ins Blickfeld des Fahrers und erhöht damit vor allem Sicherheit und Komfort,* 2014.

[LÜDERS 06]..... LÜDERS, Klaus (Hg.). *Pohls Einführung in die Physik - Elektrizitätslehre und Optik, 22. Auflage*. Springer Verlag, 2006.

[MARSLAND 11] ..... MARSLAND, Stephen. *Machine learning: an algorithmic perspective*. CRC Press, 2011.

[MEROTH & TOLG 07] ..... MEROTH, Ansgar; TOLG, Boris. *Infotainmentsysteme im Kraftfahrzeug*. Springer Verlag, 2007.

[MEYER 08]..... MEYER, Hendrik-Marten. *Empfindlichkeit, Adaptation und Schnelligkeit von Photorezeptoren*. 2008.

[MILICIC 10] ..... MILICIC, Natasa. *Sichere und ergonomische Nutzung von Head-Up Displays im Fahrzeug*. Technische Universität München, 2010.

[NEUE STATISTIK 03] ..... PROJEKT, *Neue Statistik 2003*. Freie Universität Berlin, Center für Digitale Systeme,

http://web.neuestatistik.de/inhalte_web/content/files/modul_23196.pdf, 2003.

[NEUMANN 12]..... NEUMANN, Alexander. *Simulationsbasierte Messtechnik zur Prüfung von Head-up Displays*. Technische Universität München, 2012.

[OTT & POGANY 09] ..... OTT, Peter; POGANY, Peter. *Optical design of head-up displays using CAD-compatible freeform surfaces. Photonik international*, 2009/1. P.42-45.

[PERSELLO & BRUZZONE 12]... PERSELLO, Claudio; BRUZZONE, Lorenzo. Active learning for domain adaptation in the supervised classification of remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 2012, 50. Jg., Nr. 11, S. 4468-4483.

[PFEIFER & SCHMITT 14]..... PFEIFER, Tilo; SCHMITT, Robert (Hg.). *Masing Handbuch Qualitätsmanagement*. Carl Hanser Verlag GmbH Co KG, 2014.

[SAKIC 12]..... SAKIC, Denijel. *Semiüberwachtes Lernen mit Ensemble-Methoden zur Erkennung von Gesten*. Technische Universität Dortmund - Fakultät für Informatik, Diplomarbeit, 2012

[SCHÄFER 09]..... SCHÄFER, Thomas. *Clusteranalyse*. *Lecture Notes.* https://www.tu-chemnitz.de/hsw/psychologie/professuren /method/homepages/ts/methodenlehre/meth11.pdf, 2009.

[SCHNEID 09] ..... SCHNEID, Marcus. *Entwicklung und Erprobung eines kontaktanalogen Head-up Displays im Fahrzeug*. Technische Universität München, 2009.

[SCHUMM & WORZISCHEK 11] . SCHUMM, Tobias; WORZISCHEK, Ralf. *Serienfertigung von Head-up Displays. ATZproduktion*, 2011, 4. Jg., Nr. 4, S. 32-37.

[SCHÜRMANN 96]..... SCHÜRMANN, Jürgen. *Pattern classification: a unified view of statistical and neural approaches*. New York: Wiley, 1996.

[SEIFERT 05] ..... SEIFFERT, Ulrich. *Vieweg Handbuch Kraftfahrzeugtechnik, 7. Auflage*. Springer Verlag, 2005.

[SETTLES 09]..... SETTLES, Burr. *Active learning literature survey. Computer Science technical report 1648.* University of Wisconsin-Madison, 2009.

[SMIA 04] ..... NOKIA & ST. *Camera characterisation specification*. SMIA 1.0 Part 5, 2004.

[SZELISKI 10]..... SZELISKI, Richard. *Computer vision: algorithms and applications*. Springer Science & Business Media, 2010.

[THÖNIß 04]..... THÖNIß, Thomas. *Abbildungsfehler und Abbildungsleistung optischer Systeme*. Technische Optik in der Praxis*,* Göttingen, 2004.

[WANG et al. 11] ..... WANG, Jian-Gang, et al. *Active learning with the furthest nearest neighbor criterion for facial age estimation*. In: *Computer Vision–ACCV 2010*. Springer Berlin Heidelberg, 2011. S. 11-24.

[WINTER 42]..... WINTER, Heinrich. *Physik und Chemie, 2. Auflage*. Springer Verlag, 1942.

[WOLLBERG 90]..... WOLBERG, George. *Digital image warping*. Los Alamitos, CA: IEEE computer society press, 1990.

[WUTTKE et al. 14] ..... WUTTKE, Sebastian, et al. *Bewertung von Strategien des aktiven Lernens am Beispiel der Landbedeckungsklassifikation*. Gemeinsame Tagung der DGfK, 2014.

[YEH & GALLAGHER 08] ..... YEH, Flora Yu-Hui; GALLAGHER, Marcus. An empirical study of the sample size variability of optimal active learning using Gaussian process regression. In: *2008 IEEE International Joint Conference on Neural Ne2rks*. IEEE, 2008. S. 3787-3794.

[ZHU & GOLDBERG 09]..... ZHU, Xiaojin; GOLDBERG, Andrew B. *Introduction to semi-supervised learning*. Synthesis lectures on artificial intelligence and machine learning, 2009, 3. Jg., Nr. 1, S. 1-130.

Chapter

A Appendix

A.1 Used abbreviations

AC..... Acceptable image quality
AL ..... Active learning
ALM ..... Acceptance limit
AUC ..... Area under curve
FPR ..... False positive rate
HMB..... Head motion box
HMI ..... Human-machine-interface
HUD..... Head-up display
LVQ ..... Learning vector quantisation
NN ..... Nearest neighbour classifier
PC..... Polynomial classifier
PCA ..... Principal component analysis
PLM ..... Perceptual limit
PVB ..... Polyvinyl butyral
RMSE ..... Root mean square error
ROC..... Receiver operator characteristic
SL ..... Supervised learning
SSL..... Semi-supervised learning
TPR ..... True positive rate
UAC ..... Unacceptable image quality
UL ..... Unsupervised learning

A.2 Frequency distribution diagrams for distortion types

71895 images are available for the investigation of distortion. To get an overview of occurring distortion type sizes, the frequency distribution diagrams are shown in Figure 155. In addition, the perception limits for each distortion type are highlighted. The perception limits are introduced in Table 5.

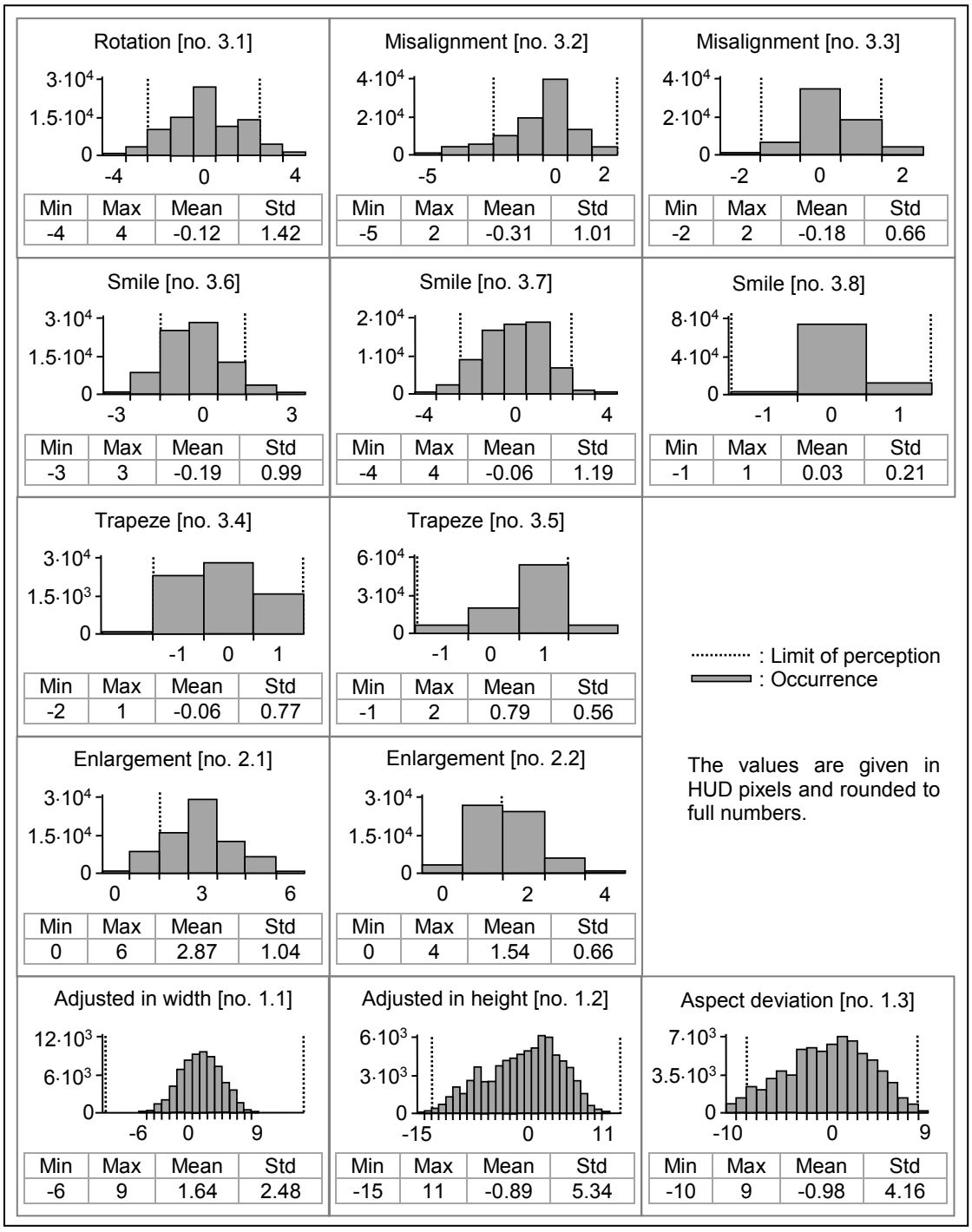


Figure 155: frequency distribution diagrams for distortion

A.3 Frequency distribution diagrams for double image types

64410 images could be used for the assessment of double images. To get an overview of occurring double image sizes the frequency distribution diagrams are shown in Figure 156. Additionally, the perception limits for each double image type are highlighted. The perception limits are introduced in Table 9.

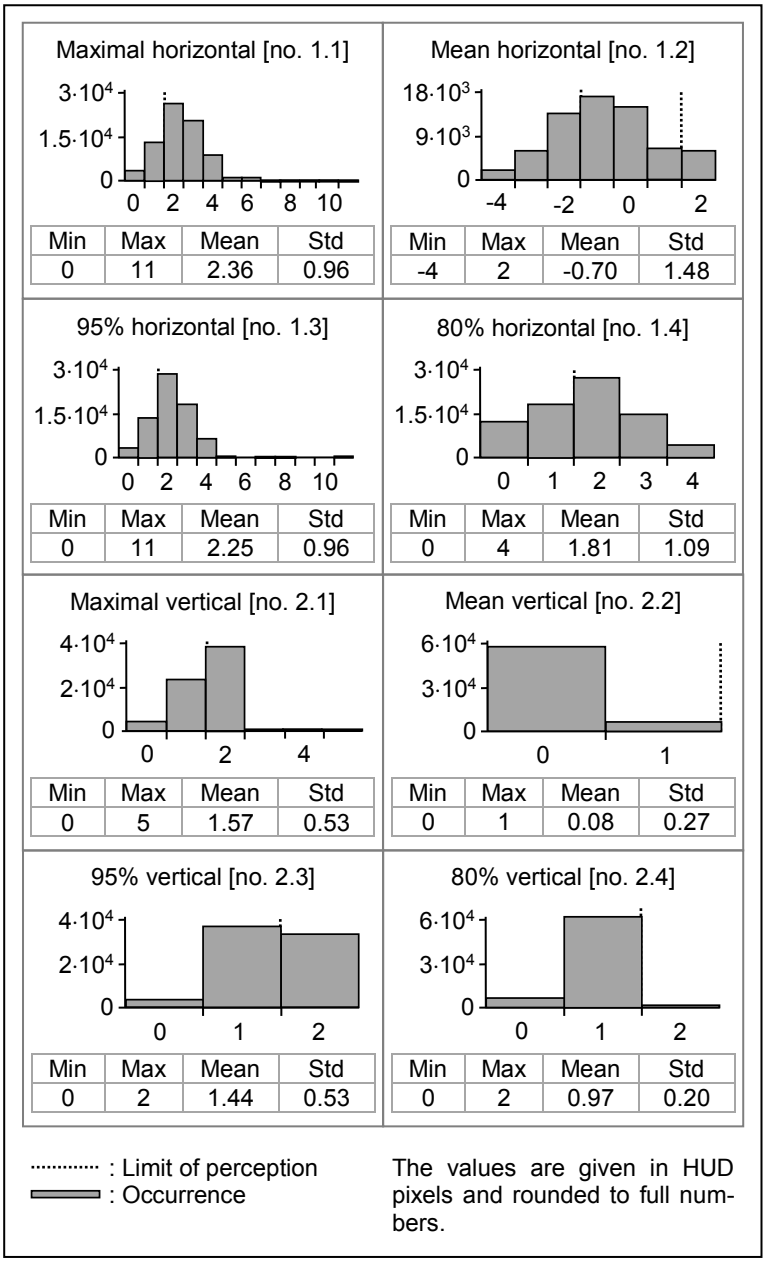


Figure 156: frequency distribution diagrams for double images

A.4 Distortion: examples of different rated images

An example image for each evaluation class is shown in Figure 157. For simplicity, the images are shown inverted. The images labelled with 5, 4 or 3 are still classified as acceptable and show minimal distortions. In contrast, the images labelled with 2 or 1 are classified as unacceptable and show clearly visible distortions.

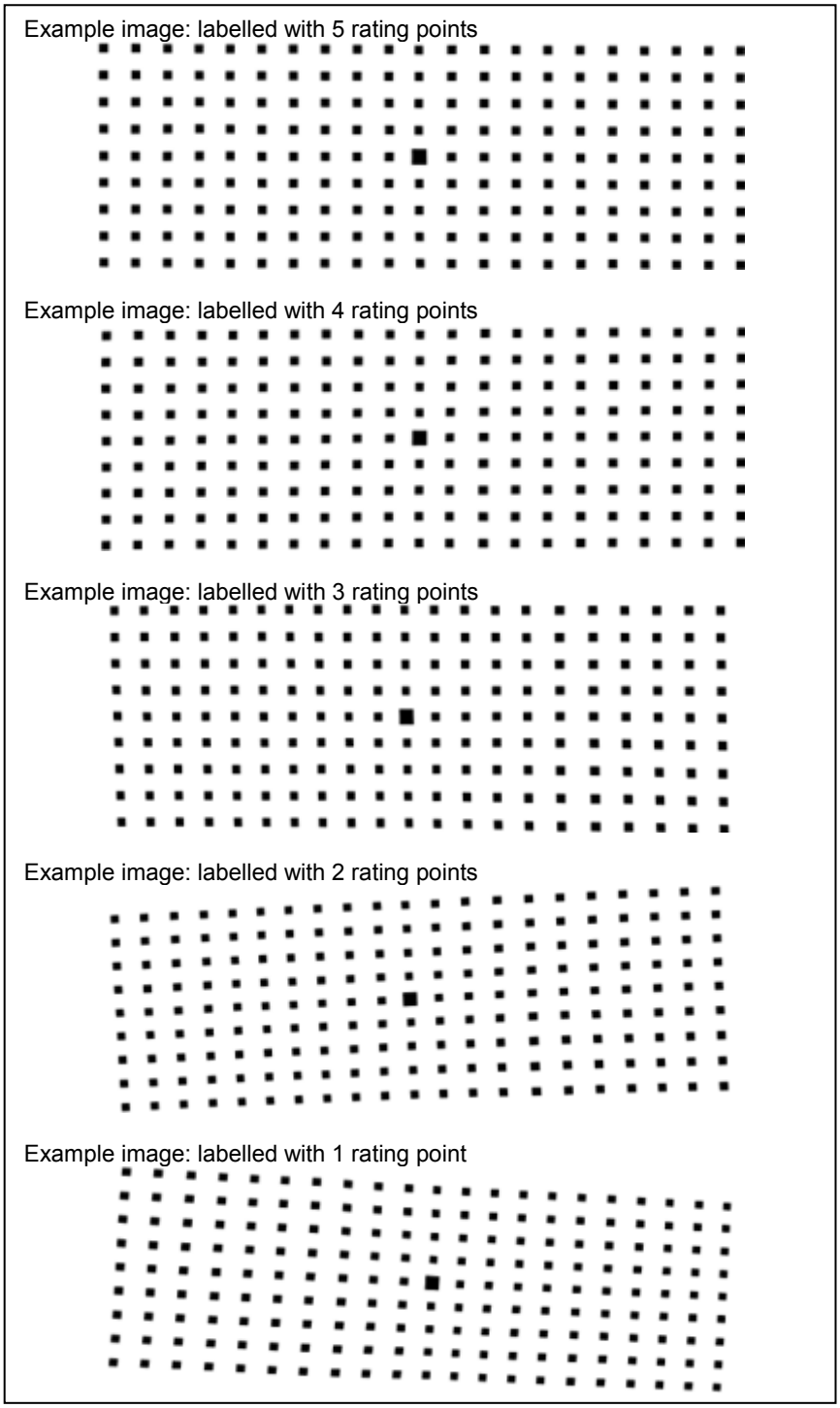


Figure 157: distortion: example images of different rating classes

A.5 Double images: examples of different rated images

An example image for each evaluation class is shown in Figure 158. For simplicity, the images are shown inverted. The images labelled with 5, 4 or 3 are still classified as acceptable and show minimal double images. In contrast, the images labelled with 2 or 1 are classified as unacceptable and show clearly visible double images.

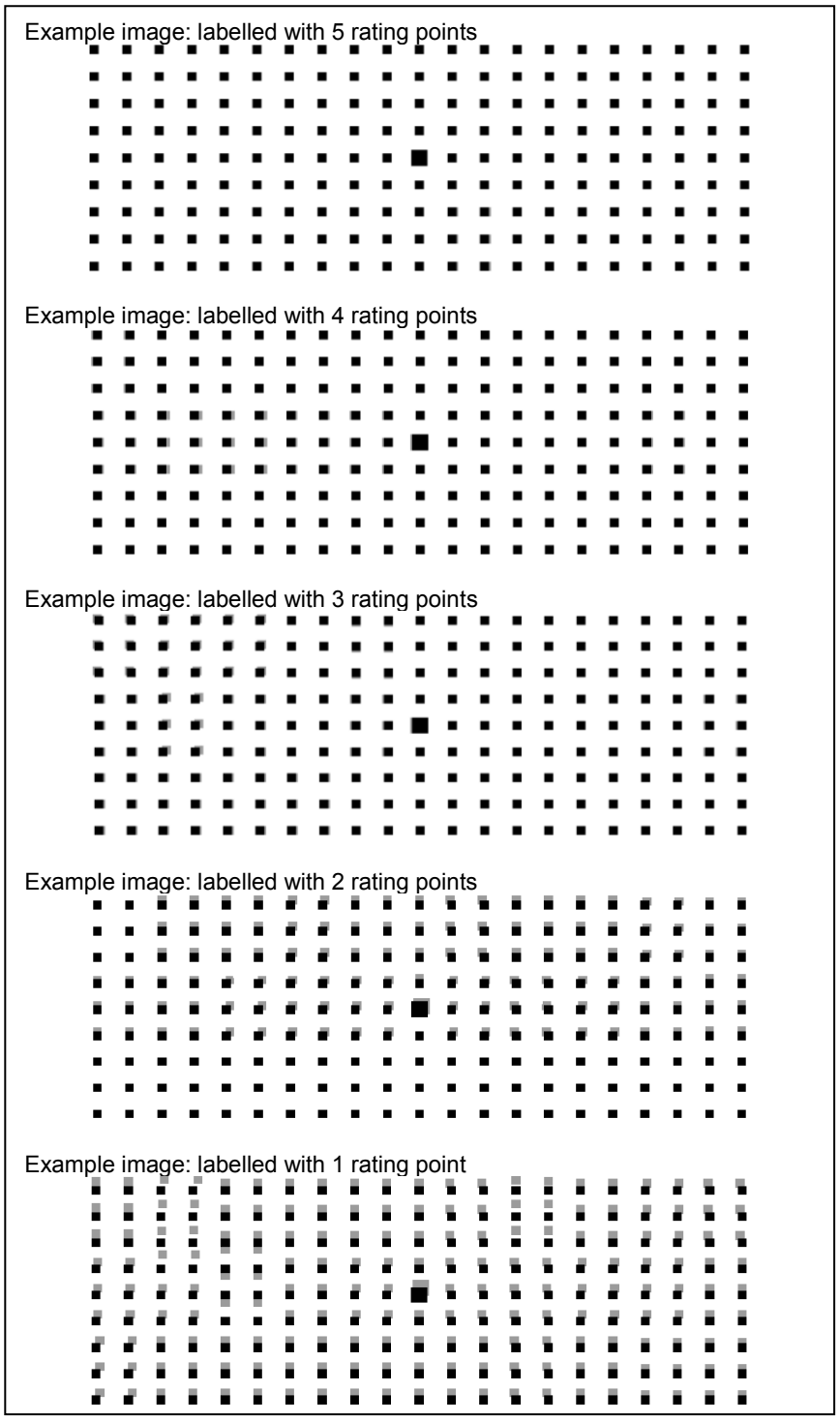


Figure 158: double images: example images of different rating classes

A.6 Distortion and double images: example images

An example image for each evaluation class is shown in Figure 159. For simplicity, the images are shown inverted. The images labelled with 5, 4 or 3 are still classified as acceptable. In contrast, the images labelled with 2 or 1 are classified as unacceptable.

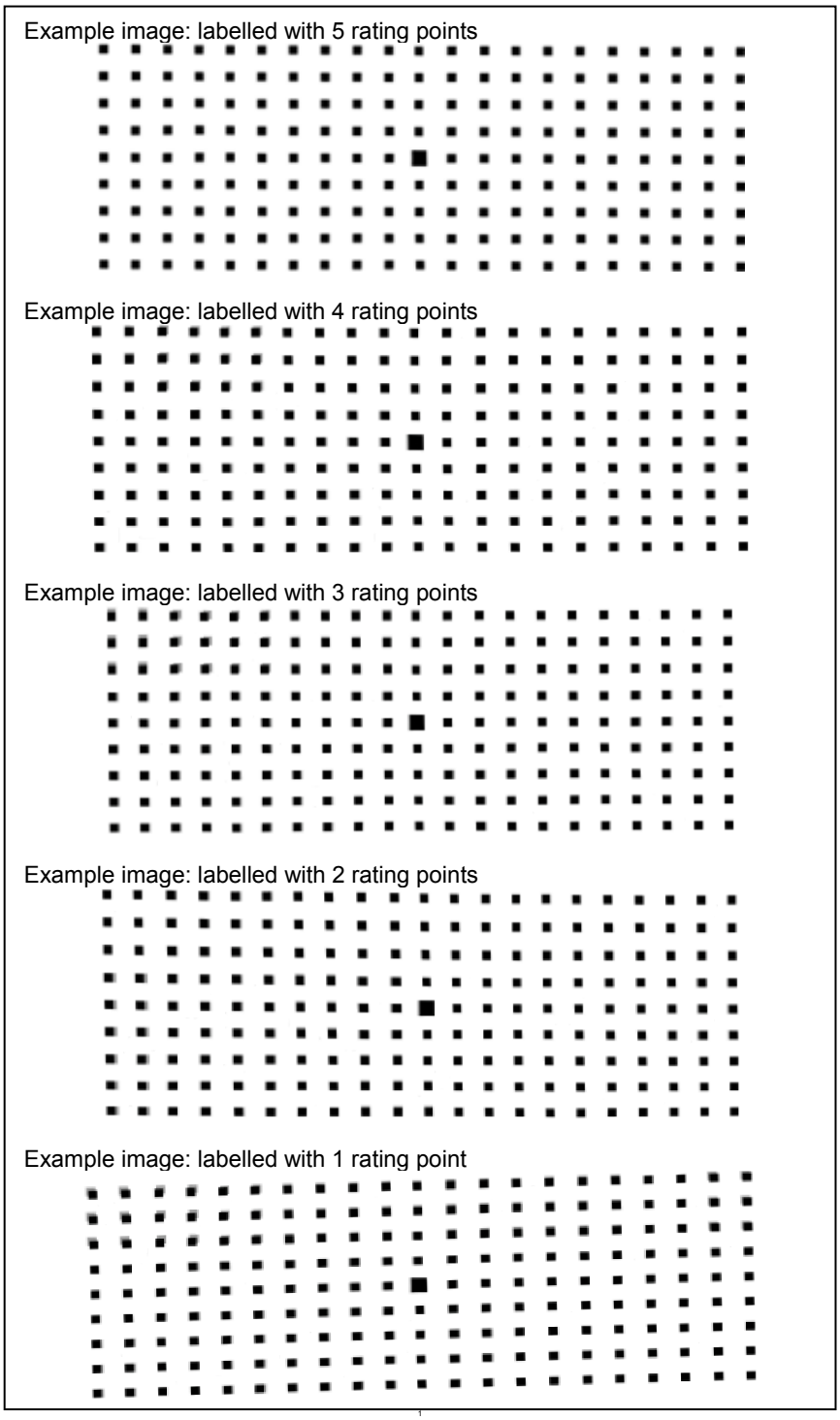


Figure 159: distortion and double images: example images of different rating classes

A.7 SSL, PC: results for the distortion dataset

In the first step, a new label is accepted and added to the training set if the maximum probability that the image belongs to the rating class is greater than the threshold θ_1. Here, θ_1 is set to 0.6, 0.7, and 0.8. The size of the initially labelled training images is set to 10%, 15%, …, 70% of all training images from each rating class, as shown in Table 32. The RMSE values and the classification accuracies after the termination of the SSL process are summarised in Figure 160. Since the selection of the labelled training data is done randomly, the average values and the standard deviations over 30 runs are calculated. The x-axes represent the number of initially labelled images in the training dataset and the y-axes the classification results.

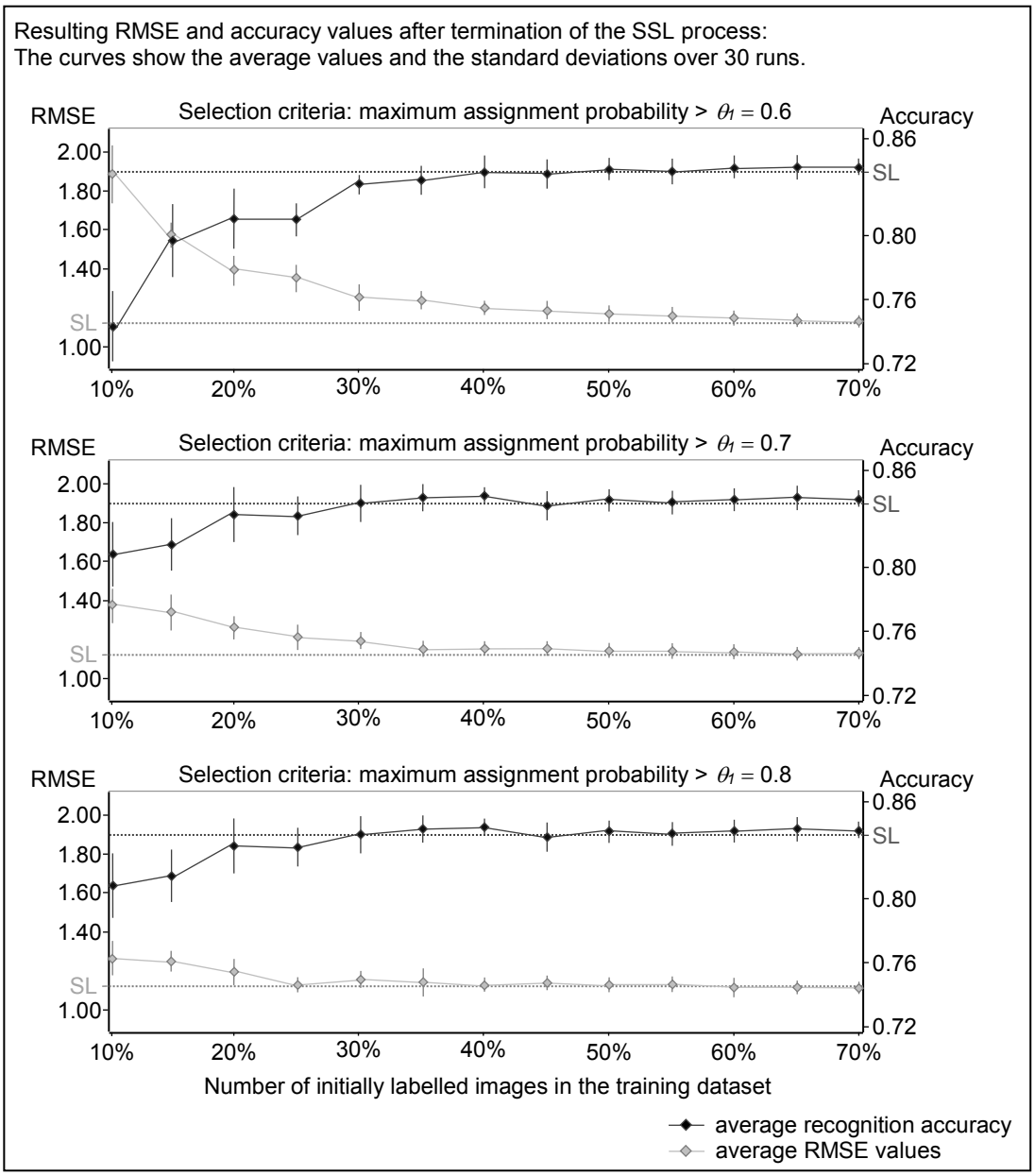


Figure 160: SSL, PC: results for the distortion dataset, θ_1, part I

The TPR and FPR values after the termination of the SSL process are summarised in Figure 161. Since the selection of the labelled training data is done randomly, the average values and the standard deviations over 30 runs are calculated. The x-axes represent the number of initially labelled images in the training dataset and the y-axes the classification results.

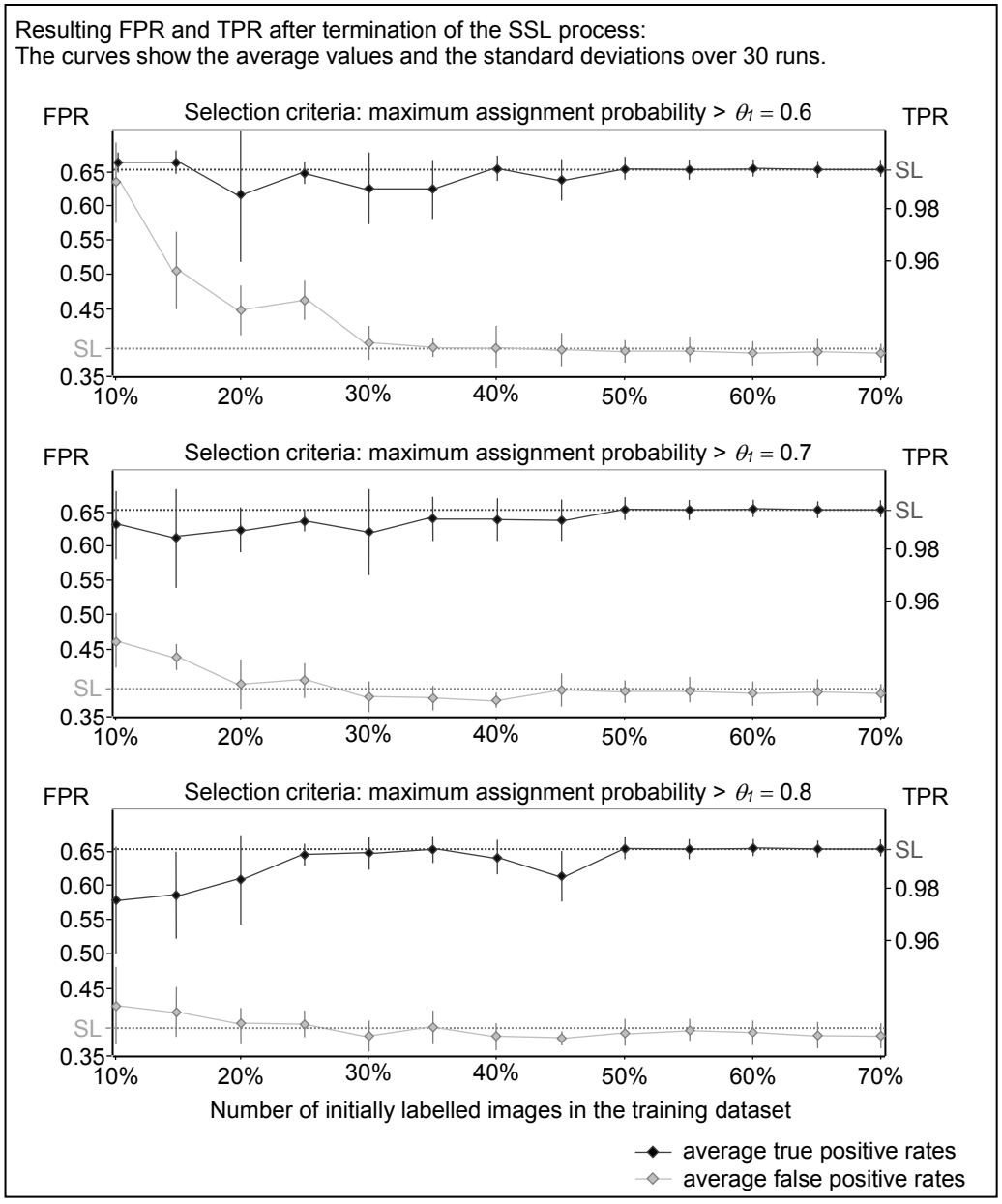


Figure 161: SSL, PC: results for the distortion dataset, θ_1, part II

In the second step, an autonomously labelled image is transferred into the training set if the difference between the largest and the second largest class probability is greater than the threshold θ_2, which is set to 0.1, 0.2, and 0.3. Again, the size of the initially labelled training images is set to 10%, 15%, …, 70% of all training images, as shown in Table 32. The RMSE values and the classification accuracies after the termination of the SSL process are summarised in Figure 162. Since the selection of the labelled training data is done randomly, the average values and the standard deviations over 30 runs are calculated. The x-axes represent the number of initially labelled images in the training dataset and the y-axes the classification results.

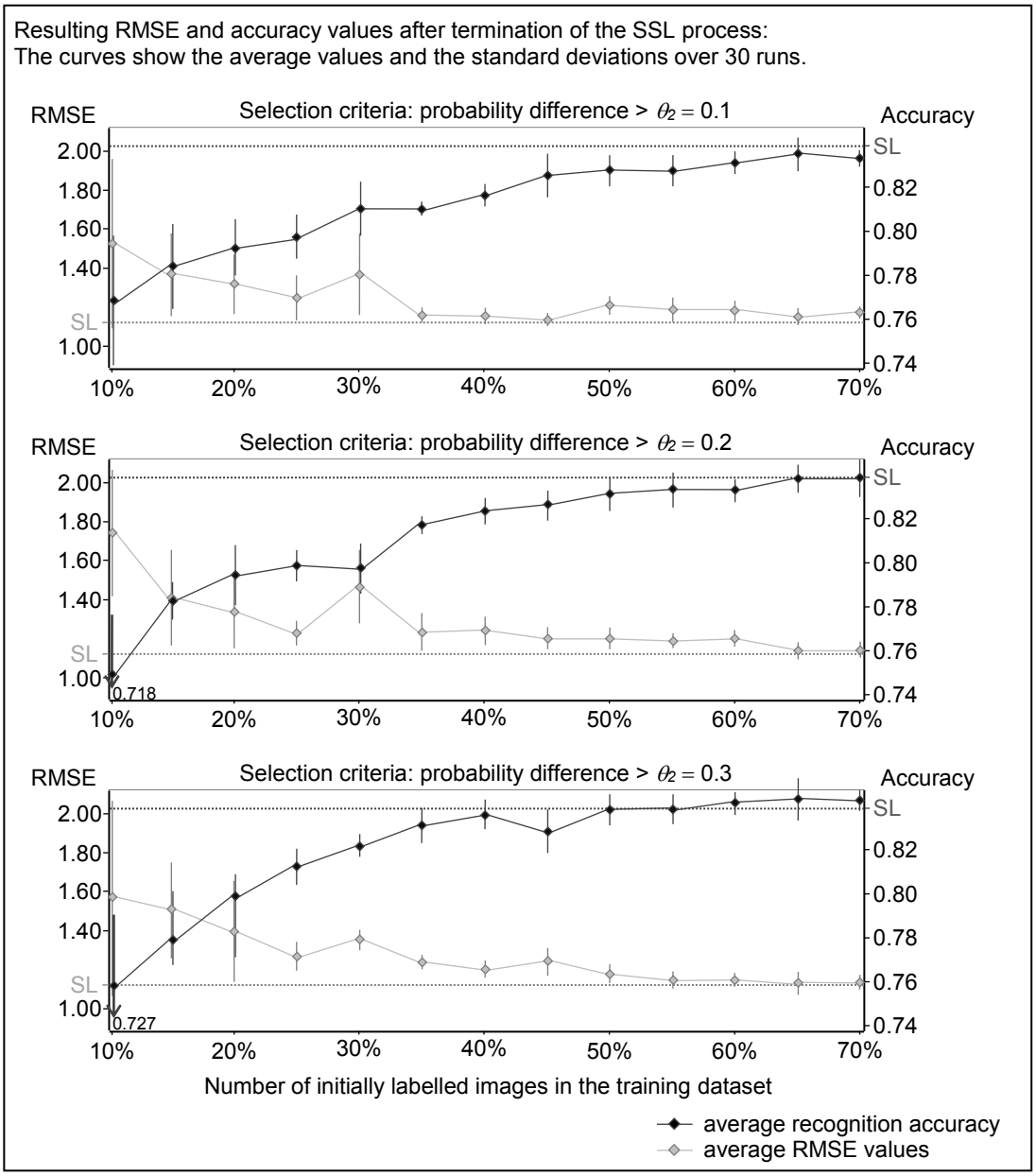


Figure 162: SSL, PC: results for the distortion dataset, θ_2, part I

The TPR and FPR values after the termination of the SSL process are summarised in Figure 163. Since the selection of the labelled training data is done randomly, the average values and the standard deviations over 30 runs are calculated. The x-axes represent the number of initially labelled images in the training dataset and the y-axes the classification results.

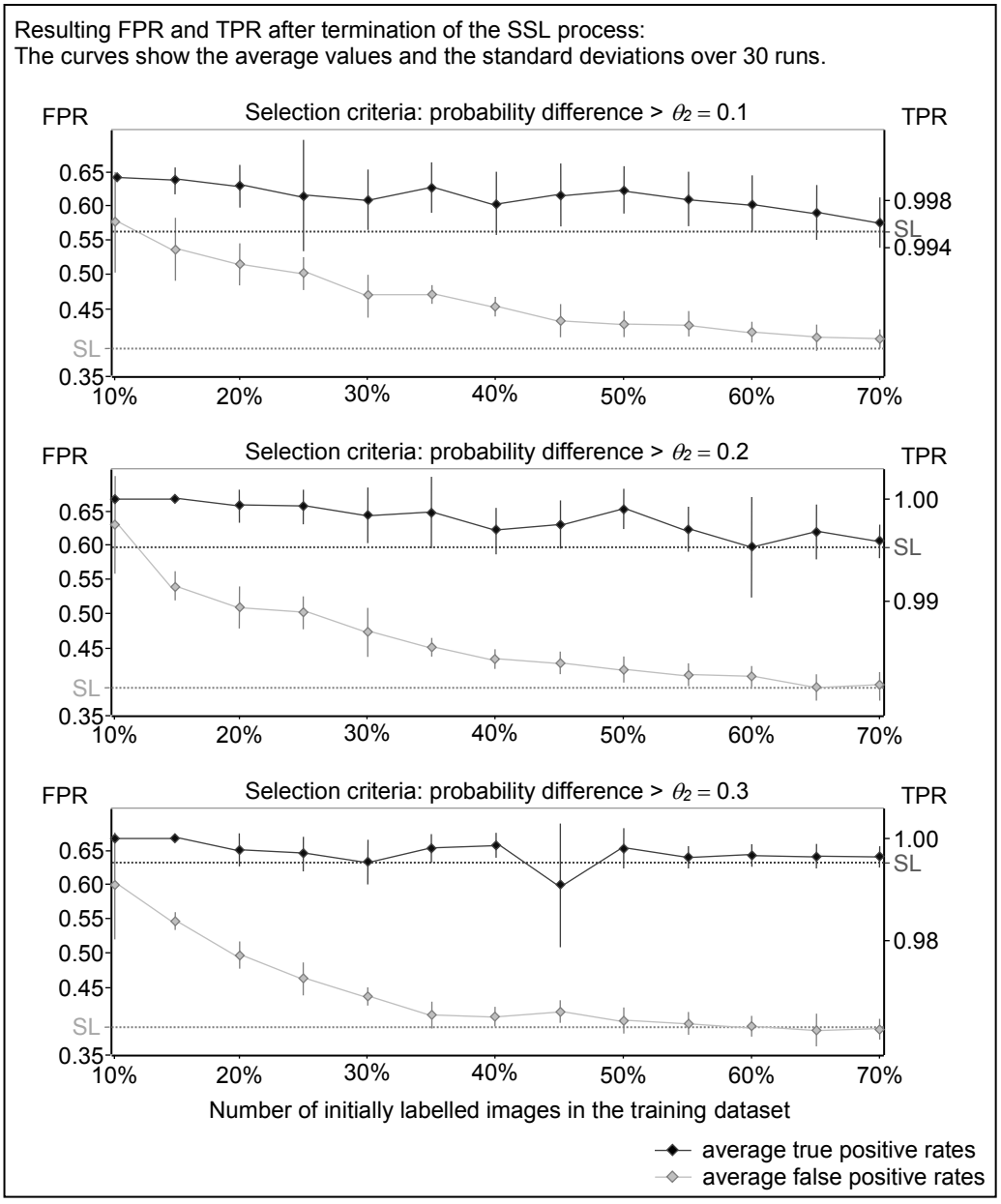


Figure 163: SSL, PC: results for the distortion dataset, θ_2, part II

In the last step, both selection criteria are combined. An autonomously labelled image is transferred into the training set if the class assignment probabilities fulfil both threshold conditions simultaneously. The maximum probability that the image belongs to the corresponding rating class has to be greater than the threshold θ_1 and the difference between the largest and the second largest class probability must be greater than the threshold θ_2. The RMSE values and the classification accuracies after the termination of the SSL process are summarised in Figure 164. Since the selection of the labelled training data is done randomly, the average values and the standard deviations over 30 runs are calculated. The x-axes represent the number of initially labelled images in the training dataset and the y-axes the classification results.

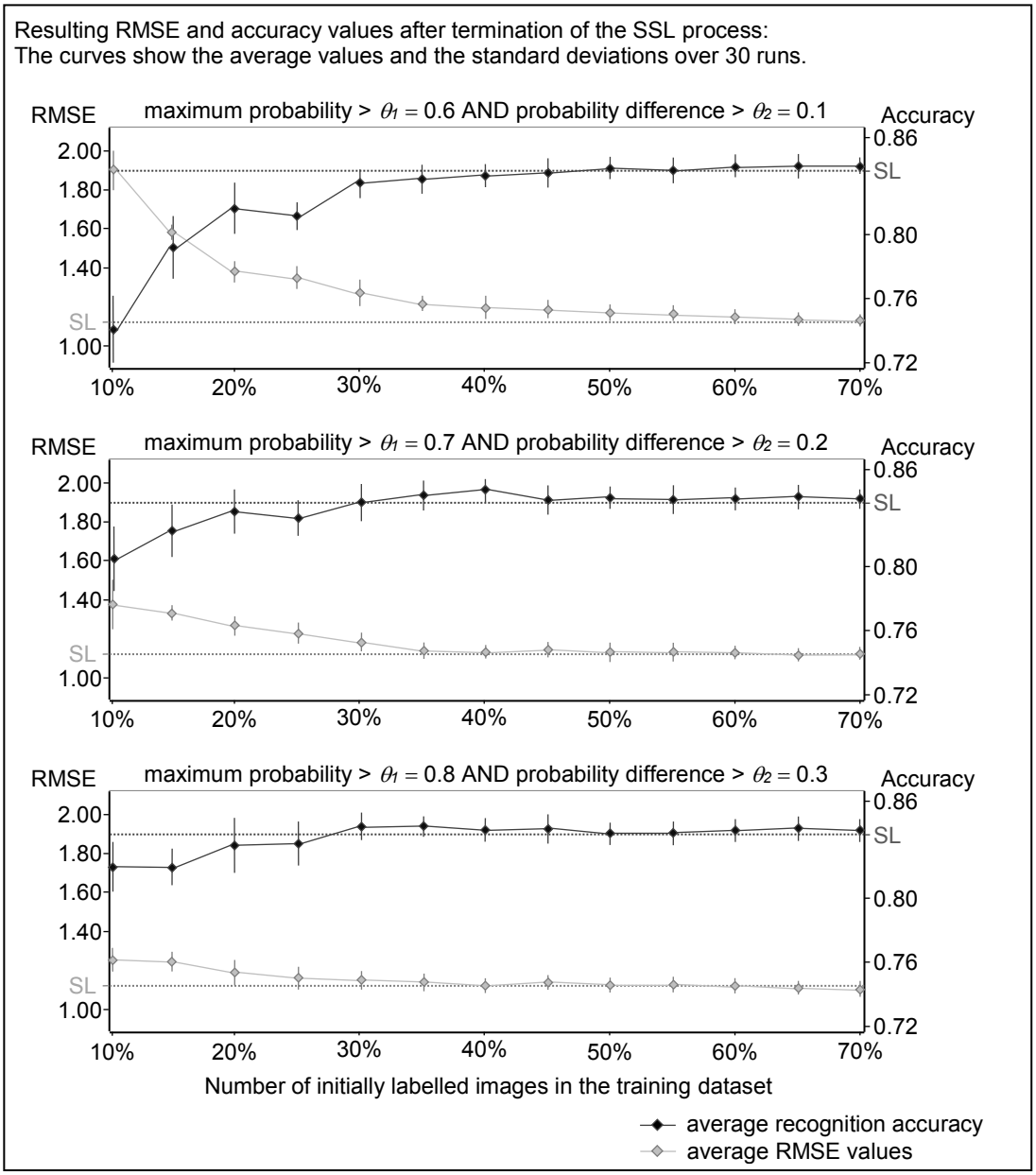


Figure 164: SSL, PC: results for the distortion dataset, θ_1 AND θ_2, part I

The TPR and FPR values after the termination of the SSL process are summarised in Figure 165. Since the selection of the labelled training data is done randomly, the average values and the standard deviations over 30 runs are calculated. The x-axes represent the number of initially labelled images in the training dataset and the y-axes the classification results.

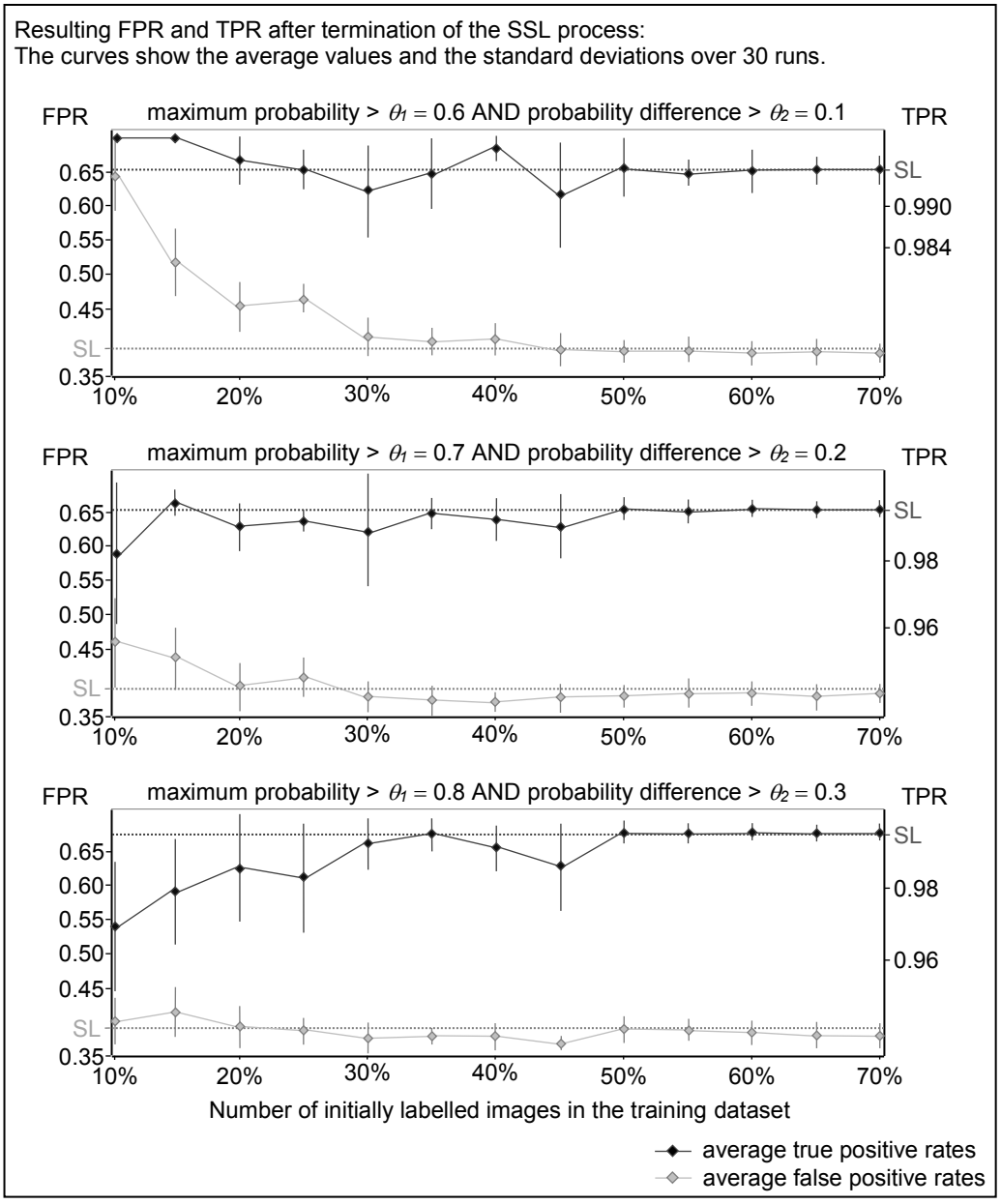


Figure 165: SSL, PC: results for the distortion dataset, θ_1 AND θ_2, part II

A.8 SSL, PC: learning curves for the distortion dataset

The learning curves for the distortion dataset for 40% manually labelled images are shown in Figure 166. A new label is accepted and added to the training set if the maximum probability that the image belongs to the corresponding rating class is greater than the threshold θ_1, which is exemplarily set to 0.7.

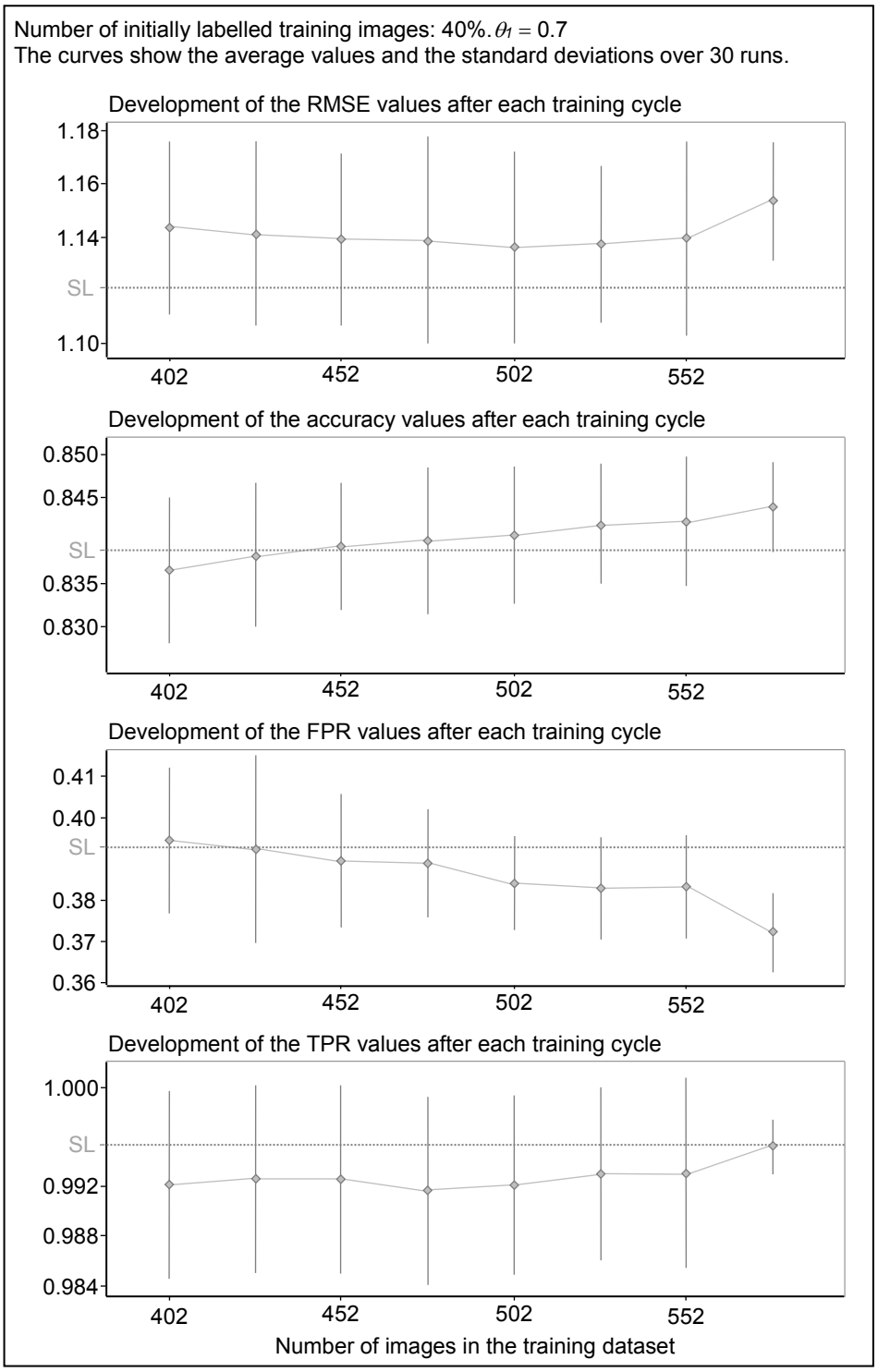


Figure 166: SSL, PC: learning curves for the distortion dataset, θ_1

In the second step, if the difference between the largest and the second largest class probability is greater than the threshold θ_2, a new label is accepted and added to the training set. Here, θ_2 is exemplarily set to 0.3. The corresponding learning curves for 40% manually labelled images are shown in Figure 167.

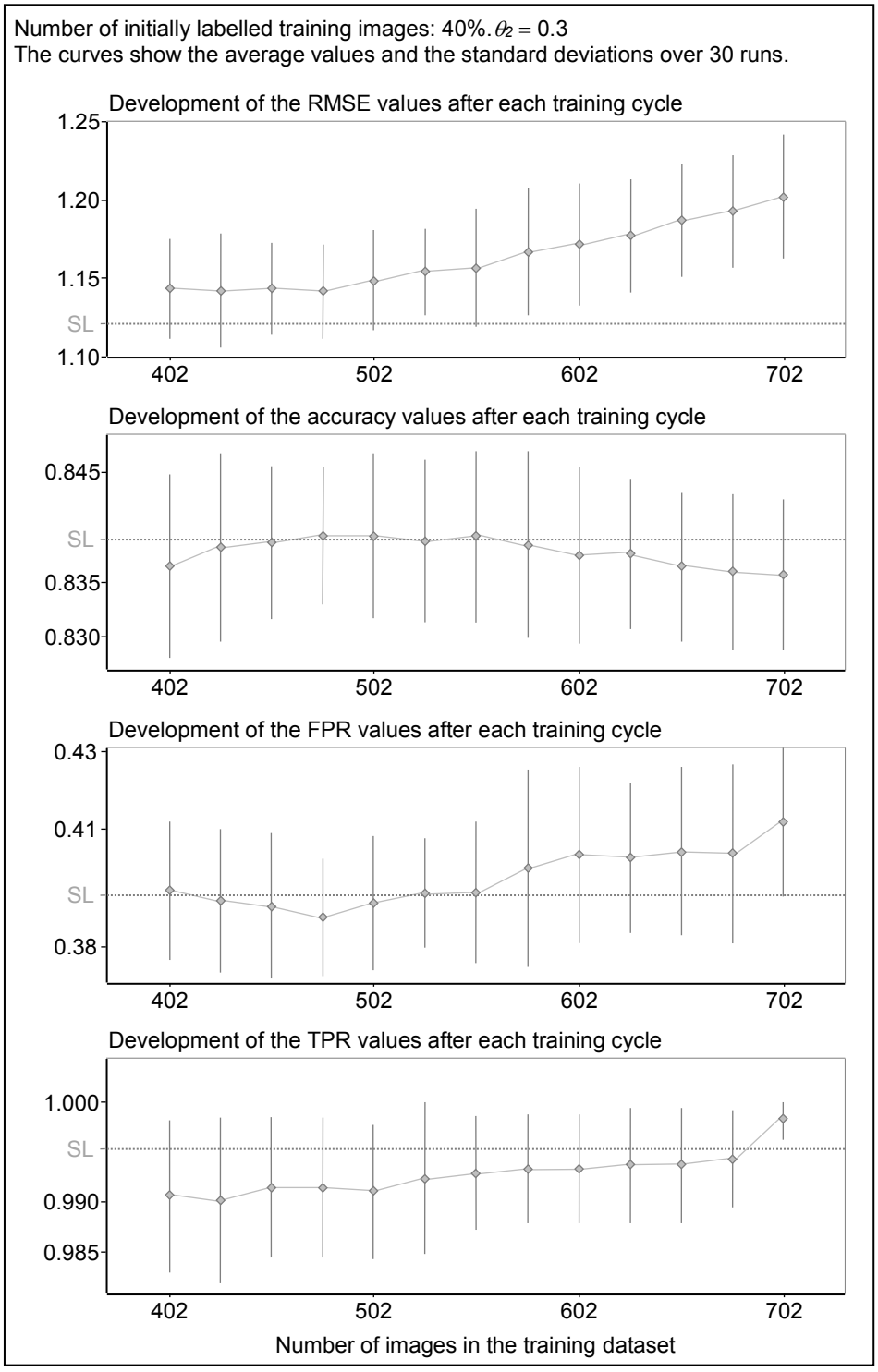


Figure 167: SSL, PC: learning curves for the distortion dataset, θ_2

Finally, a new label is accepted and added to the training set if the maximum probability that the image belongs to the corresponding rating class is greater than the threshold θ_1 and if the difference between the largest and the second largest class probability is greater than the threshold θ_2. Exemplarily, the resulting learning curves for 40% manually labelled images and $\theta_1 = 0.7$ and $\theta_2 = 0.2$ are shown in Figure 168.

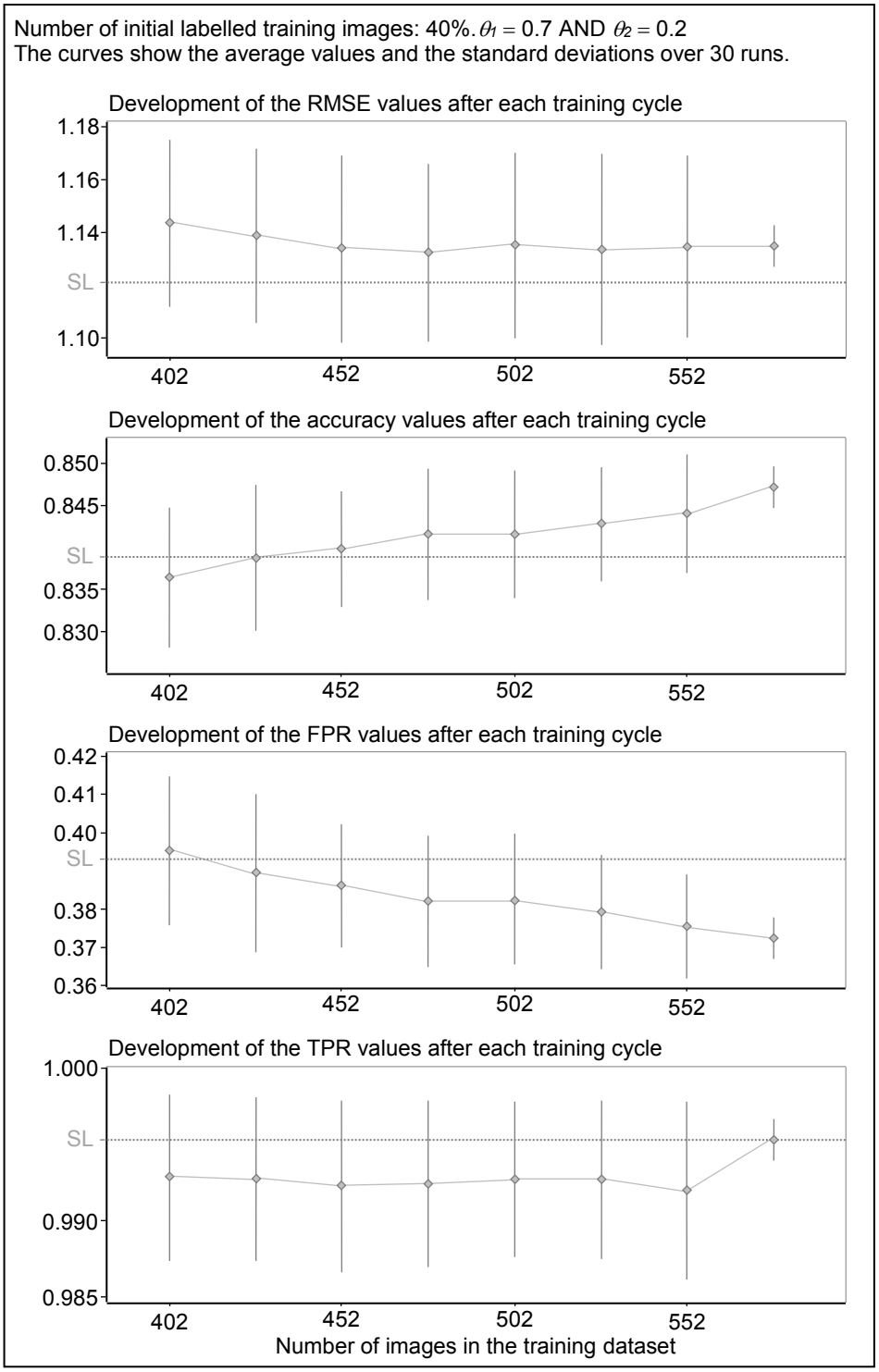


Figure 168: SSL, PC: learning curves for the distortion dataset, θ_1 AND θ_2

A.9 SSL, kNN: results for the distortion dataset

During the SSL process, an autonomously labelled image is transferred into the training dataset if the distance to the nearest reference pattern is smaller than a given threshold distance θ. Successively, the minimal, mean, and maximal determined distances, as shown in Table 33, are used as thresholds during the SSL process. The size of the initially labelled training set is set to 10%, 15%, …, 70% of all training images, as shown in Table 32. Due to the random selection of the initially labelled training images, the mean and the standard deviation over 30 runs are calculated. The RMSE values and the classification accuracies after the termination of the SSL process are summarised in Figure 169. The x-axes represent the number of initially labelled images in the training dataset and the y-axes the classification results.

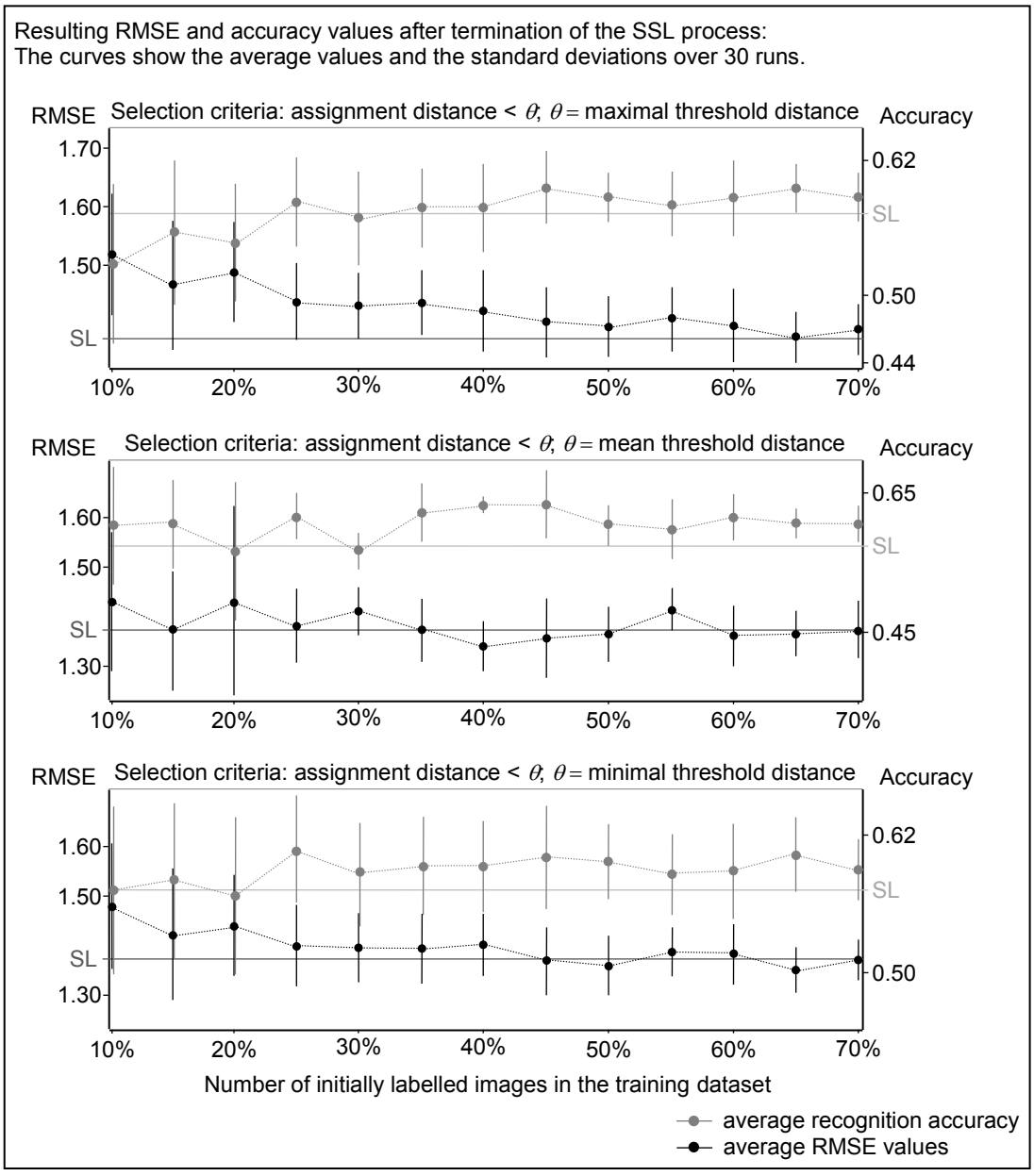


Figure 169: SSL, kNN: results for the distortion dataset, part I

The TPR and FPR values after the termination of the SSL process are summarised in Figure 170. Since the selection of the labelled training data is done randomly, the average values and the standard deviations over 30 runs are calculated. The x-axes represent the number of initially labelled images in the training dataset and the y-axes the classification results.

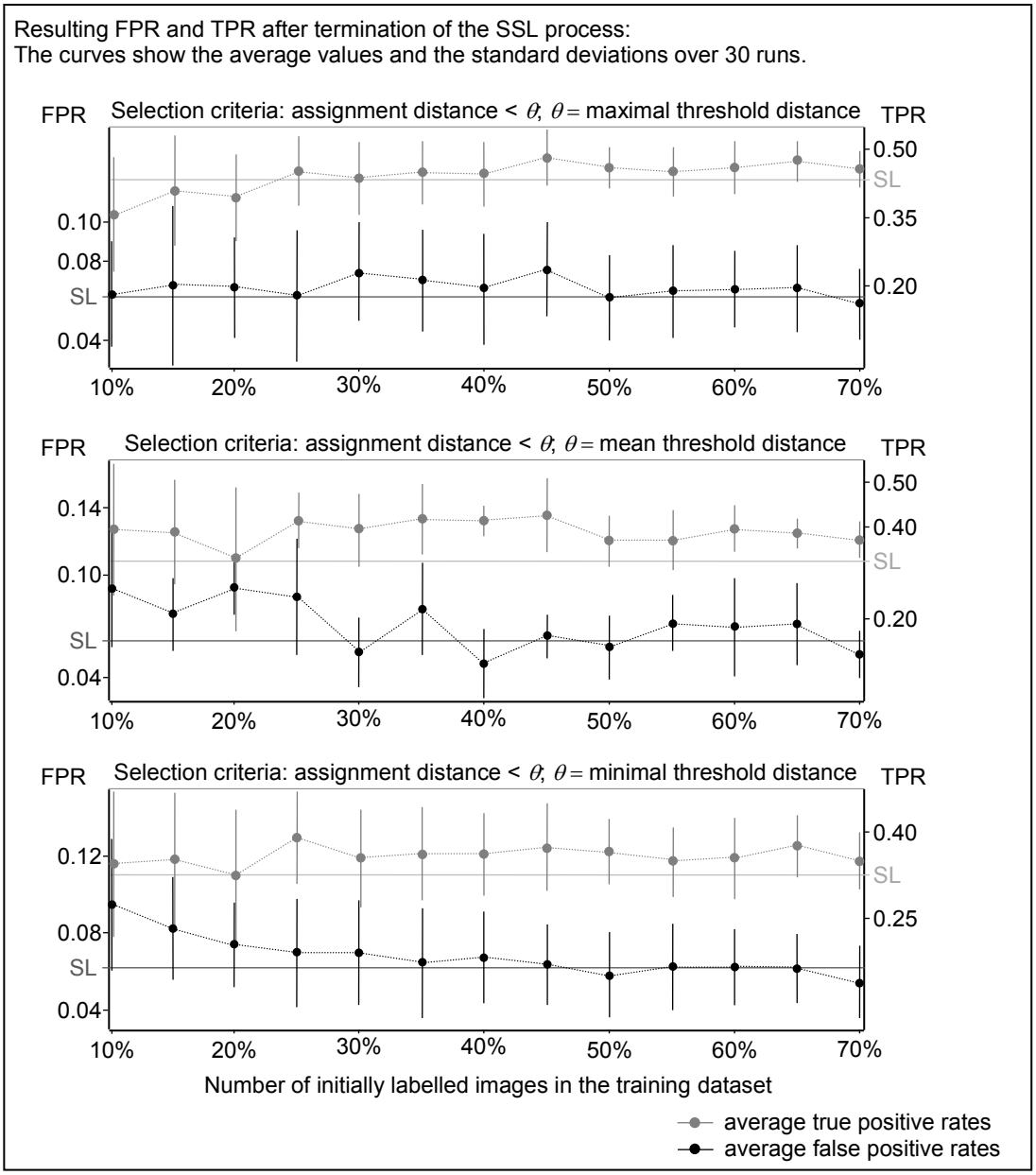


Figure 170: SSL, kNN: results for the distortion dataset, part II

A.10 SSL, kNN: learning curves for the distortion dataset

The learning curves for the distortion dataset for 45% manually labelled images are shown in Figure 171. A new label is accepted and added to the training set if the distance to the next training sample is smaller than the threshold distance.

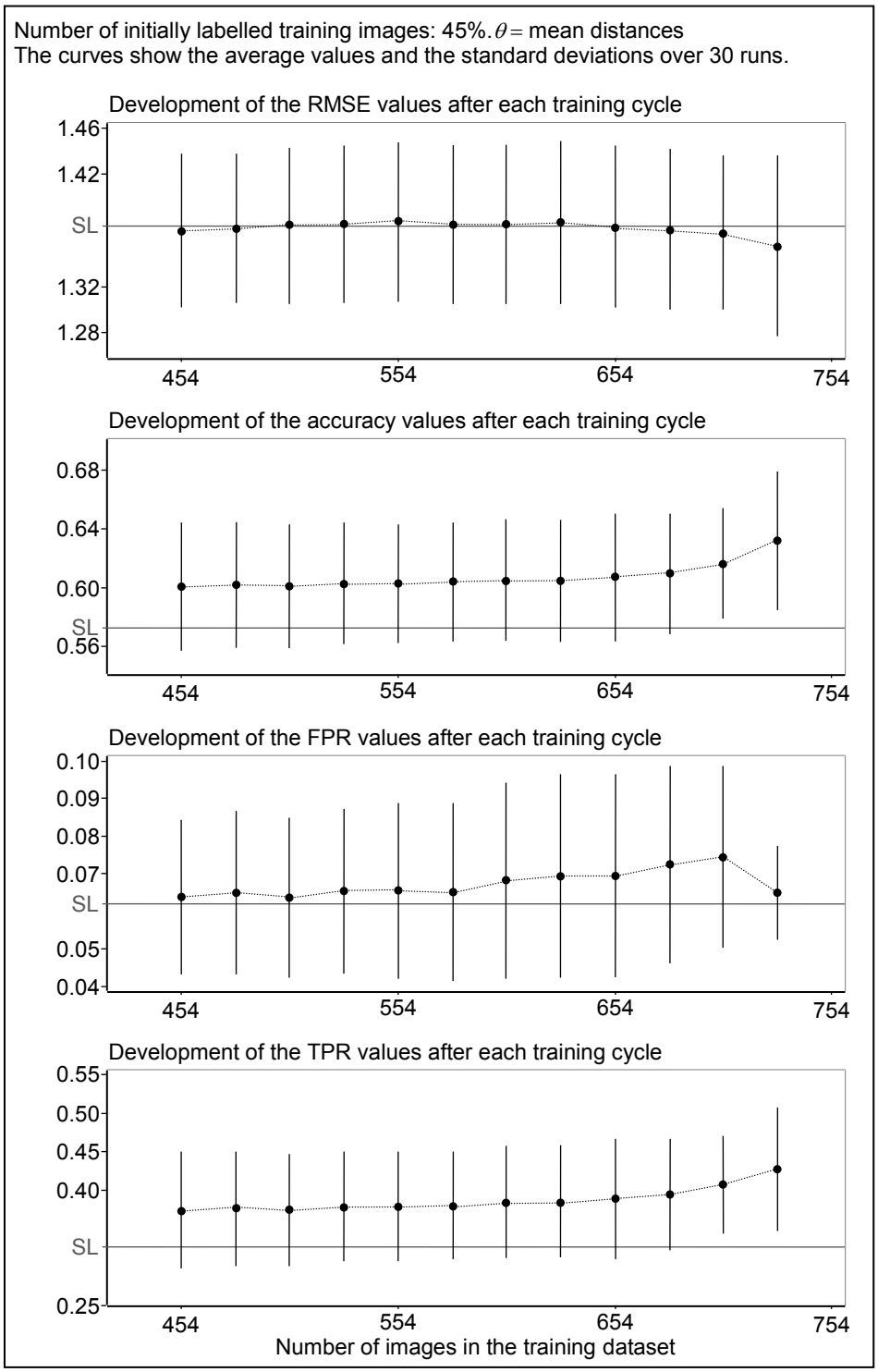


Figure 171: SSL, kNN: learning curves for the distortion dataset

A.11 SSL, LVQ: results for the distortion dataset

During the SSL process, an autonomously labelled image is transferred into the training dataset if the distance to the nearest prototype is smaller than a given threshold distance θ. Successively, the minimal, mean, and maximal determined distances, as shown in Table 34, are used as thresholds during the SSL process. The size of the initially labelled training set is set to 10%, 15%, …, 70% of all training images, as shown in Table 32. Due to the random selection of the initially labelled training images, the mean and the standard deviation over 30 runs are calculated. The RMSE values and the classification accuracies after the termination of the SSL process are summarised in Figure 172. The x-axes represent the number of initially labelled images in the training dataset and the y-axes the classification results.

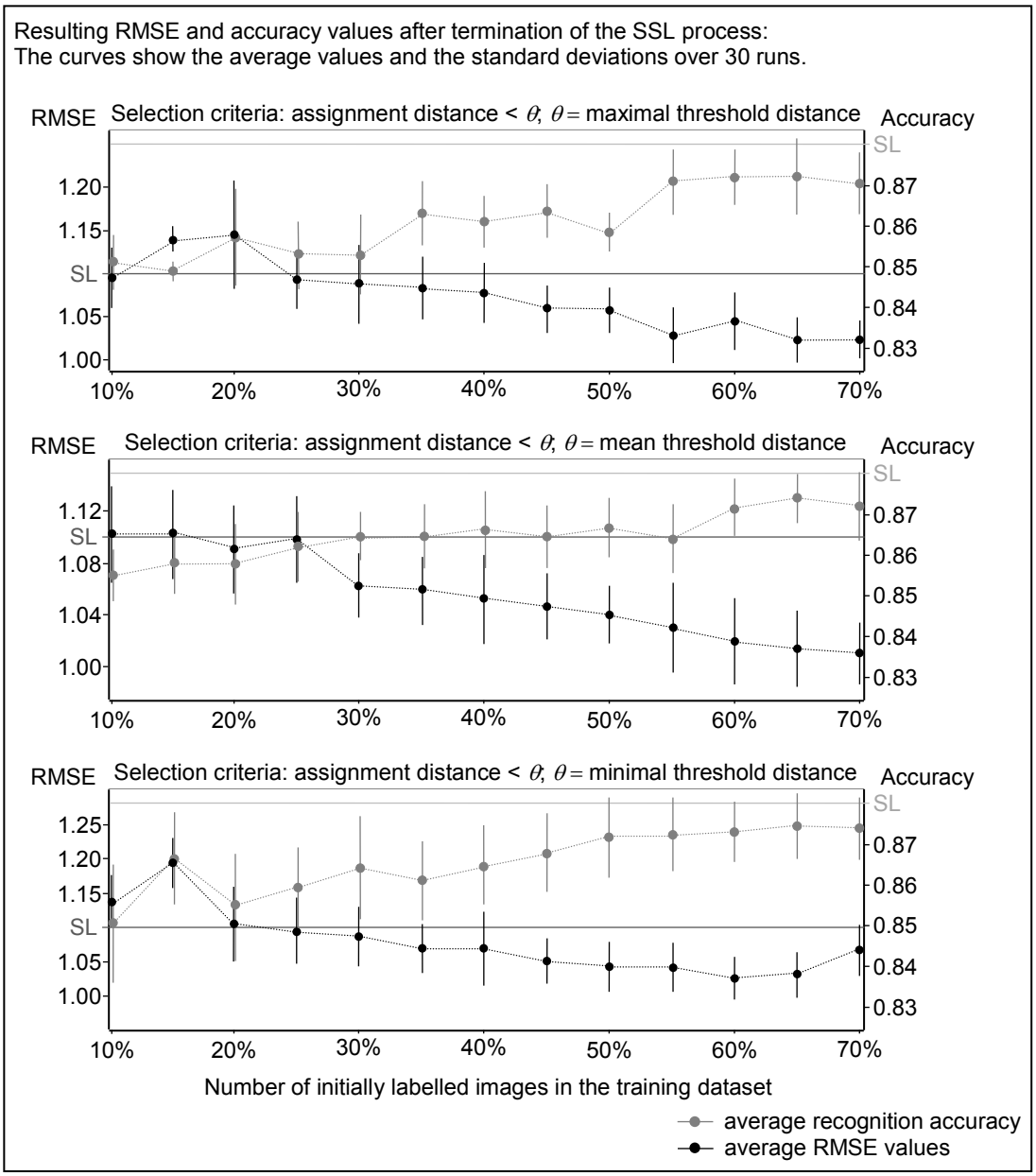


Figure 172: SSL, LVQ: results for the distortion dataset, part I

The TPR and FPR values after the termination of the SSL process are summarised in Figure 173. Since the selection of the labelled training data is done randomly, the average values and the standard deviations over 30 runs are calculated. The x-axes represent the number of initially labelled images in the training dataset and the y-axes the classification results.

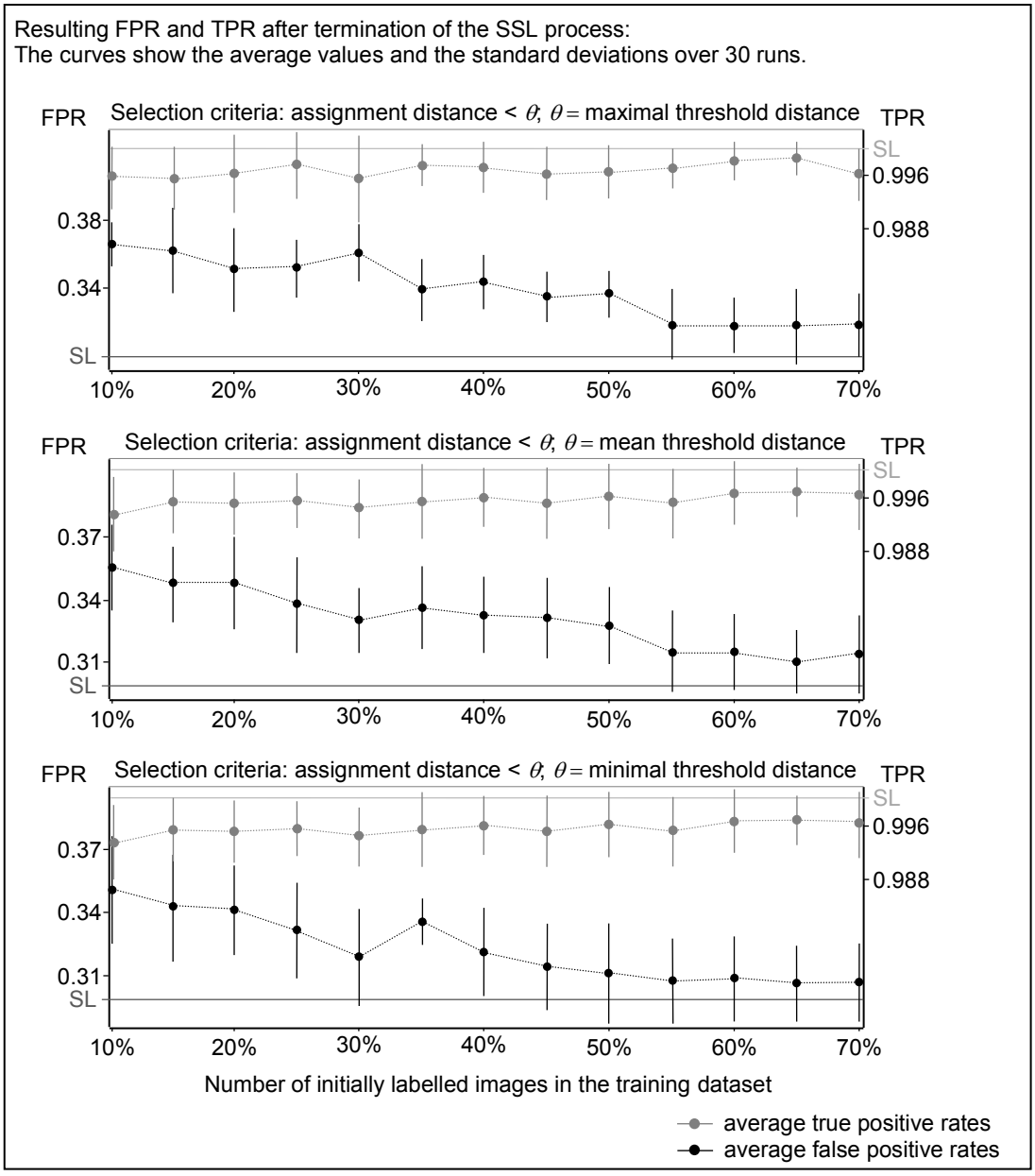


Figure 173: SSL, LVQ: results for the distortion dataset, part II

A.12 SSL, LVQ: learning curves for the distortion dataset

The learning curves for the distortion dataset for 55% manually labelled images are exemplarily shown in Figure 174. A new label is accepted and added to the training set if the distance to the next training sample is smaller than the threshold distance.

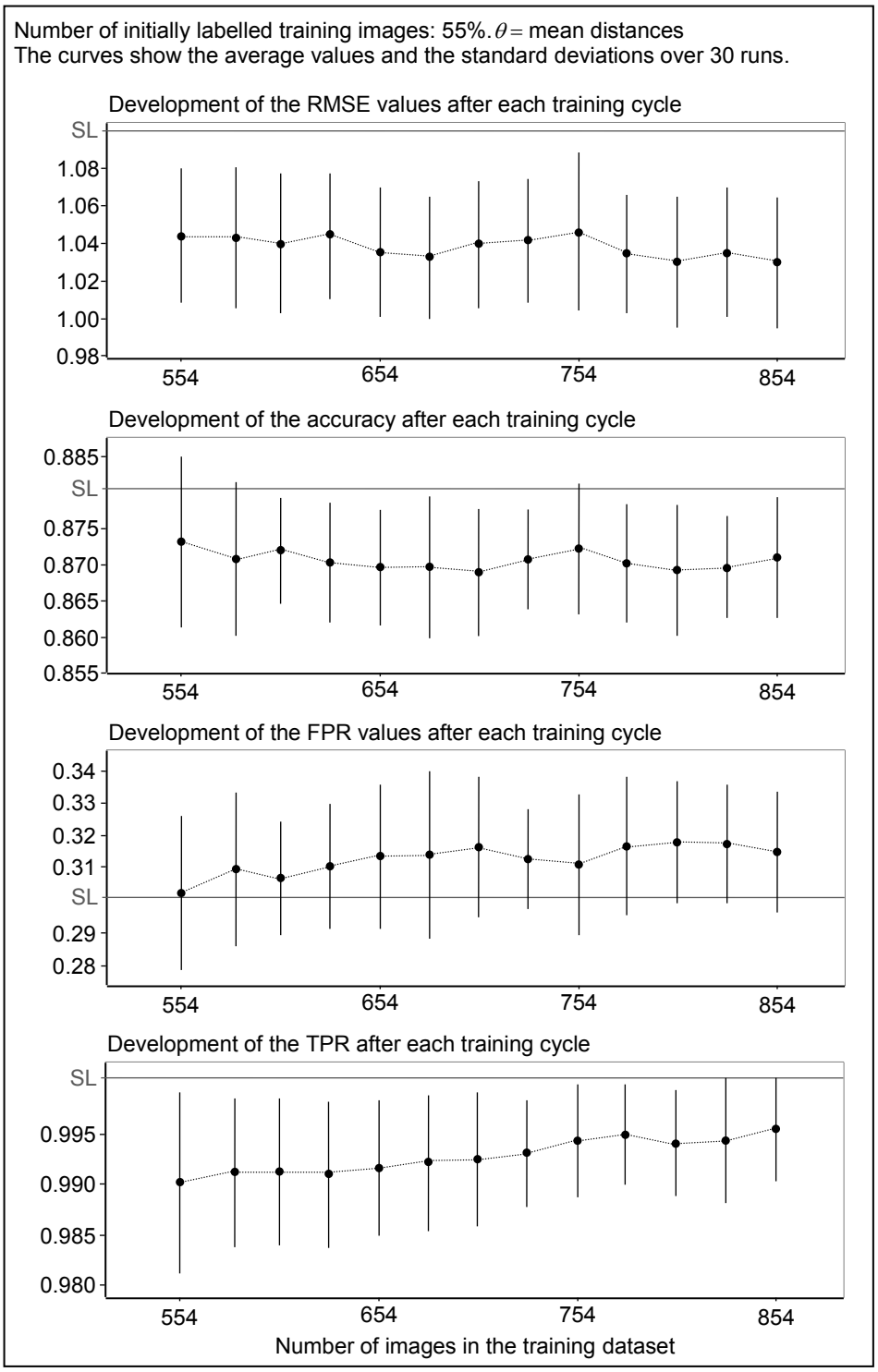


Figure 174: SSL, LVQ: learning curves for the distortion dataset

A.13 SSL, PC: results for the double image dataset

In the first step, a new label is accepted and added to the training set if the maximum probability that the image belongs to the rating class is greater than the threshold θ_1. Here, θ_1 is set to 0.6, 0.7, and 0.8. The size of the initially labelled training images is set to 10%, 15%, …, 70% of all training images from each rating class, as shown in Table 35. The RMSE values and the classification accuracies after the termination of the SSL process are summarised in Figure 175. Since the selection of the labelled training data is done randomly, the average values and the standard deviations over 30 runs are calculated. The x-axes represent the number of initially labelled images in the training dataset and the y-axes the classification results.

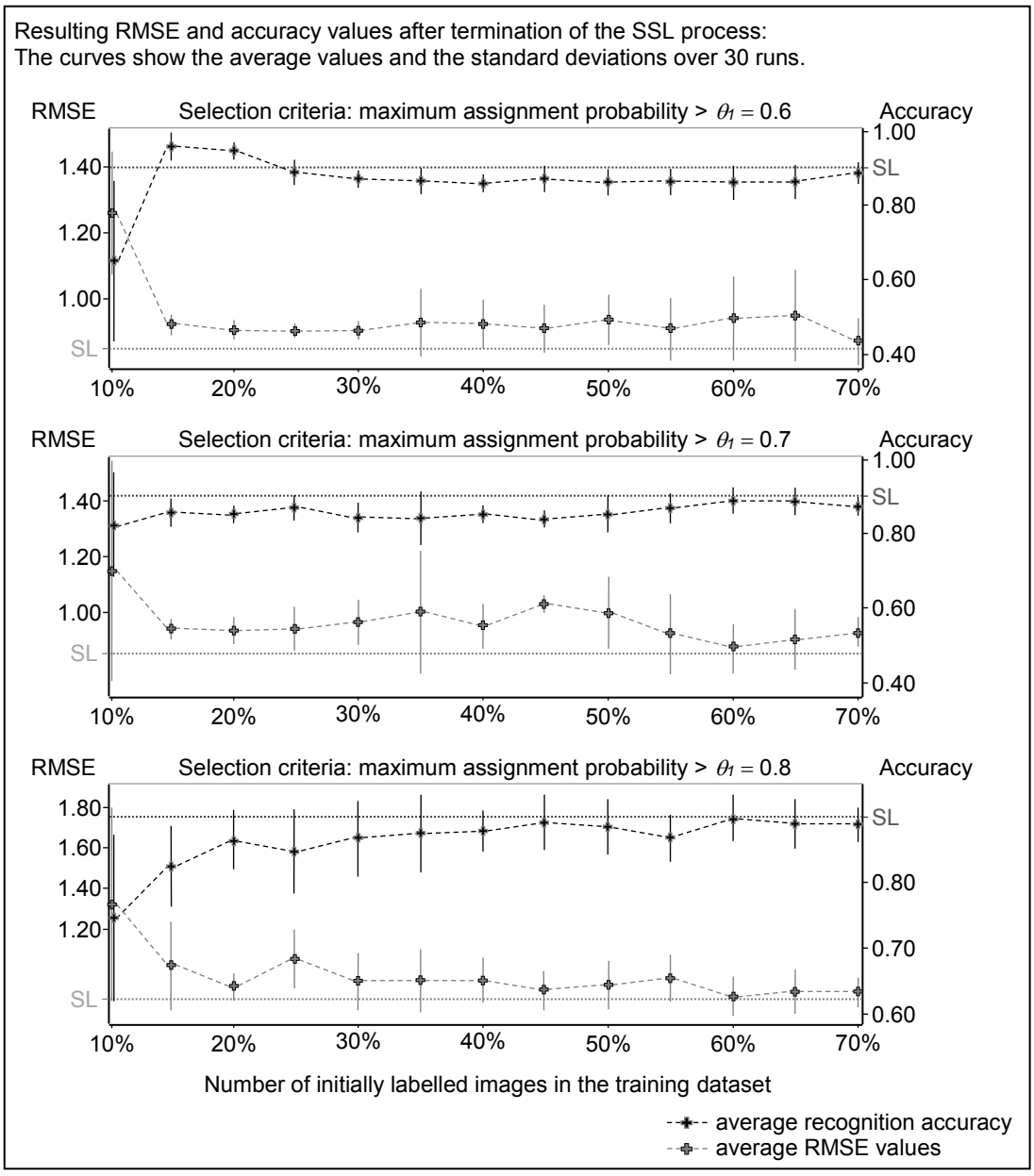


Figure 175: SSL, PC: results for the double image dataset, θ_1, part I

The TPR and FPR values after the termination of the SSL process are summarised in Figure 176. Since the selection of the labelled training data is done randomly, the average values and the standard deviations over 30 runs are calculated. The x-axes represent the number of initially labelled images in the training dataset and the y-axes the classification results.

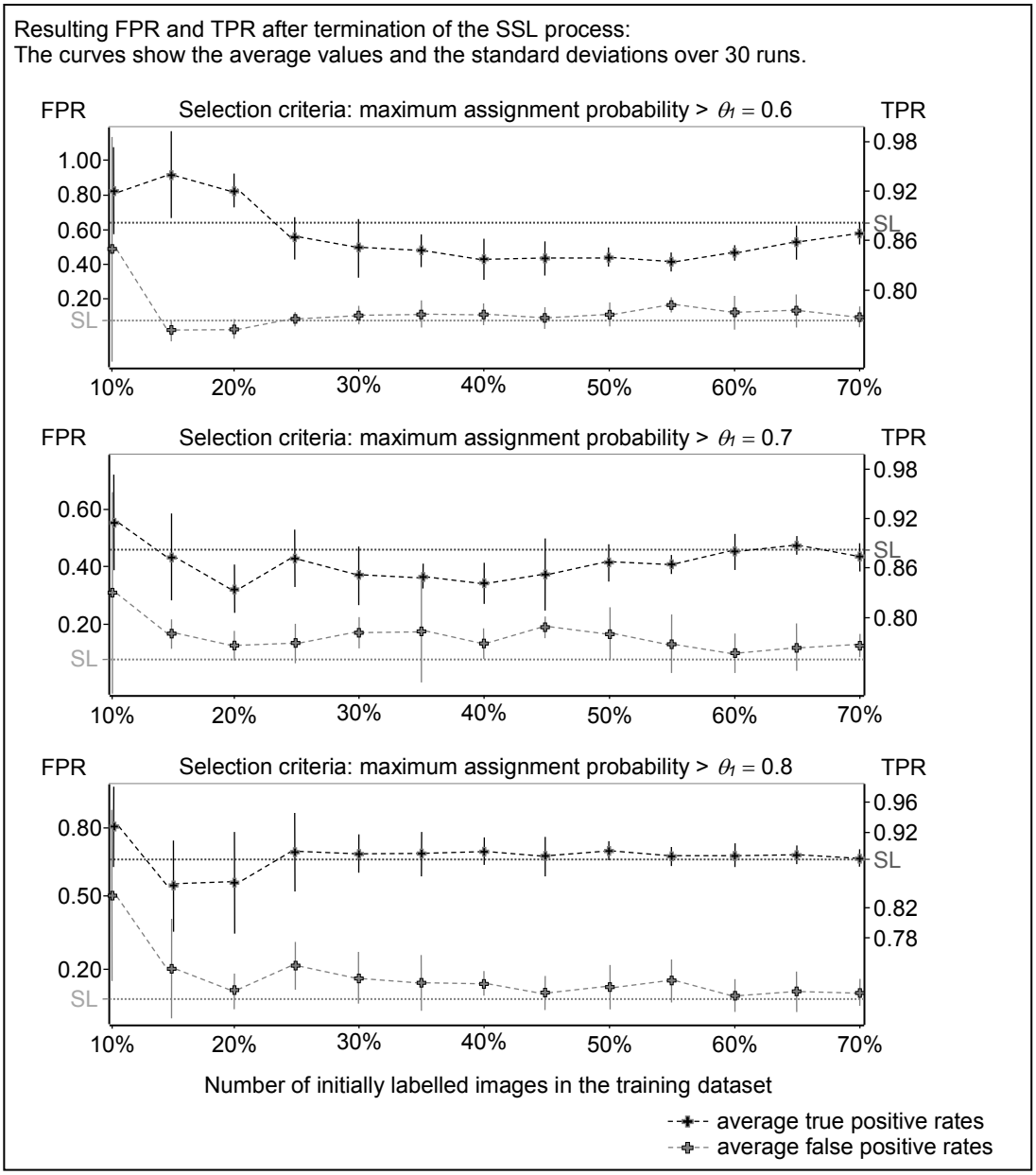


Figure 176: SSL, PC: results for the double image dataset, θ_1, part II

In the second step, an autonomously labelled image is transferred into the training set if the difference between the largest and the second largest class probability is greater than the threshold θ_2, which is set to 0.1, 0.2, and 0.3. Again, the size of the initially labelled training images is set to 10%, 15%, …, 70% of all training images, as shown in Table 35. The RMSE values and the classification accuracies after the termination of the SSL process are summarised in Figure 177. Since the selection of the labelled training data is done randomly, the average values and the standard deviations over 30 runs are calculated. The x-axes represent the number of initially labelled images in the training dataset and the y-axes the classification results.

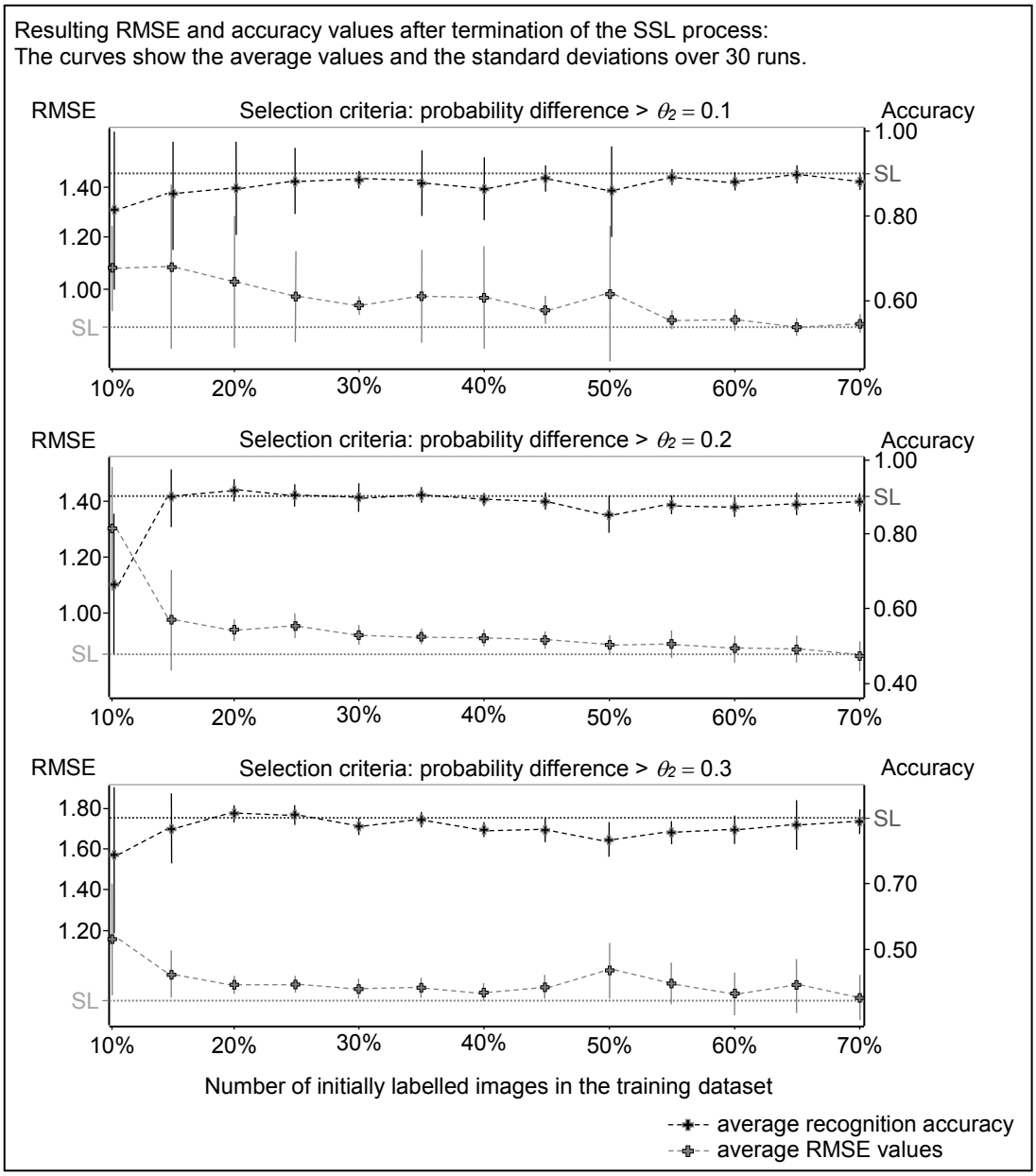


Figure 177: SSL, PC: results for the double image dataset, θ_2, part I

The TPR and FPR values after the termination of the SSL process are summarised in Figure 178. Since the selection of the labelled training data is done randomly, the average values and the standard deviations over 30 runs are calculated. The x-axes represent the number of initially labelled images in the training dataset and the y-axes the classification results.

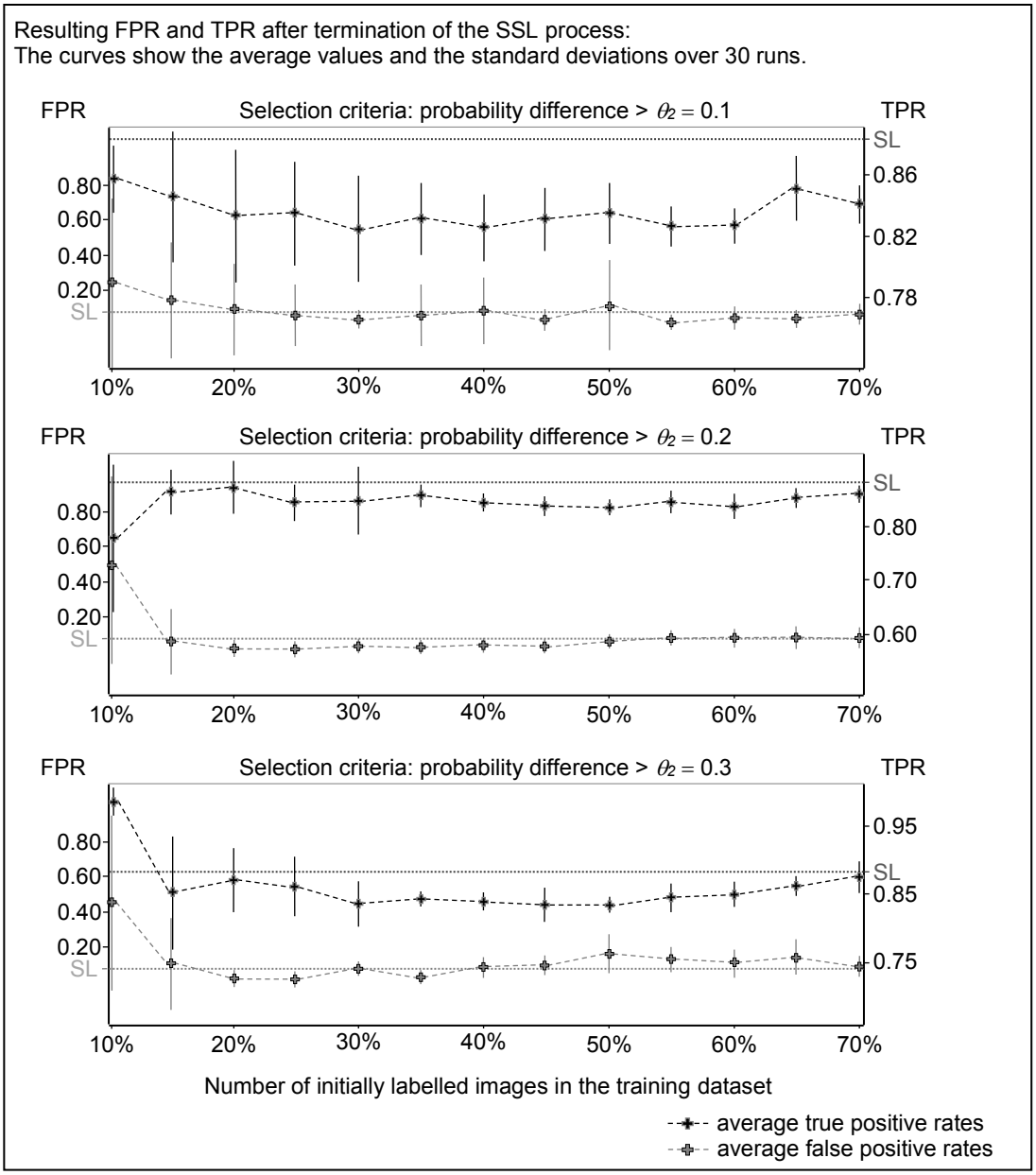


Figure 178: SSL, PC: results for the double image dataset, θ_2, part II

In the last step, both selection criteria are combined. An autonomously labelled image is transferred into the training set if the class assignment probabilities fulfil both threshold conditions simultaneously. The maximum probability that the image belongs to the corresponding rating class has to be greater than the threshold θ_1 and the difference between the largest and the second largest class probability must be greater than the threshold θ_2. The RMSE values and the classification accuracies after the termination of the SSL process are summarised in Figure 179. Since the selection of the labelled training data is done randomly, the average values and the standard deviations over 30 runs are calculated. The x-axes represent the number of initially labelled images in the training dataset and the y-axes the classification results.

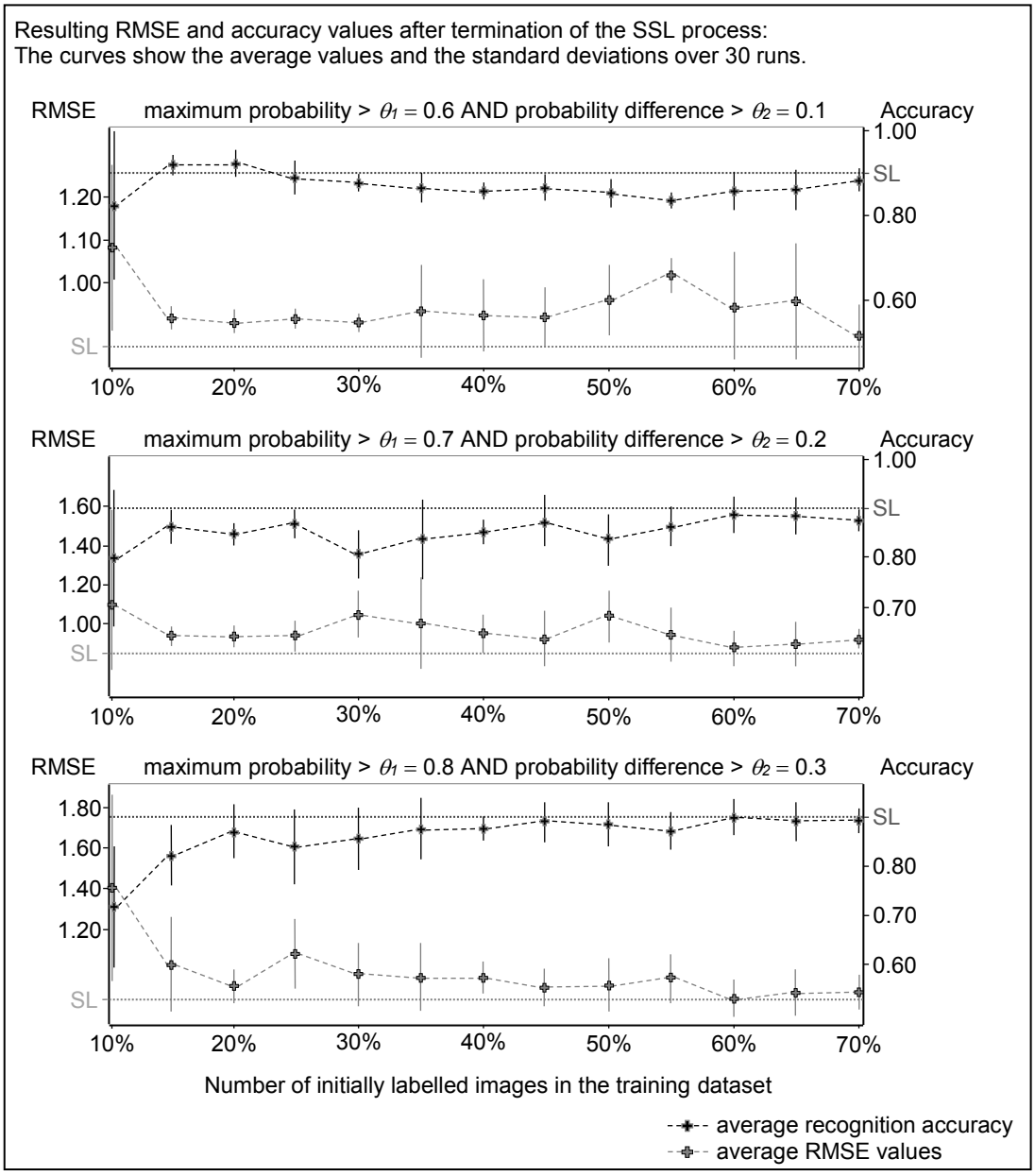


Figure 179: SSL, PC: results for the double image dataset, θ_1 AND θ_2, part I

The TPR and FPR values after the termination of the SSL process are summarised in Figure 180. Since the selection of the labelled training data is done randomly, the average values and the standard deviations over 30 runs are calculated. The x-axes represent the number of initially labelled images in the training dataset and the y-axes the classification results.

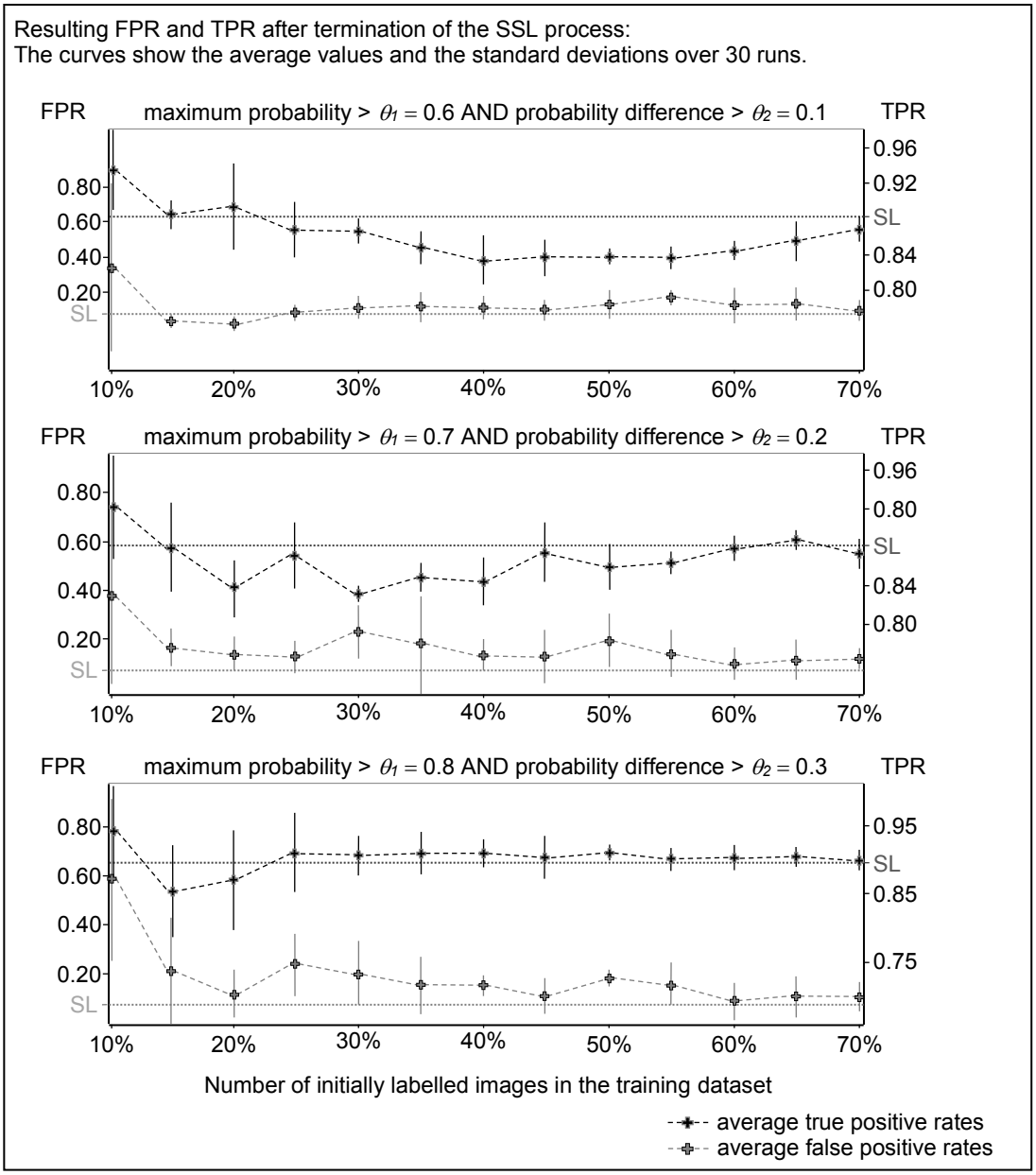


Figure 180: SSL, PC: results for the double image dataset, θ_1 AND θ_2, part II

A.14 SSL, PC: learning curves for the double image dataset

The learning curves for the distortion dataset for 15% manually labelled images are shown in Figure 181. A new label is accepted and added to the training set if the maximum probability that the image belongs to the corresponding rating class is greater than the threshold θ_1, which is exemplarily set to 0.6.

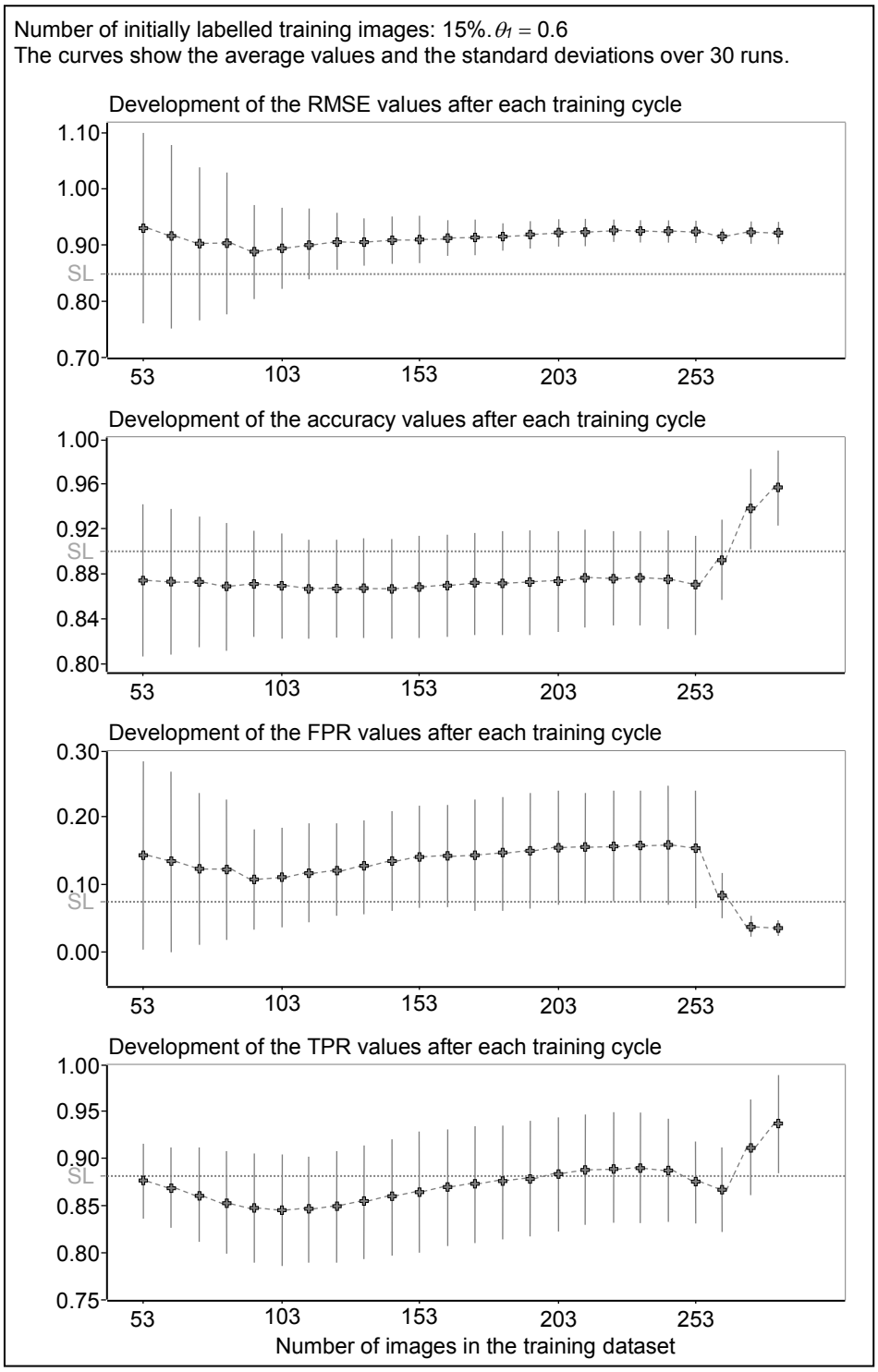


Figure 181: SSL, PC: learning curves for the double image dataset, θ_1

In the second step, if the difference between the largest and the second largest class probability is greater than the threshold θ_2, a new label is accepted and added to the training set. Here, θ_2 is exemplarily set to 0.2. The corresponding learning curves for 20% manually labelled images are shown in Figure 182.

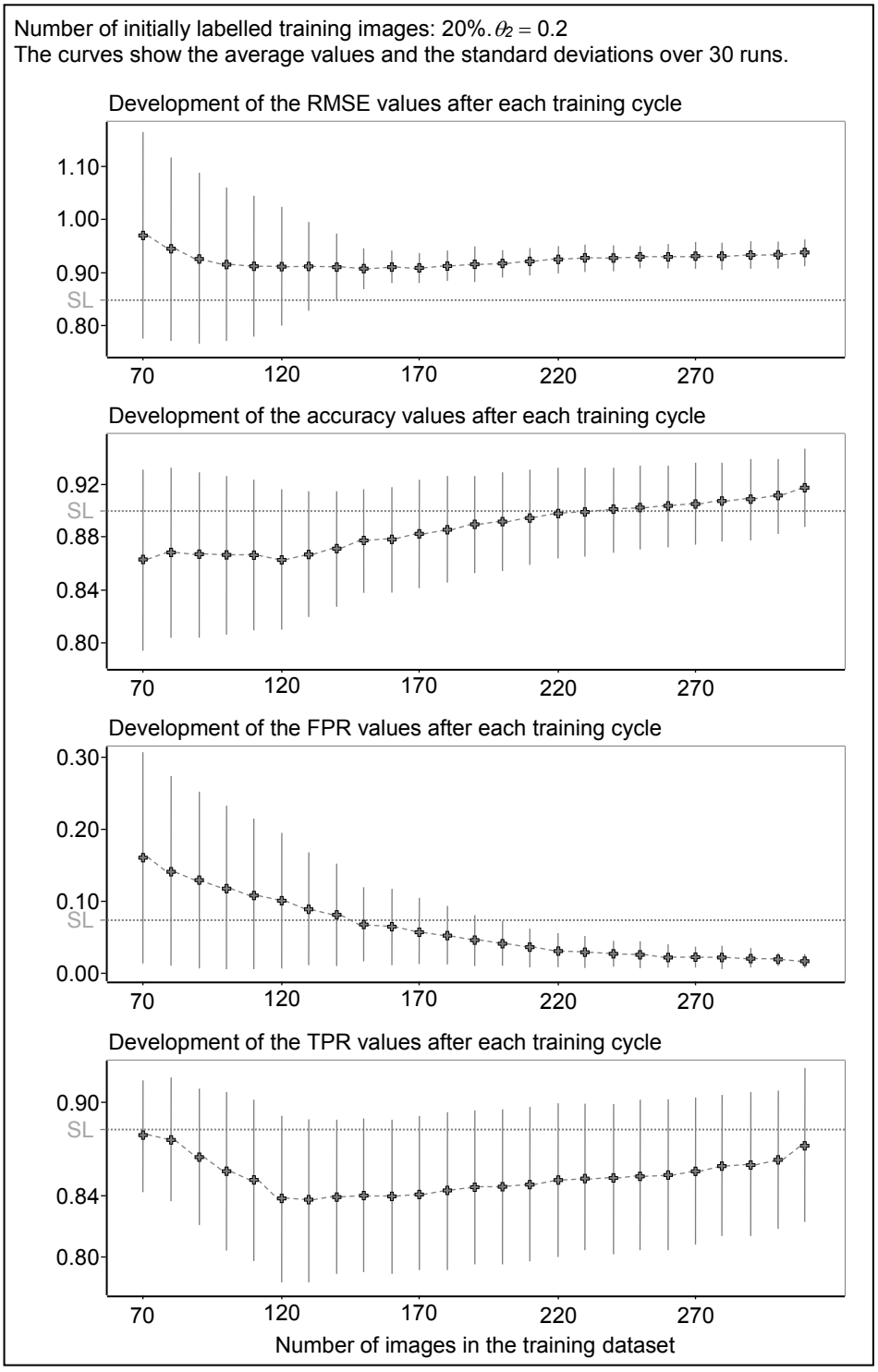


Figure 182: SSL, PC: learning curves for the double image dataset, θ_2

Finally, a new label is accepted and added to the training set if the maximum probability that the image belongs to the corresponding rating class is greater than the threshold θ_1 and if the difference between the largest and the second largest class probability is greater than the threshold θ_2. Exemplarily, the resulting learning curves for 20% manually labelled images and $\theta_1 = 0.6$ and $\theta_2 = 0.1$ are shown in Figure 183.

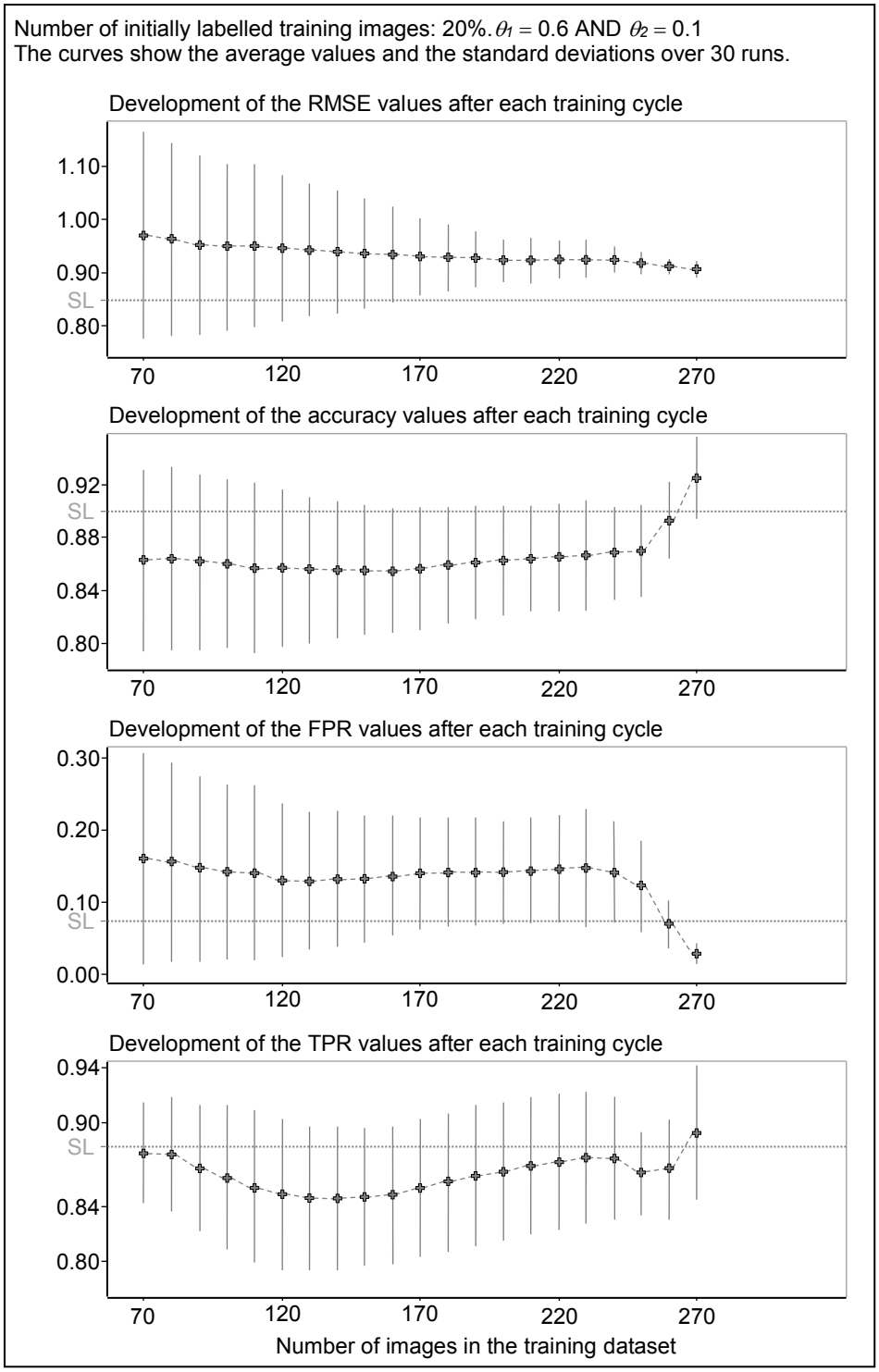


Figure 183: SSL, PC: learning curves for the double image dataset, θ_1 AND θ_2

A.15 SSL, kNN: results for the double image dataset

During the SSL process, an autonomously labelled image is transferred into the training dataset if the distance to the nearest reference pattern is smaller than a given threshold distance θ. Successively, the minimal, mean, and maximal determined distances, as shown in Table 36, are used as thresholds during the SSL process. The size of the initially labelled training set is set to 10%, 15%, …, 70% of all training images, as shown in Table 35. Due to the random selection of the initially labelled training images, the mean and the standard deviation over 30 runs are calculated. The RMSE values and the classification accuracies after the termination of the SSL process are summarised in Figure 184. The x-axes represent the number of initially labelled images in the training dataset and the y-axes the classification results.

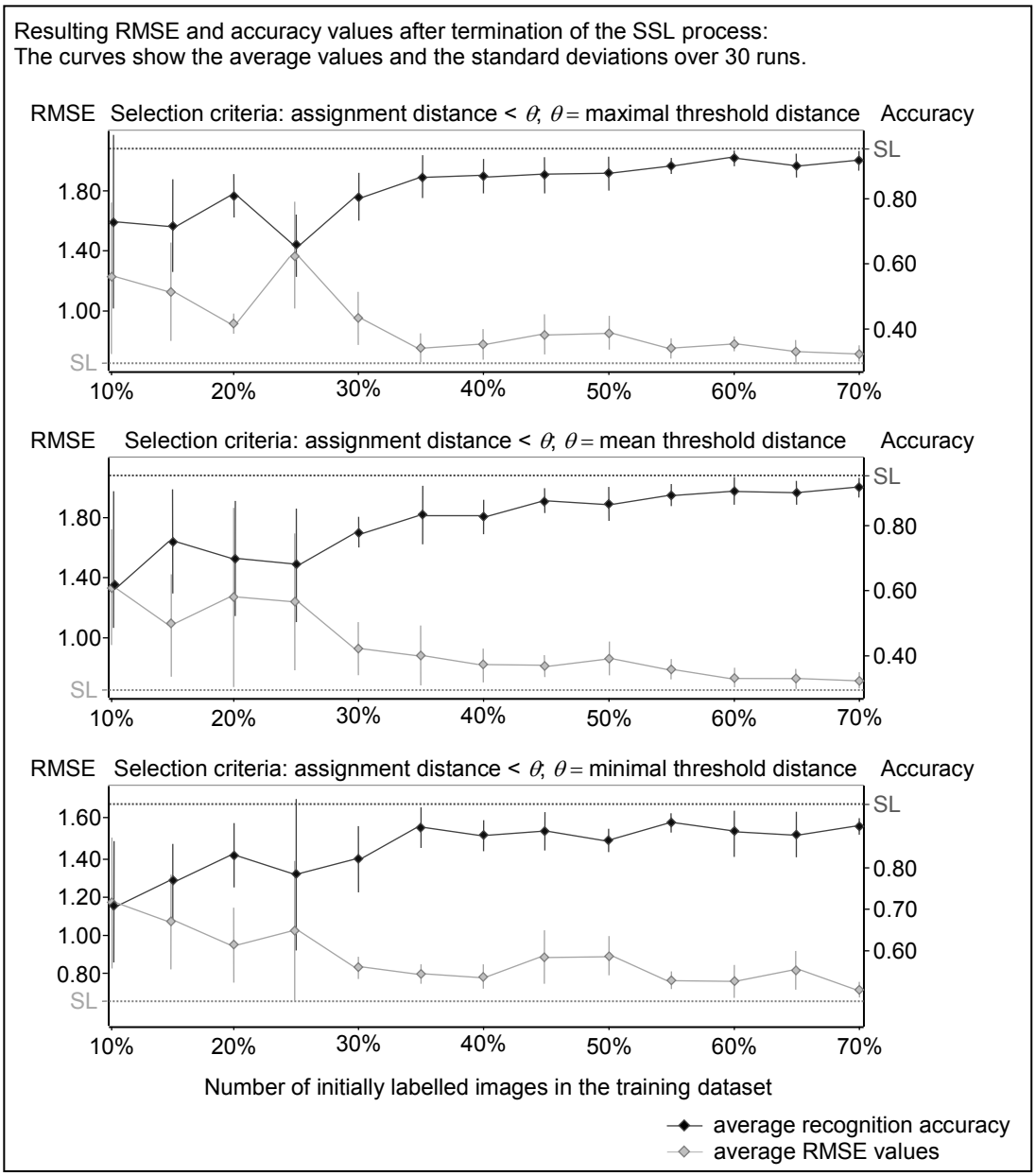


Figure 184: SSL, kNN: results for the double image dataset, part I

The TPR and FPR values after the termination of the SSL process are summarised in Figure 185. Since the selection of the labelled training data is done randomly, the average values and the standard deviations over 30 runs are calculated. The x-axes represent the number of initially labelled images in the training dataset and the y-axes the classification results.

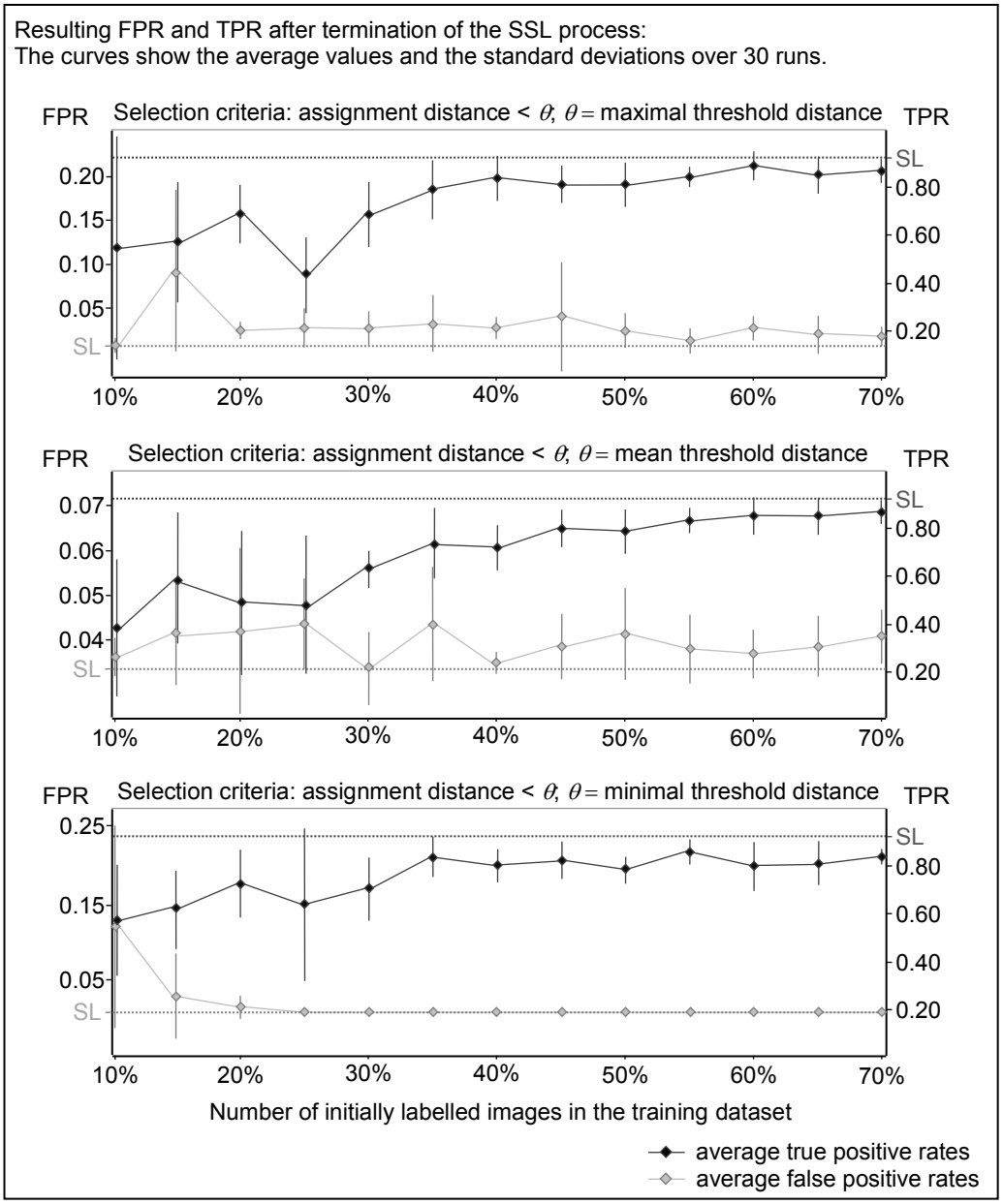


Figure 185: SSL, kNN: results for the double image dataset, part II

A.16 SSL, kNN: learning curves for the double image dataset

The learning curves for the distortion dataset for 35% manually labelled images are shown in Figure 186. A new label is accepted and added to the training set if the distance to the next training sample is smaller than the threshold distance.

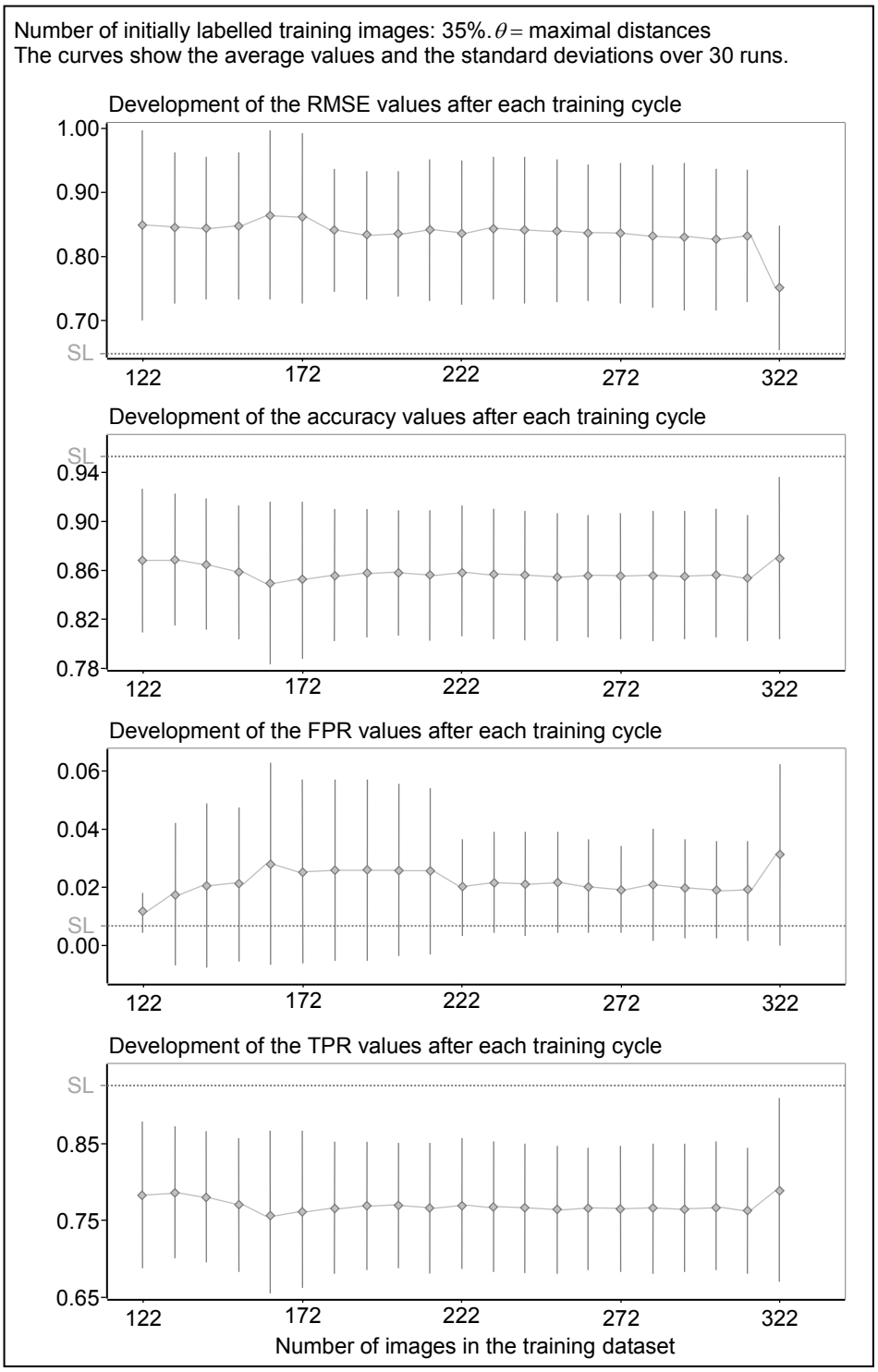


Figure 186: SSL, kNN: learning curves for the double image dataset

A.17 SSL, LVQ: results for the double image dataset

During the SSL process, an autonomously labelled image is transferred into the training dataset if the distance to the nearest prototype is smaller than a given threshold distance θ. Successively, the minimal, mean, and maximal determined distances, as shown in Table 36, are used as thresholds during the SSL process. The size of the initially labelled training set is set to 10%, 15%, …, 70% of all training images, as shown in Table 37. Due to the random selection of the initially labelled training images, the mean and the standard deviation over 30 runs are calculated. The RMSE values and the classification accuracies after the termination of the SSL process are summarised in Figure 187. The x-axes represent the number of initially labelled images in the training dataset and the y-axes the classification results.

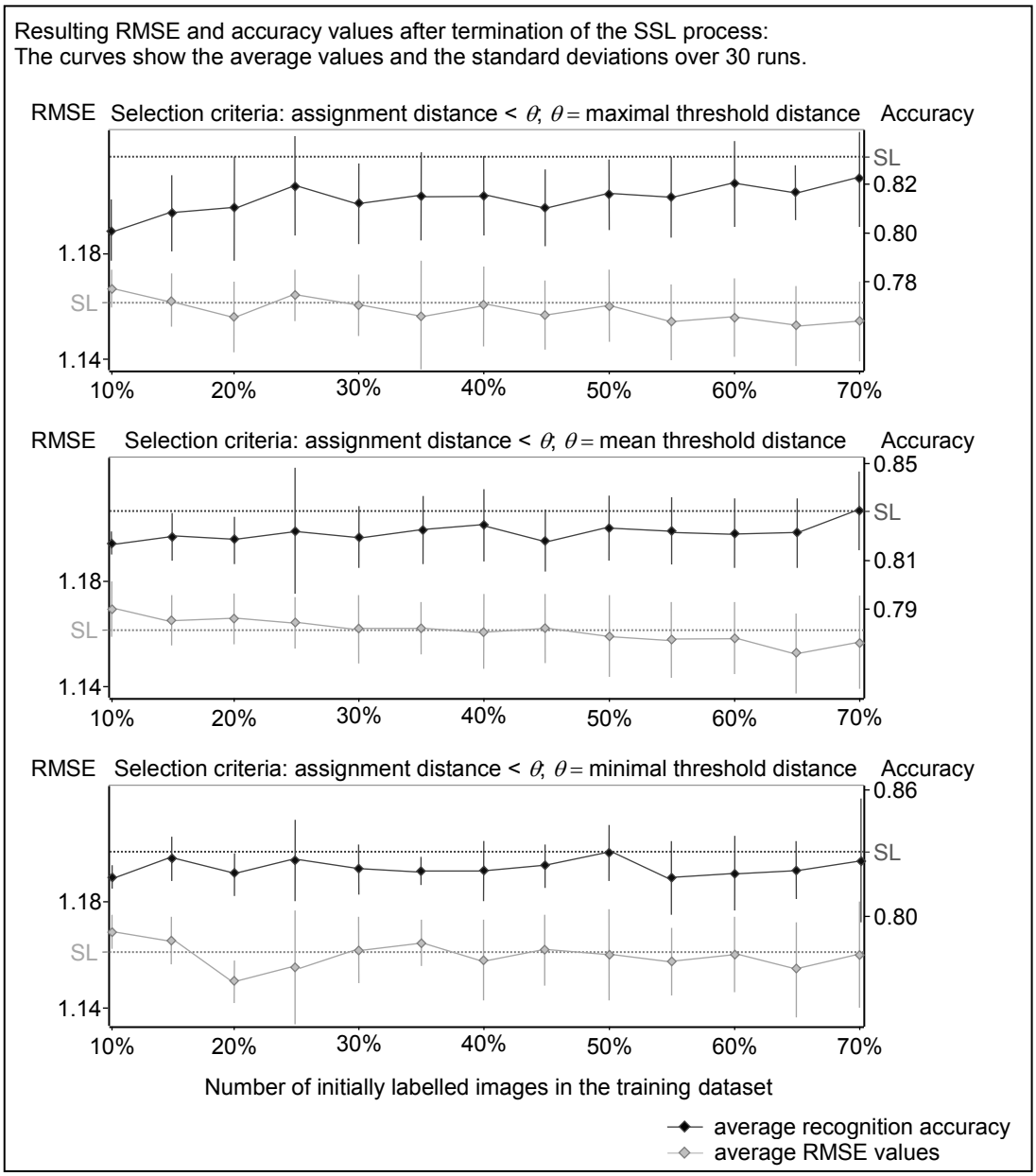


Figure 187: SSL, LVQ: results for the double image dataset, part I

The TPR and FPR values after the termination of the SSL process are summarised in Figure 188. Since the selection of the labelled training data is done randomly, the average values and the standard deviations over 30 runs are calculated. The x-axes represent the number of initially labelled images in the training dataset and the y-axes the classification results.

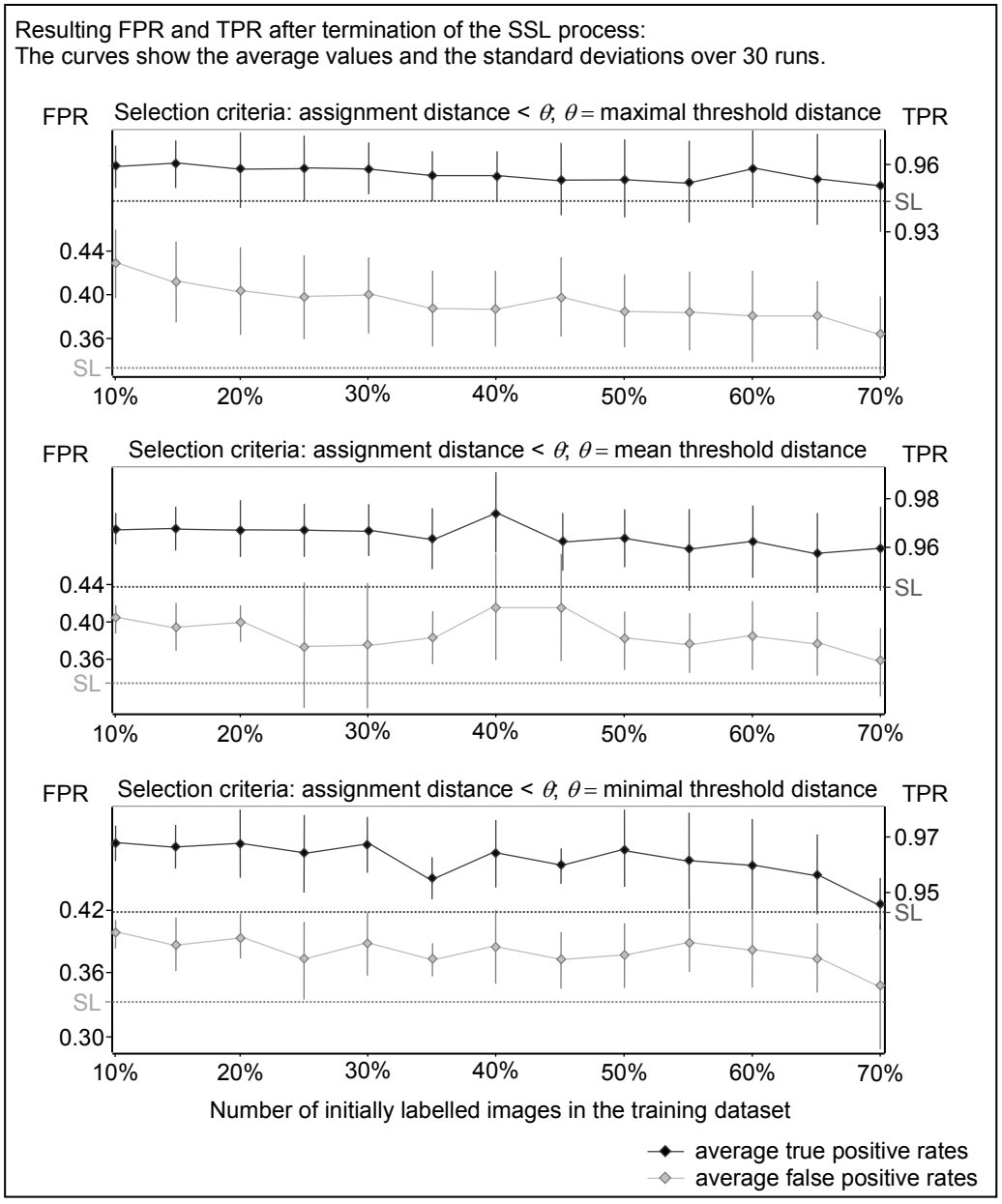


Figure 188: SSL, LVQ: results for the double image dataset, part II

A.18 SSL, LVQ: learning curves for the double image dataset

The learning curves for the distortion dataset for 25% manually labelled images are exemplarily shown in Figure 189. A new label is accepted and added to the training set if the distance to the next training sample is smaller than the threshold distance.

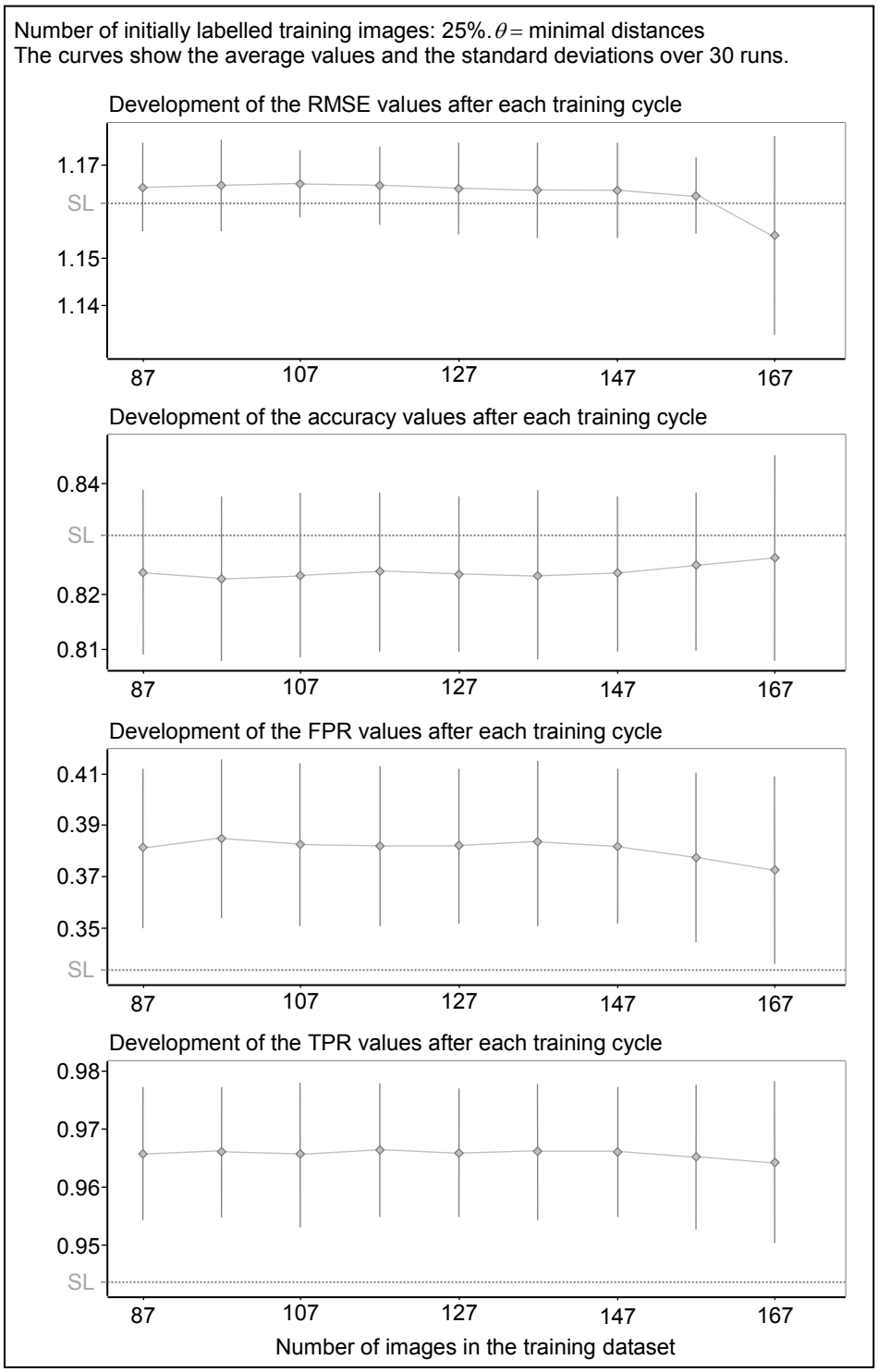


Figure 189: SSL, LVQ: learning curves for the double image dataset

A.19 SSL, PC: results for the distortion and double image dataset

In the first step, a new label is accepted and added to the training set if the maximum probability that the image belongs to the rating class is greater than the threshold θ_1. Here, θ_1 is set to 0.6, 0.7, and 0.8. The RMSE values and the classification accuracies after the termination of the SSL process are summarised in Figure 190. Since the selection of the labelled training data is done randomly, the average values and the standard deviations over 30 runs are calculated. The x-axes represent the number of initially labelled images in the training dataset and the y-axes the classification results.

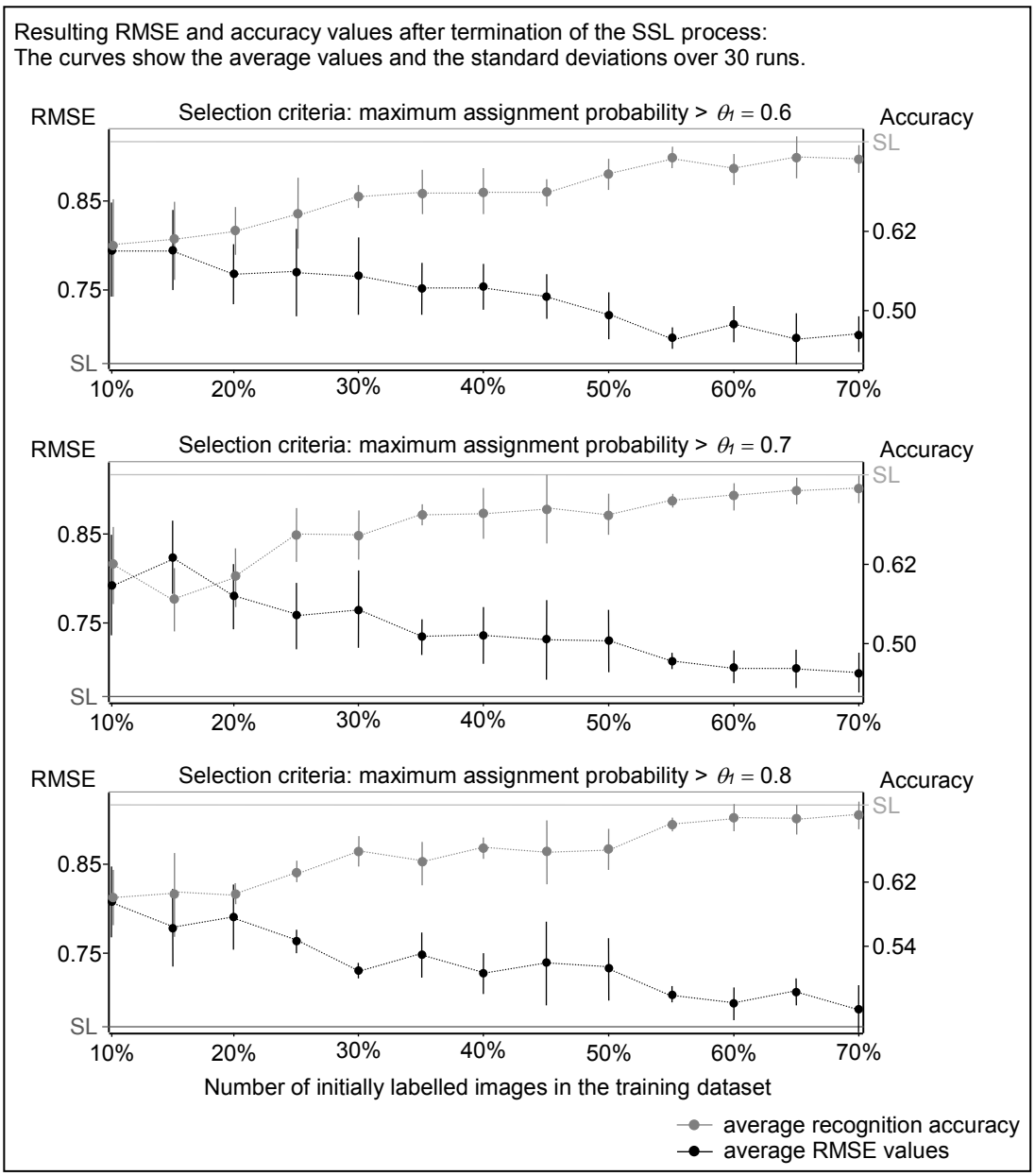


Figure 190: SSL, PC: results for the distortion and double image dataset θ_1, part I

The TPR and FPR values after the termination of the SSL process are summarised in Figure 191. Since the selection of the labelled training data is done randomly, the average values and the standard deviations over 30 runs are calculated. The x-axes represent the number of initially labelled images in the training dataset and the y-axes the classification results.

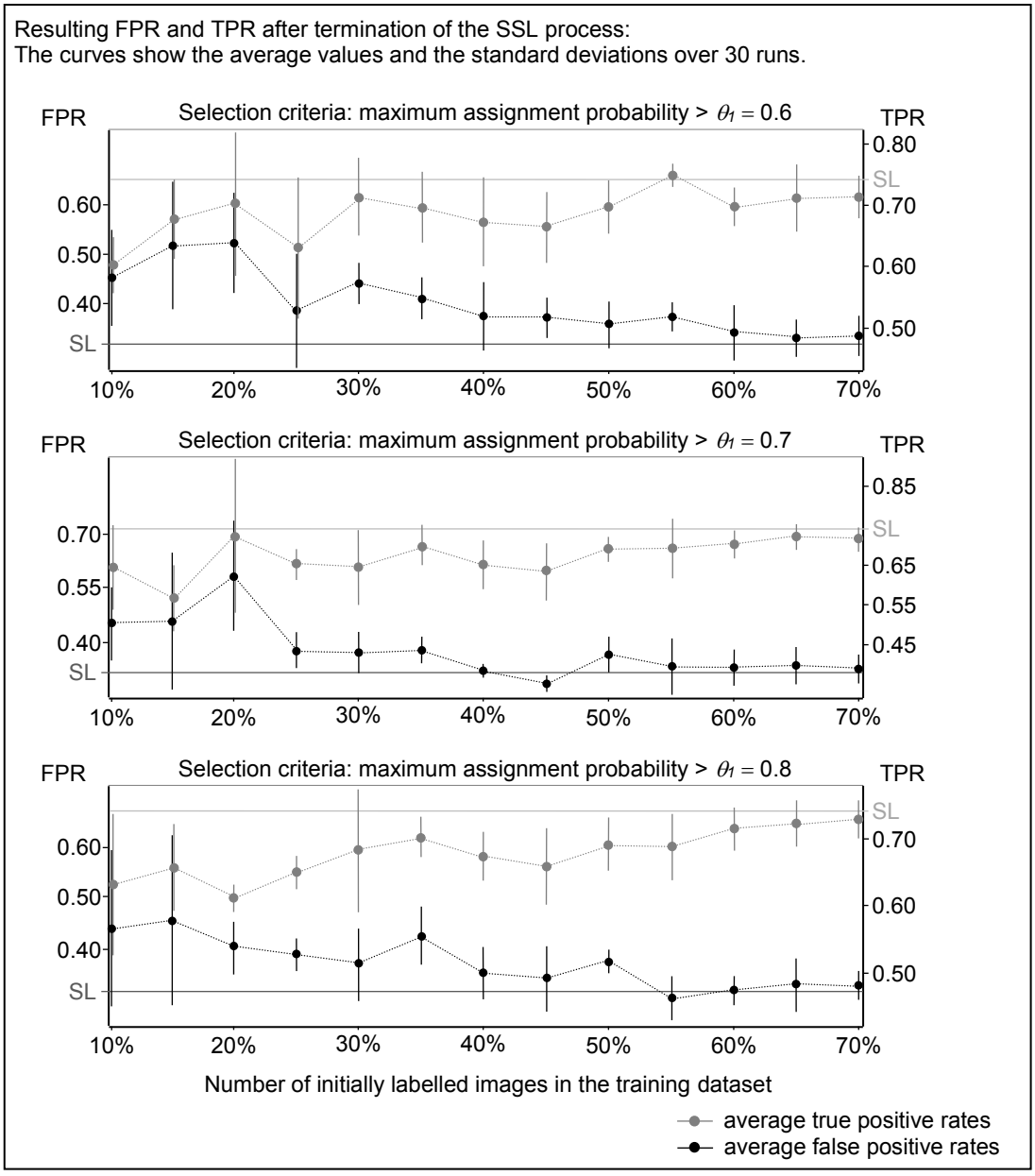


Figure 191: SSL, PC: results for the distortion and double image dataset θ_1, part II

In the second step, an autonomously labelled image is transferred into the training set if the difference between the largest and the second largest class probability is greater than the threshold θ_2, which is set to 0.1, 0.2, and 0.3. Again, the size of the initially labelled training images is set to 10%, 15%, …, 70% of all training images, as shown in Table 38. The RMSE values and the classification accuracies after the termination of the SSL process are summarised in Figure 192. Since the selection of the labelled training data is done randomly, the average values and the standard deviations over 30 runs are calculated. The x-axes represent the number of initially labelled images in the training dataset and the y-axes the classification results.

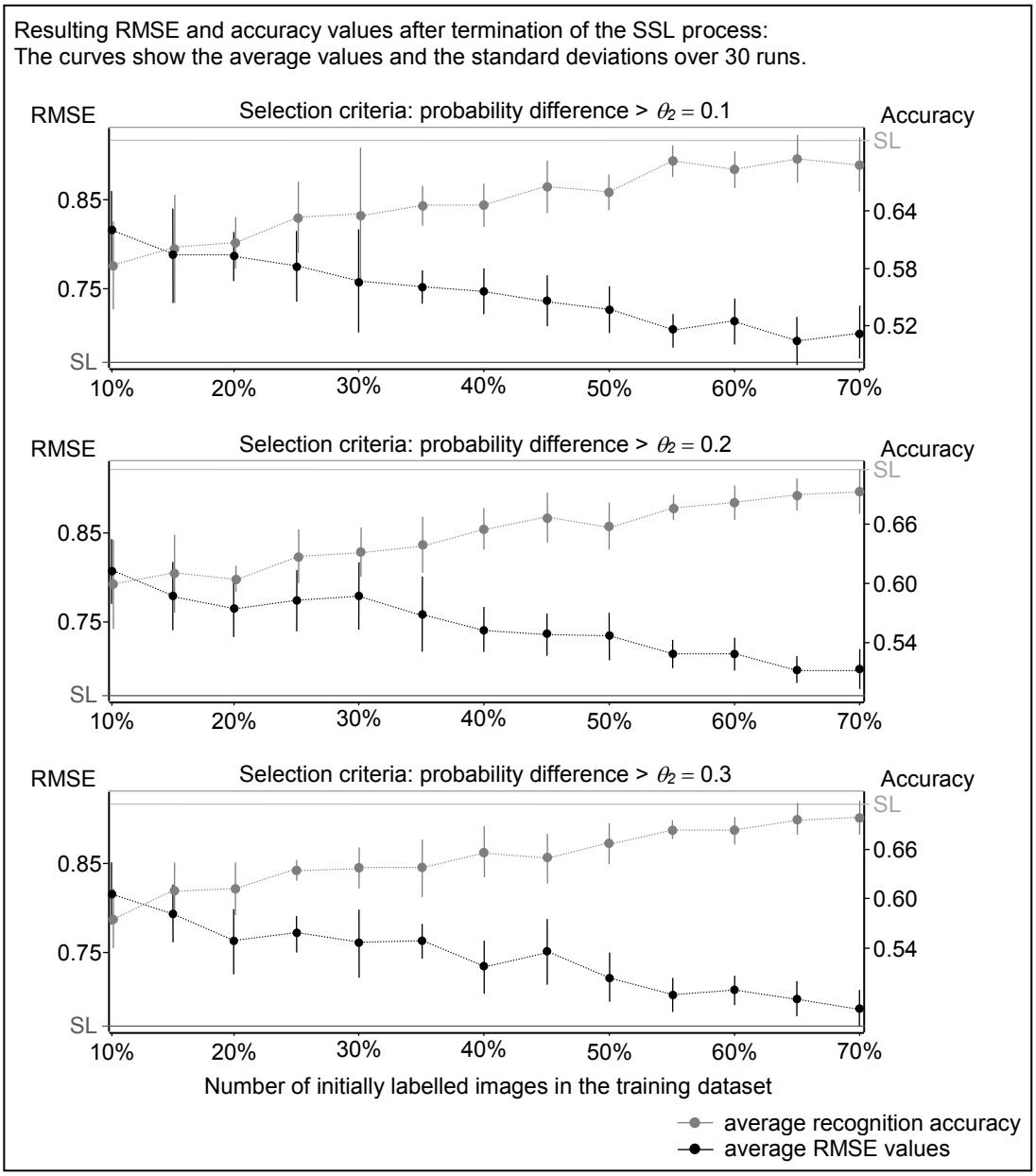


Figure 192: SSL, PC: results for the distortion and double image dataset, θ_2, part I

The TPR and FPR values after the termination of the SSL process are summarised in Figure 193. Since the selection of the labelled training data is done randomly, the average values and the standard deviations over 30 runs are calculated. The x-axes represent the number of initially labelled images in the training dataset and the y-axes the classification results.

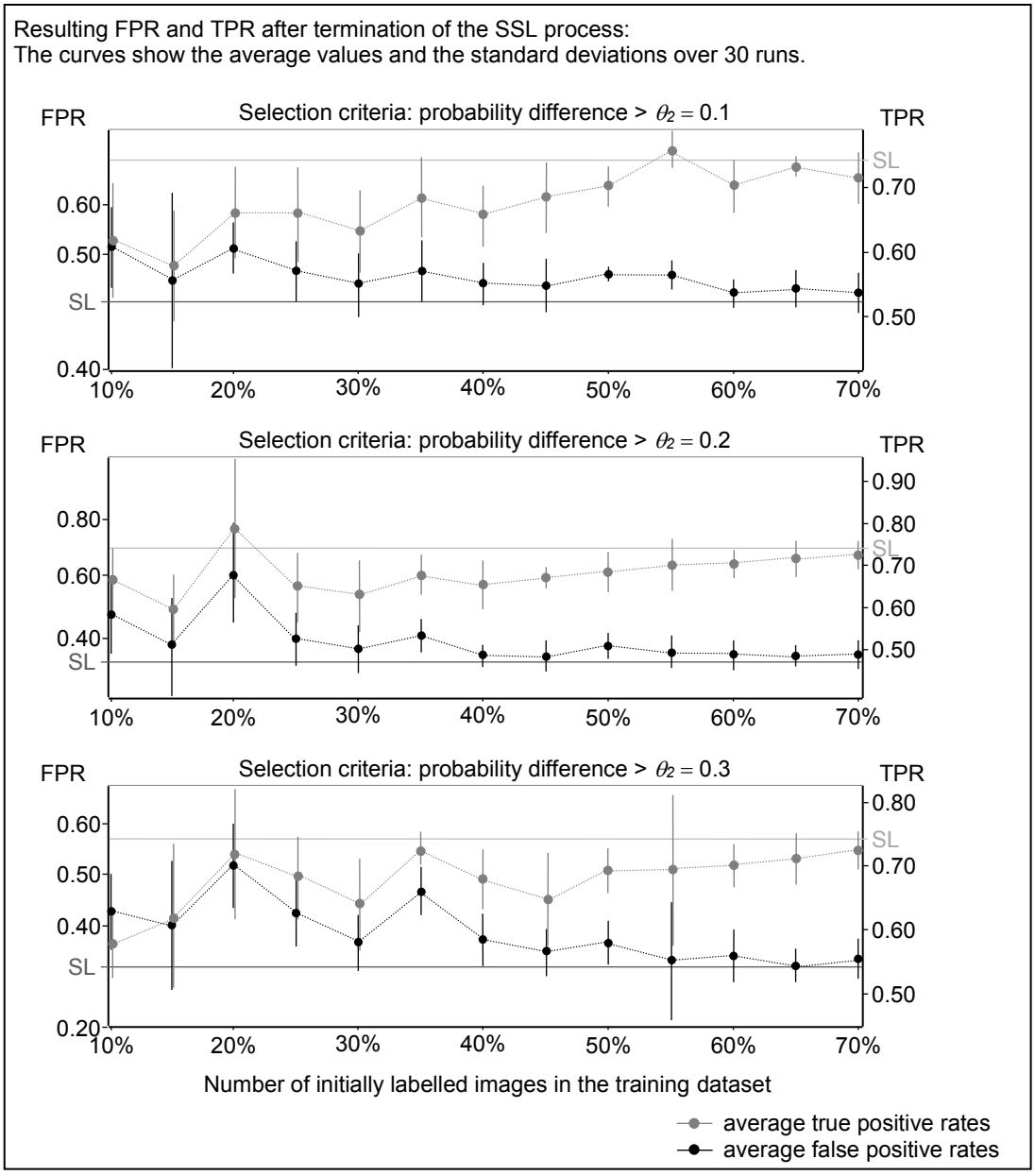


Figure 193: SSL, PC: results for the distortion and double image dataset θ_2, part II

In the last step, both selection criteria are combined. An autonomously labelled image is transferred into the training set if the class assignment probabilities fulfil both threshold conditions simultaneously. The maximum probability that the image belongs to the corresponding rating class has to be greater than the threshold θ_1 and the difference between the largest and the second largest class probability must be greater than the threshold θ_2. The RMSE values and the classification accuracies after the termination of the SSL process are summarised in Figure 194. Since the selection of the labelled training data is done randomly, the average values and the standard deviations over 30 runs are calculated. The x-axes represent the number of initially labelled images in the training dataset and the y-axes the classification results.

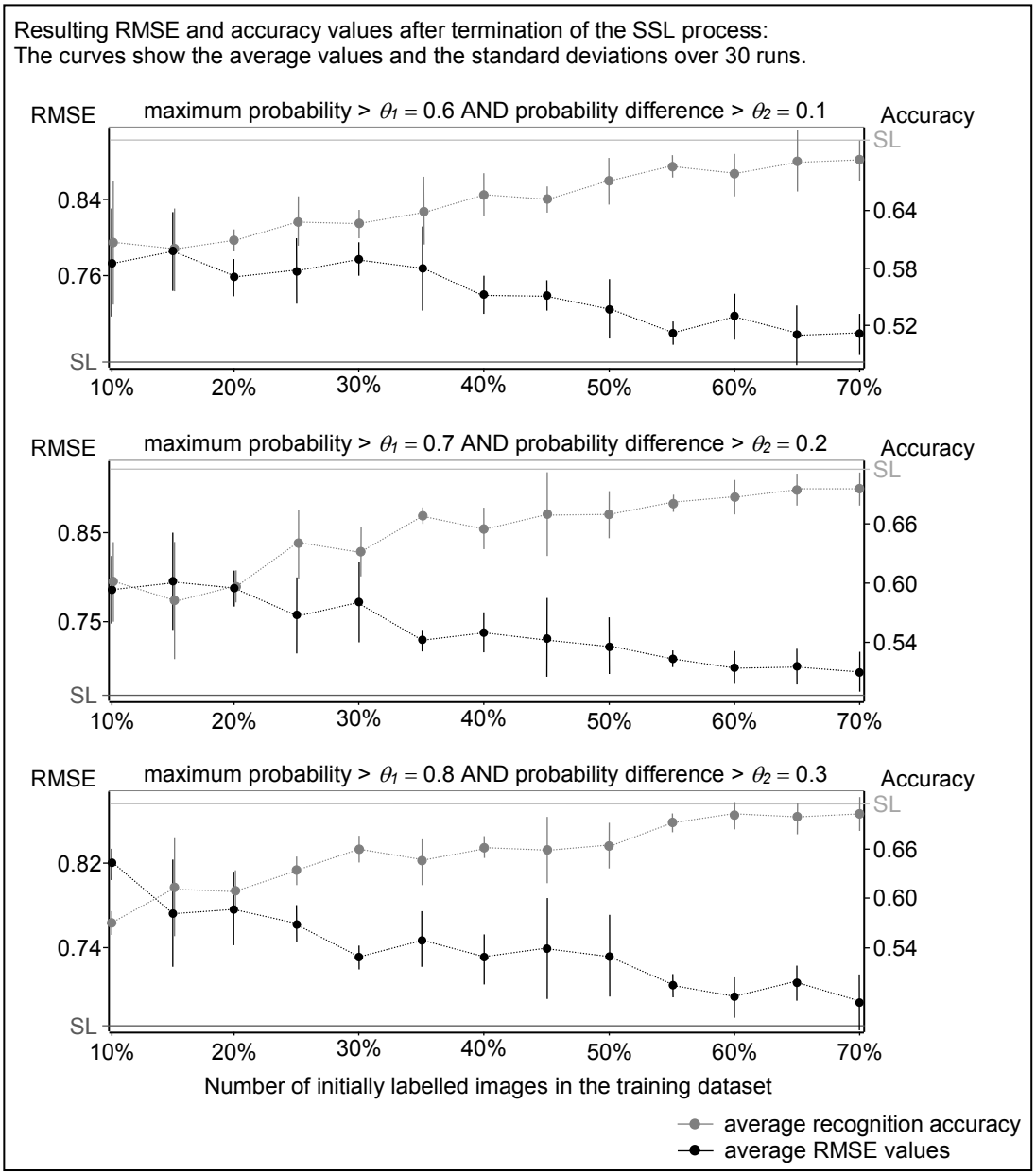


Figure 194: SSL, PC: results for the distortion and double image dataset, θ_1 AND θ_2, part I

The TPR and FPR values after the termination of the SSL process are summarised in Figure 195. Since the selection of the labelled training data is done randomly, the aver-

age values and the standard deviations over 30 runs are calculated. The x-axes represent the number of initially labelled images in the training dataset and the y-axes the classification results.

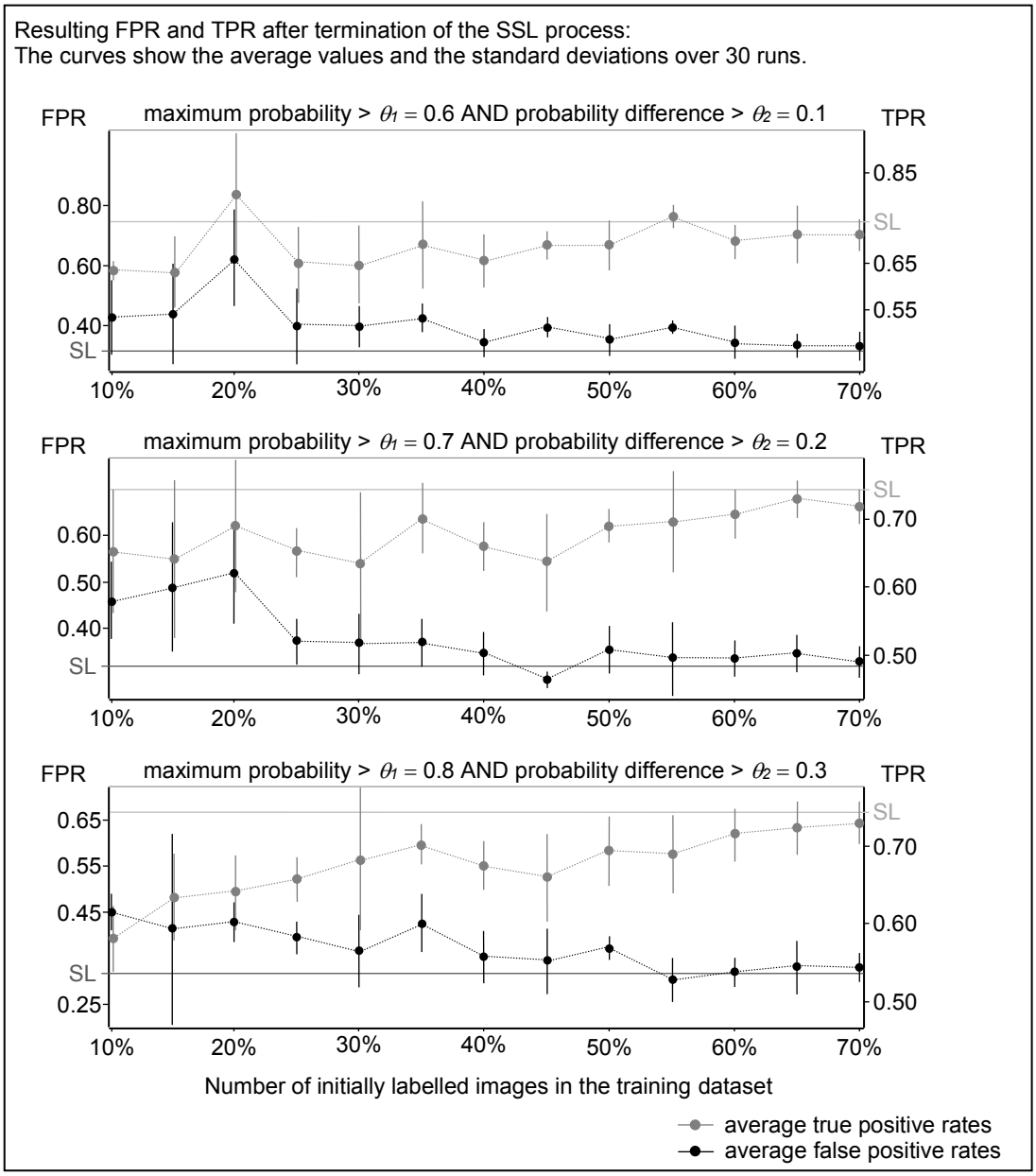


Figure 195: SSL, PC: results for the distortion and double image dataset, θ_1 AND θ_2, part II

A.20 SSL, PC: learning curves for the distortion and double image dataset

A new label is added to the training set if the maximum probability that the image belongs to the corresponding rating class is greater than the threshold θ_1, which is set to 0.7. The learning curves for 35% manually labelled images are shown in Figure 196.

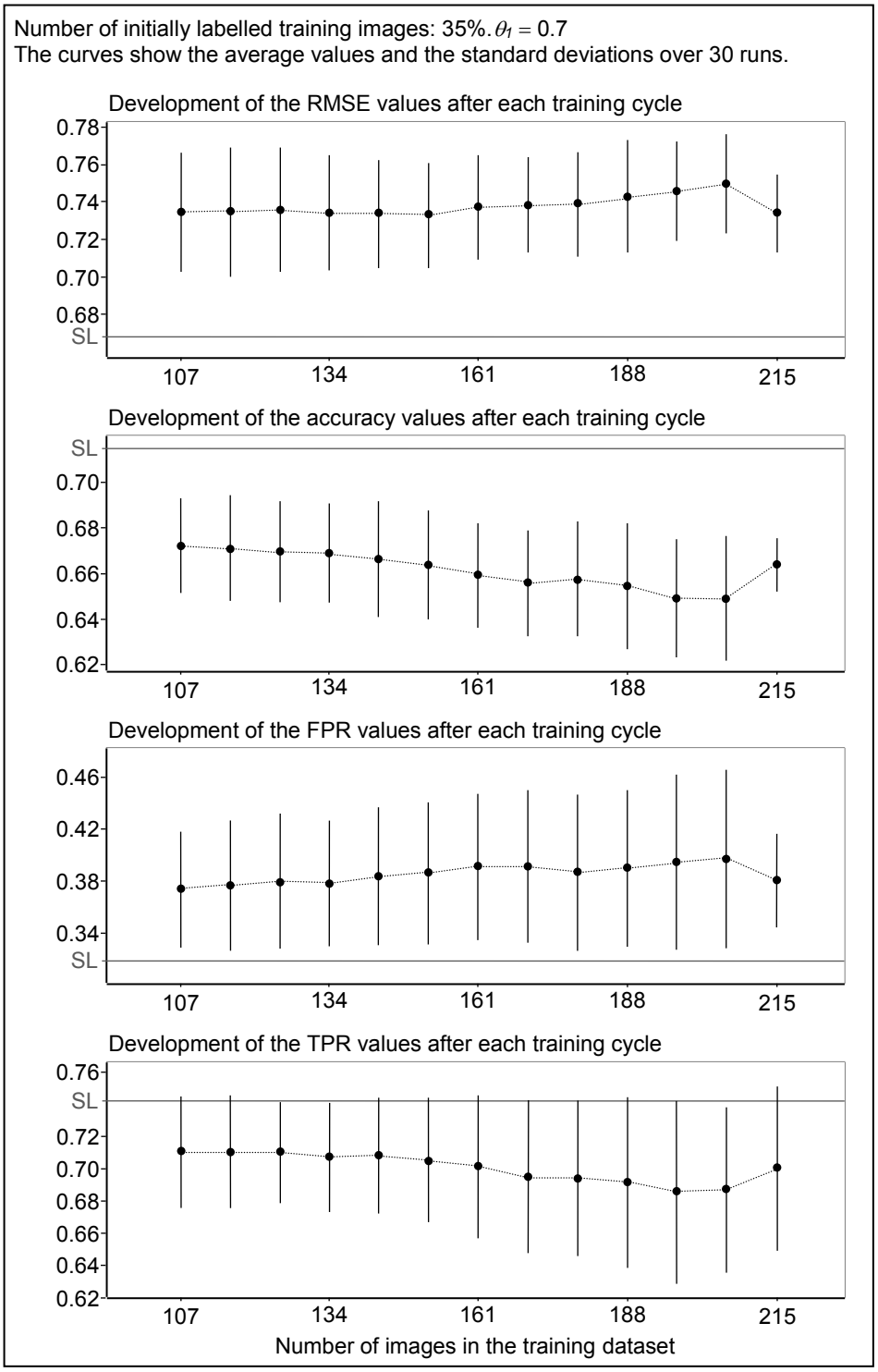


Figure 196: SSL, PC: learning curves for the distortion and double image dataset, θ_1

In the second step, if the difference between the largest and the second largest class probability is greater than the threshold θ_2, a new label is accepted and added to the training set. Here, θ_2 is exemplarily set to 0.2. The corresponding learning curves for 15% manually labelled images are shown in Figure 197.

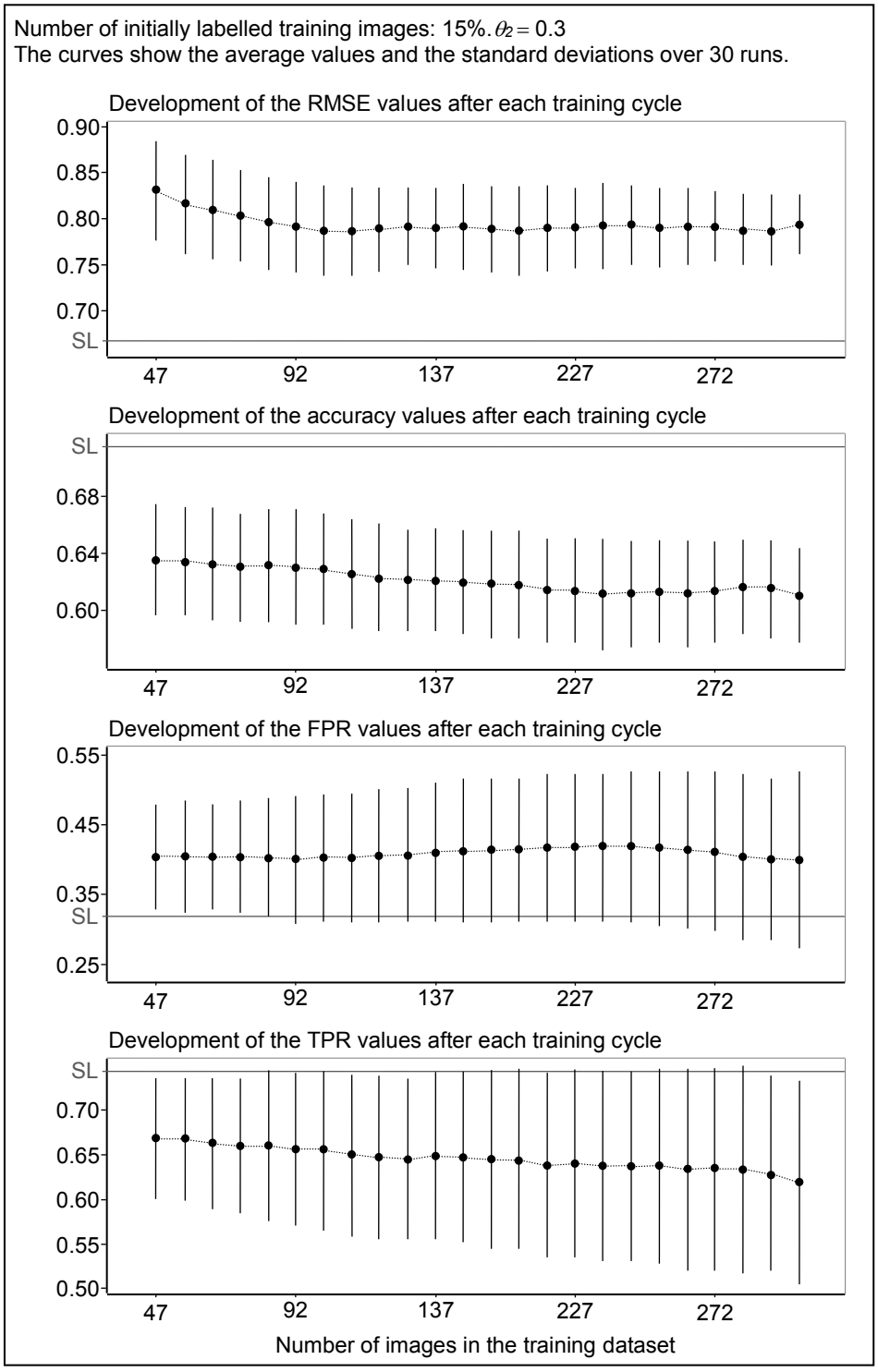


Figure 197: SSL, PC: learning curves for the distortion and double image dataset, θ_2

Finally, a new label is accepted and added to the training set if the maximum probability that the image belongs to the corresponding rating class is greater than the threshold θ_1 and if the difference between the largest and the second largest class probability is greater than the threshold θ_2. The resulting learning curves for 30% manually labelled images and $\theta_1 = 0.8$ and $\theta_2 = 0.3$ are shown in Figure 198.

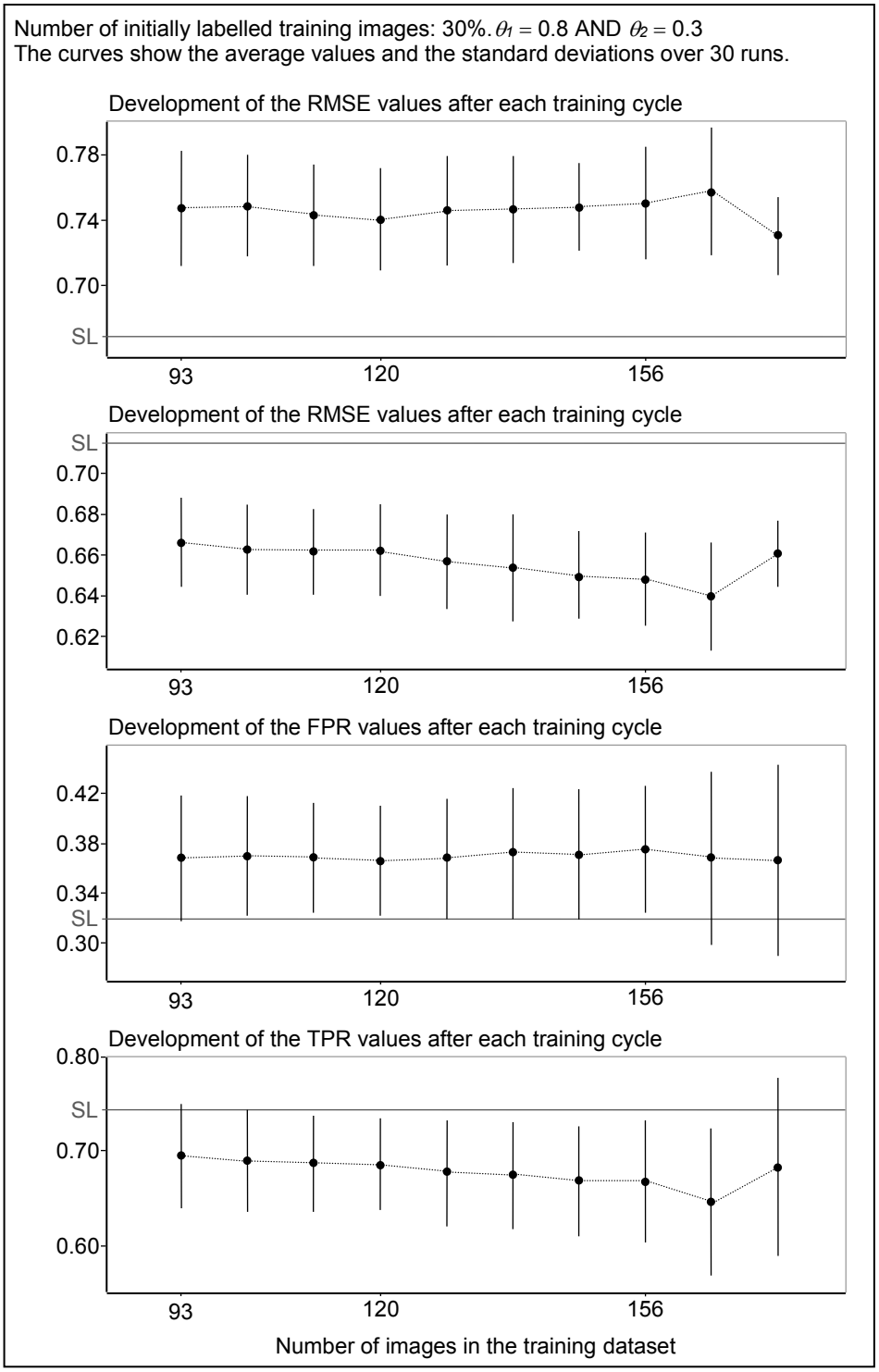


Figure 198: SSL, PC: learning curves for the distortion and double image dataset θ_1 AND θ_2

A.21 SSL, kNN: results for the distortion and double image dataset

During the SSL process, an autonomously labelled image is transferred into the training dataset if the distance to the nearest reference pattern is smaller than a given threshold distance θ. Successively, the minimal, mean, and maximal determined distances, as shown in Table 39, are used as thresholds during the SSL process. Due to the random selection of the initially labelled training images, the mean and the standard deviation over 30 runs are calculated. The RMSE values and the classification accuracies after the termination of the SSL process are summarised in Figure 199. The x-axes represent the number of initially labelled images in the training dataset and the y-axes the classification results.

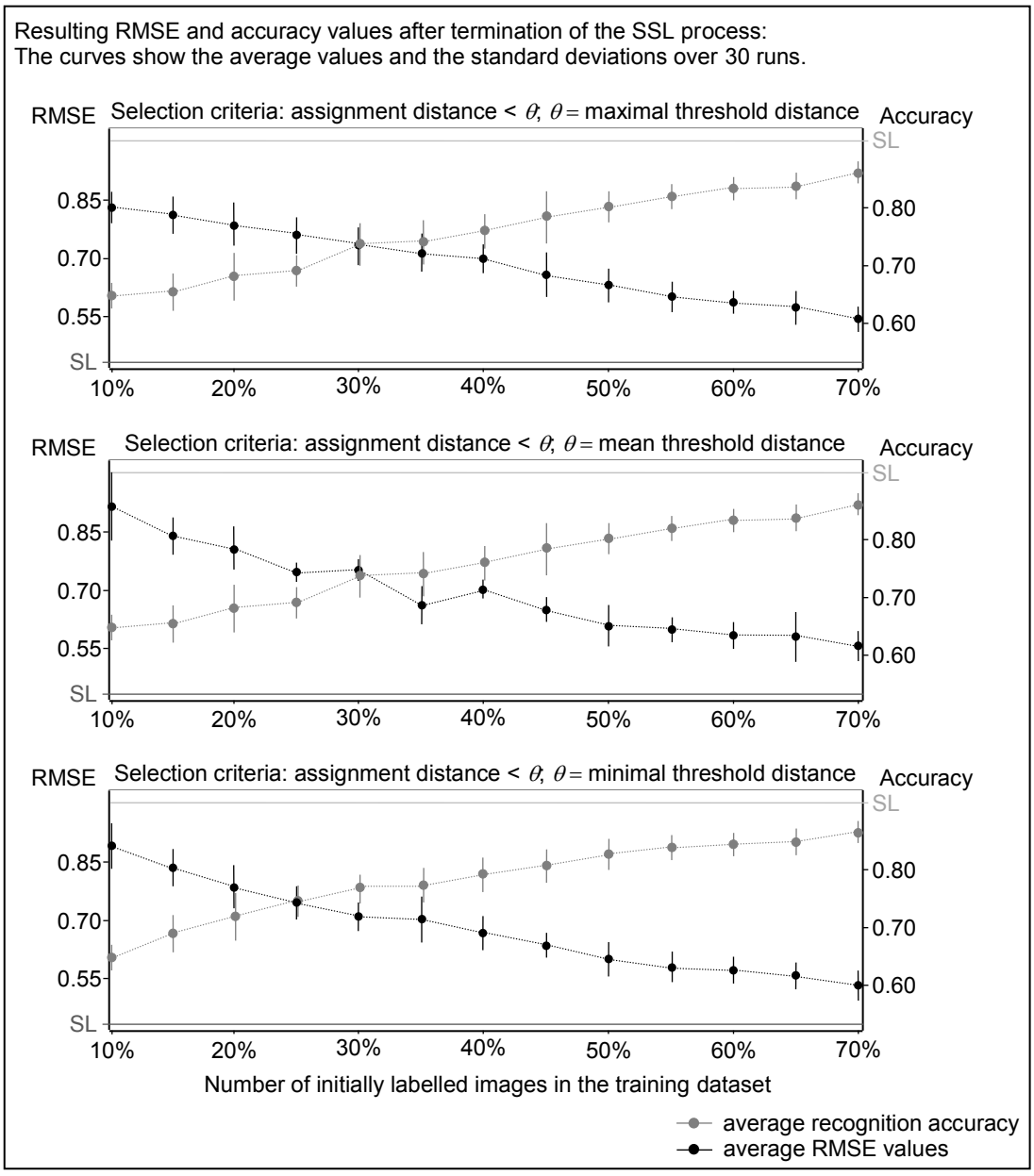


Figure 199: SSL, kNN: results for the distortion and double image dataset part I

The TPR and FPR values after the termination of the SSL process are summarised in Figure 200. Since the selection of the labelled training data is done randomly, the average values and the standard deviations over 30 runs are calculated. The x-axes represent the number of initially labelled images in the training dataset and the y-axes the classification results.

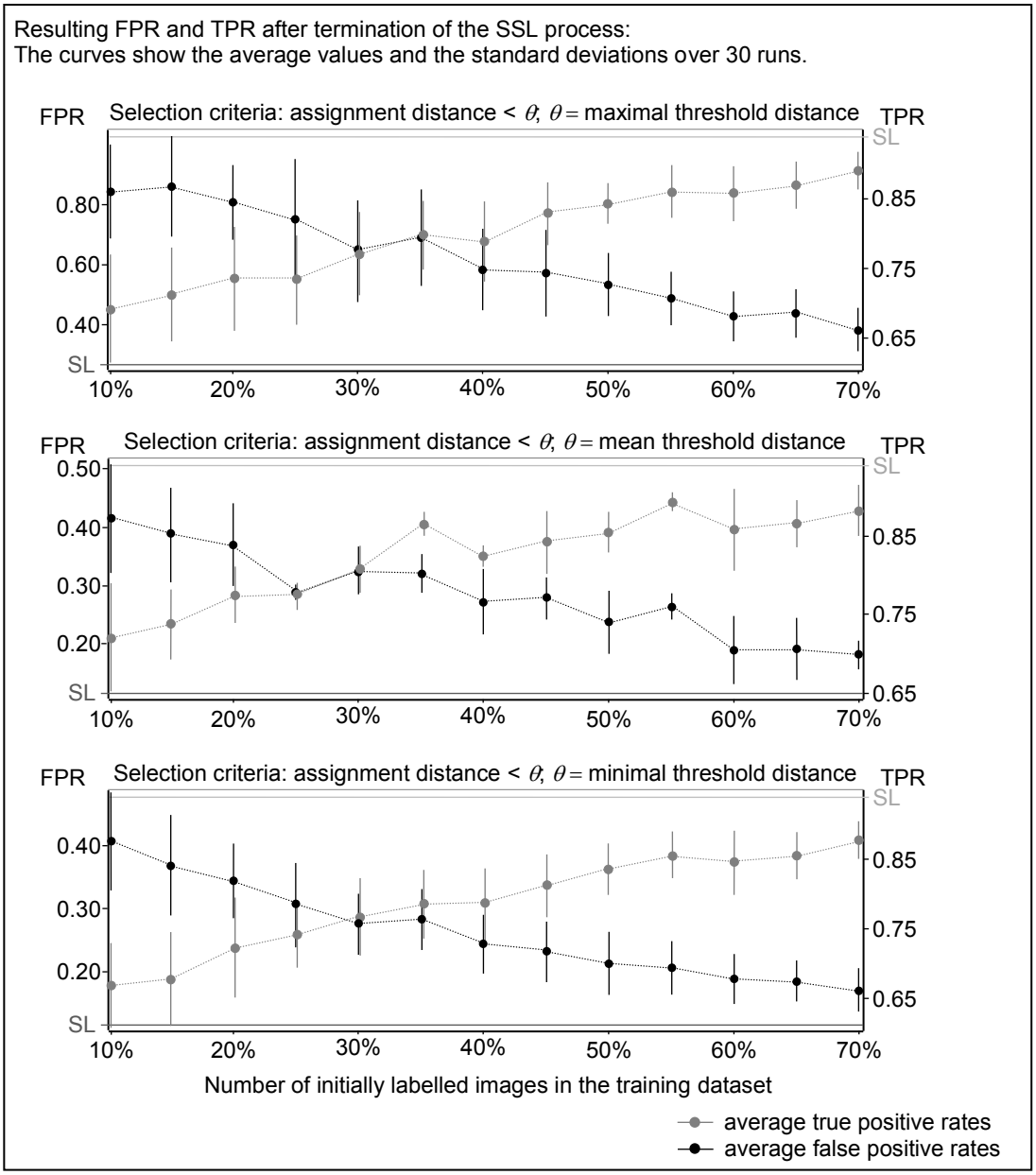


Figure 200: SSL, kNN: results for the distortion and double image dataset part II

A.22 SSL, kNN: learning curves for the distortion and double image dataset

The learning curves for the distortion dataset for 35% manually labelled images are shown in Figure 201. A new label is accepted and added to the training set if the distance to the next training sample is smaller than the threshold distance.

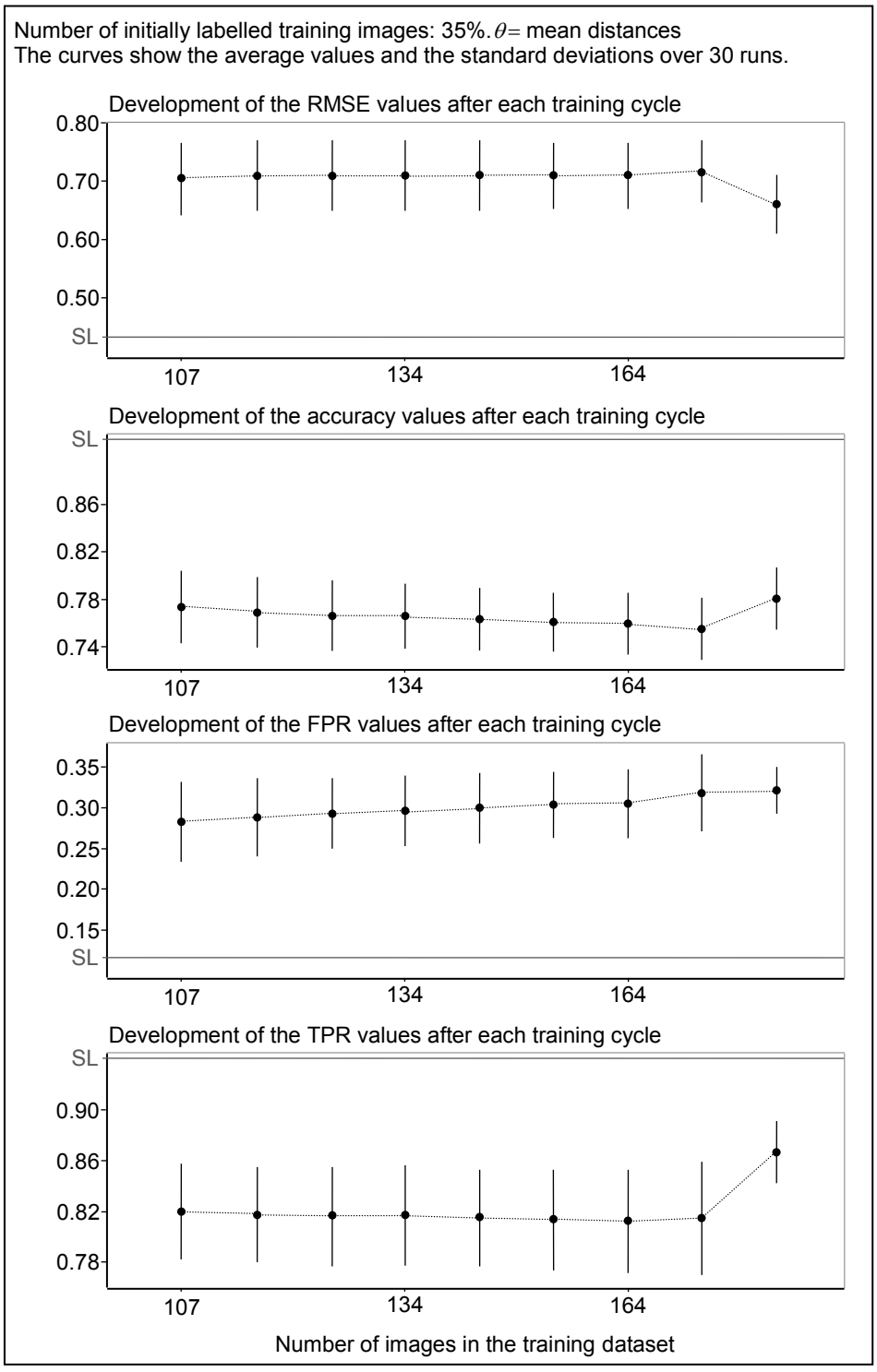


Figure 201: SSL, kNN: learning curves for the distortion and double image dataset

A.23 SSL, LVQ: results for the distortion and double image dataset

During the SSL process, an autonomously labelled image is transferred into the training dataset if the distance to the nearest prototype is smaller than a given threshold distance θ. Successively, the minimal, mean, and maximal determined distances, as shown in Table 40, are used as thresholds during the SSL process. Due to the random selection of the initially labelled training images, the mean and the standard deviation over 30 runs are calculated. The RMSE values and the classification accuracies after the termination of the SSL process are summarised in Figure 202. The x-axes represent the number of initially labelled images in the training dataset and the y-axes the classification results.

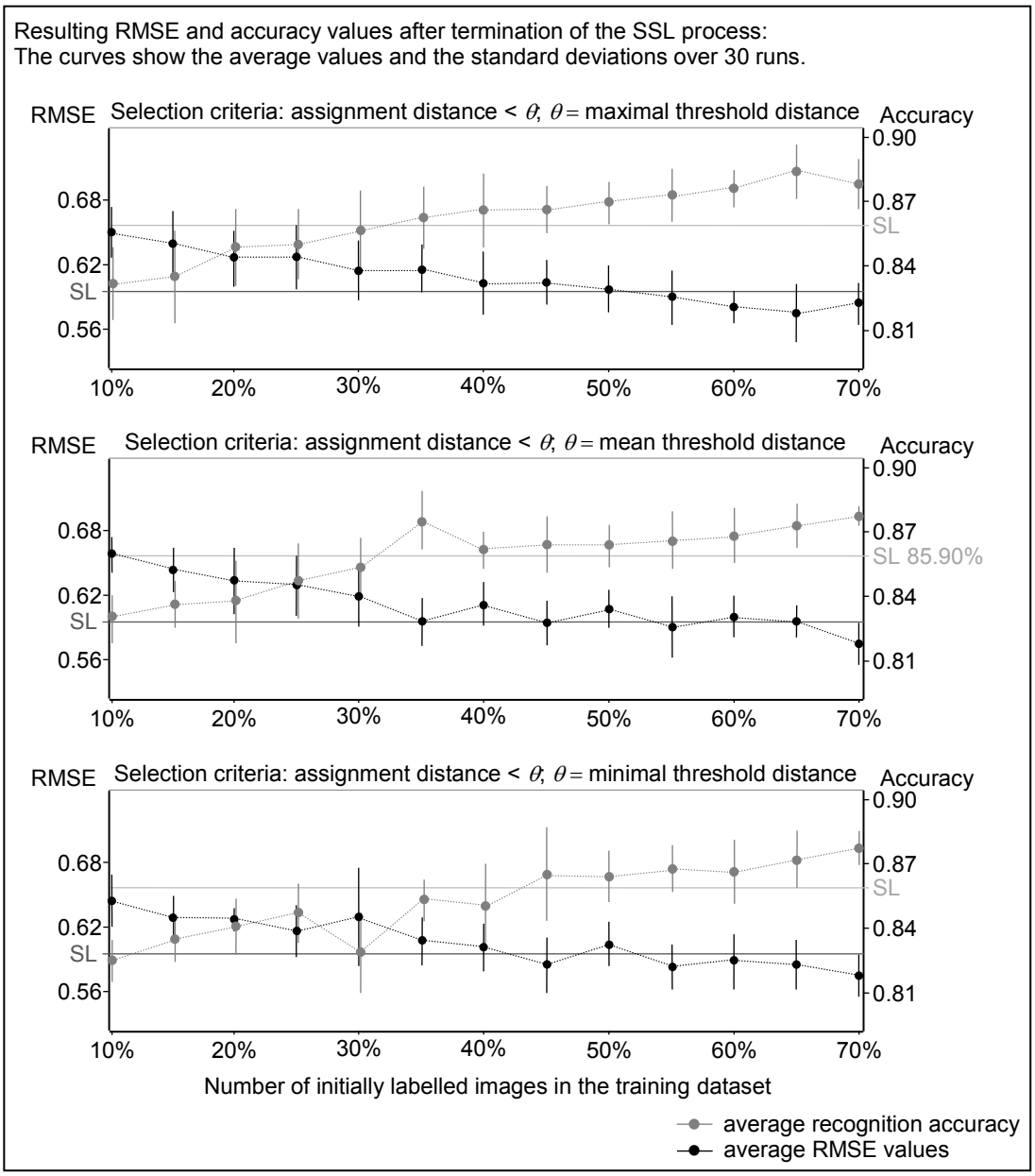


Figure 202: SSL, LVQ: results for the distortion and double image dataset, part I

The TPR and FPR values after the termination of the SSL process are summarised in Figure 203. Since the selection of the labelled training data is done randomly, the average values and the standard deviations over 30 runs are calculated. The x-axes represent the number of initially labelled images in the training dataset and the y-axes the classification results.

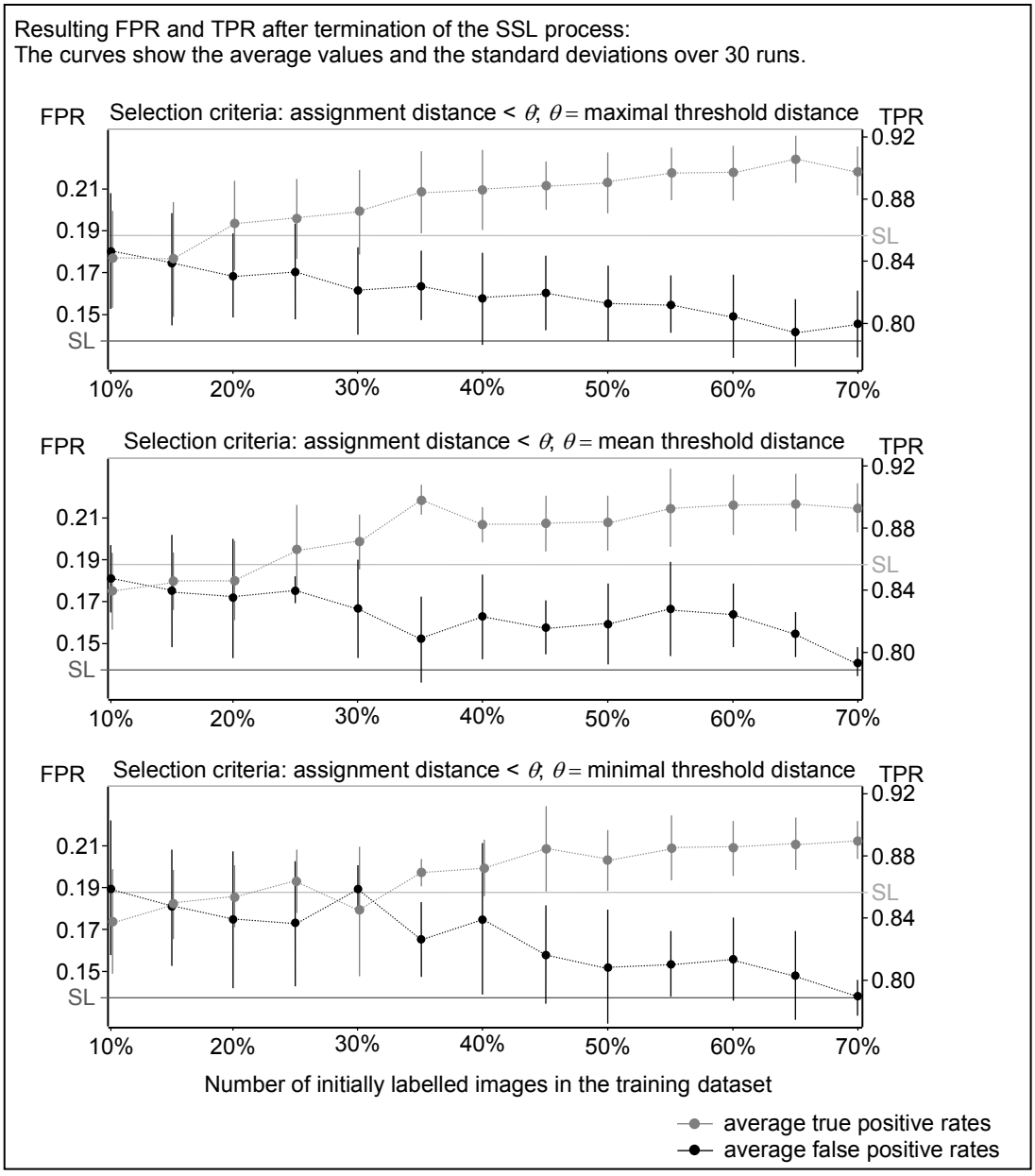


Figure 203: SSL, LVQ: results for the distortion and double image dataset part II

A.24 SSL, LVQ: learning curves for the distortion and double image dataset

The learning curves for the distortion dataset for 35% manually labelled images are exemplarily shown in Figure 204. A new label is accepted and added to the training set if the distance to the next training sample is smaller than the threshold distance.

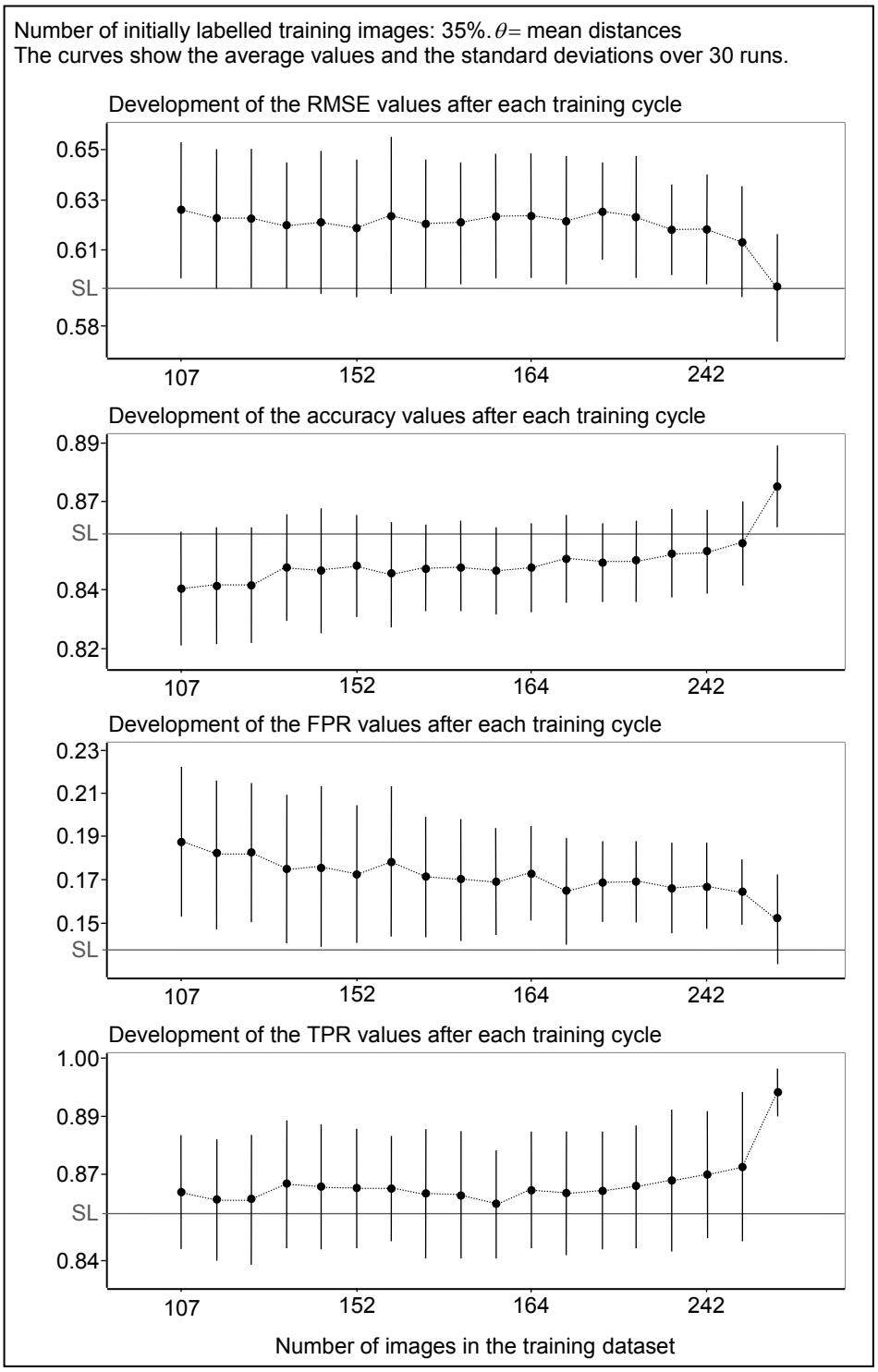


Figure 204: SSL, LVQ: learning curves for the distortion and double image dataset

A.25 AL, PC: results for the distortion dataset

First, a training image is labelled by the Oracle if the maximum probability that the image belongs to the corresponding rating class is lower than the threshold θ_1. Here, θ_1 is set to 0.6, 0.7, and 0.8. The size of the initial training set is to 0.5%, 2%, and 5% of all training images from each rating class, as shown in Table 41. The classification results after the termination of the AL process are summarised in Figure 205 and Figure 206. Since the selection of the labelled training data is done randomly, the average values and the standard deviations over 30 runs are calculated. The x-axes represent the number of initially labelled images in the training dataset and the y-axes the classification results.

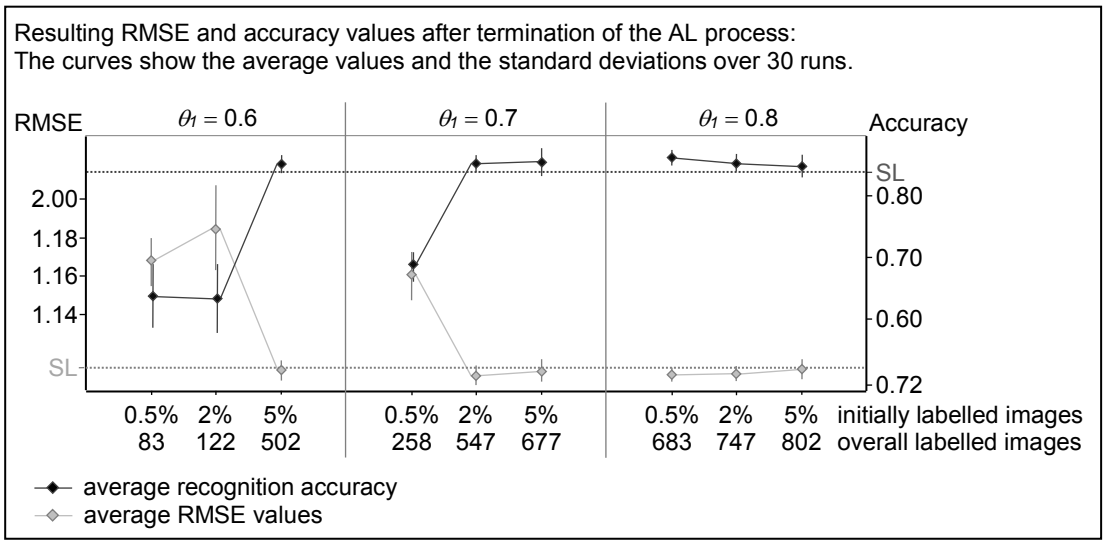


Figure 205: AL, PC: results for the distortion dataset, θ_1, part I

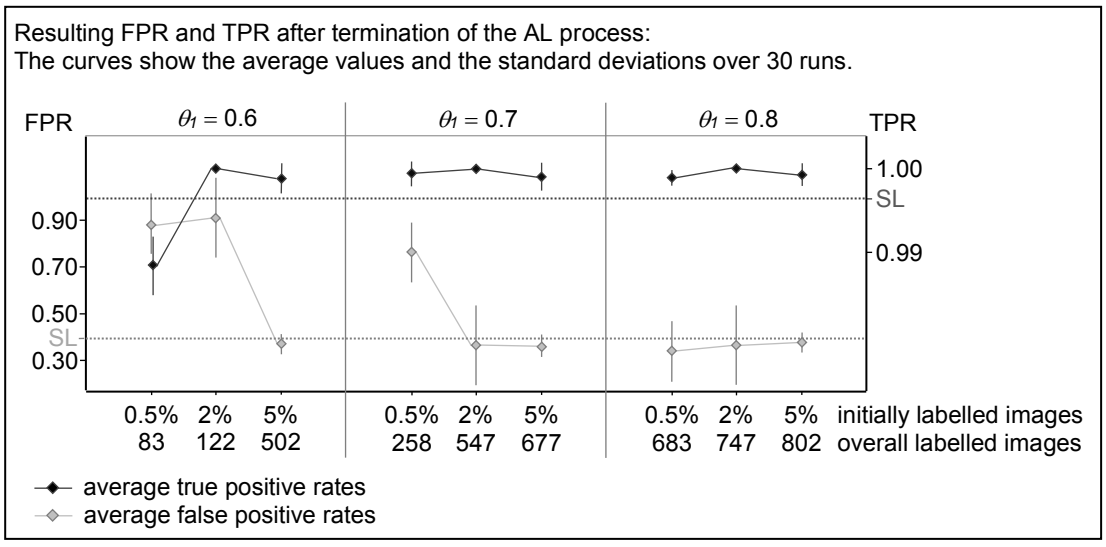


Figure 206: AL, PC: results for the distortion dataset, θ_1, part II

In the second step, the Oracle is asked for the right label if the difference between the largest and the second largest class probability is less than a threshold θ_2, which is set to 0.1, 0.2, and 0.3. The size of the initial training set is to 0.5%, 2%, and 5% of all training images from each rating class, as shown in Table 41. The classification results after the termination of the AL process are summarised in Figure 207 and Figure 208. Since the selection of the labelled training data is done randomly, the average values and the standard deviations over 30 runs are calculated. The x-axes represent the number of initially labelled images in the training dataset and the y-axes the classification results.

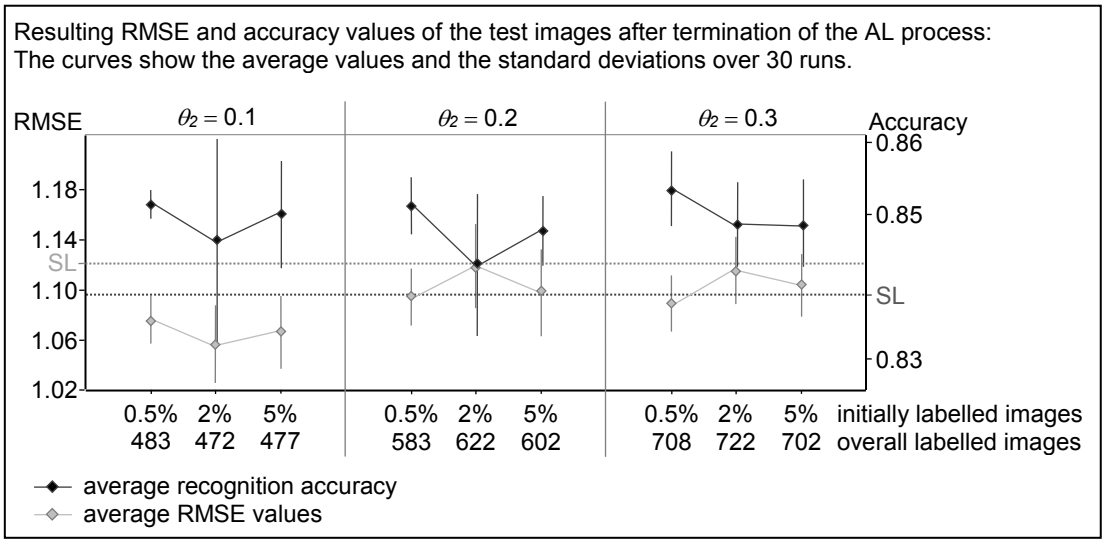


Figure 207: AL, PC: results for the distortion dataset, θ_2, part I

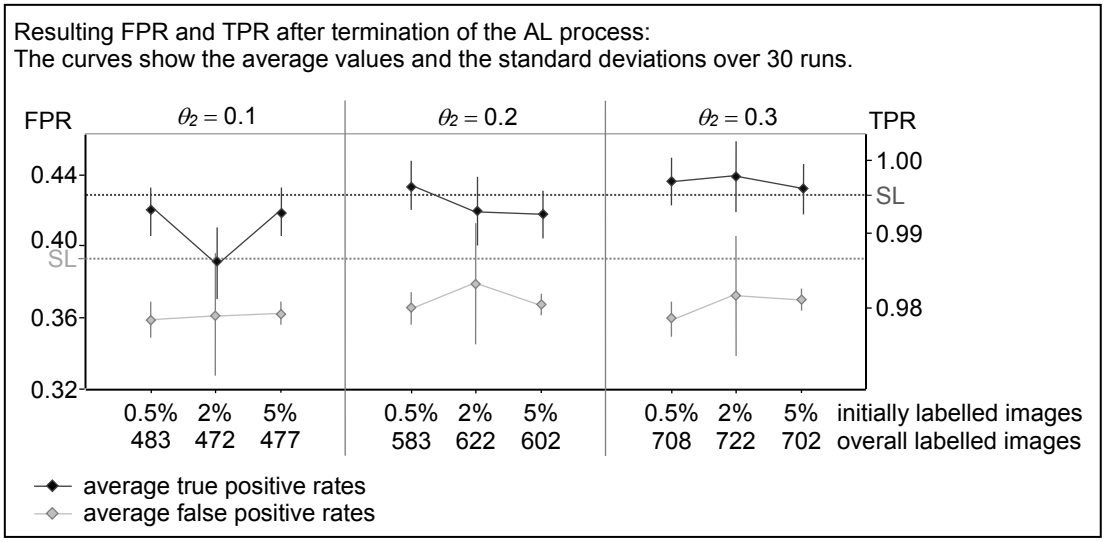


Figure 208: AL, PC: results for the distortion dataset, θ_2, part II

Finally, the 2 selection criteria, which determine if the training image needs to the labelled by the Oracle, are combined. A training image needs to be labelled manually if the maximum probability that the image belongs to the corresponding rating class is lower than the threshold θ_1 and if the difference between the largest and the second largest class probability is less than a threshold θ_2. The size of the initial training set is to 0.5%, 2%, and 5% of all training images from each rating class, as shown in Table 41. The classification results after the termination of the AL process are summarised in Figure 209 and Figure 210. Since the selection of the labelled training data is done randomly, the average values and the standard deviations over 30 runs are calculated. The x-axes represent the number of initially labelled images in the training dataset and the y-axes the classification results.

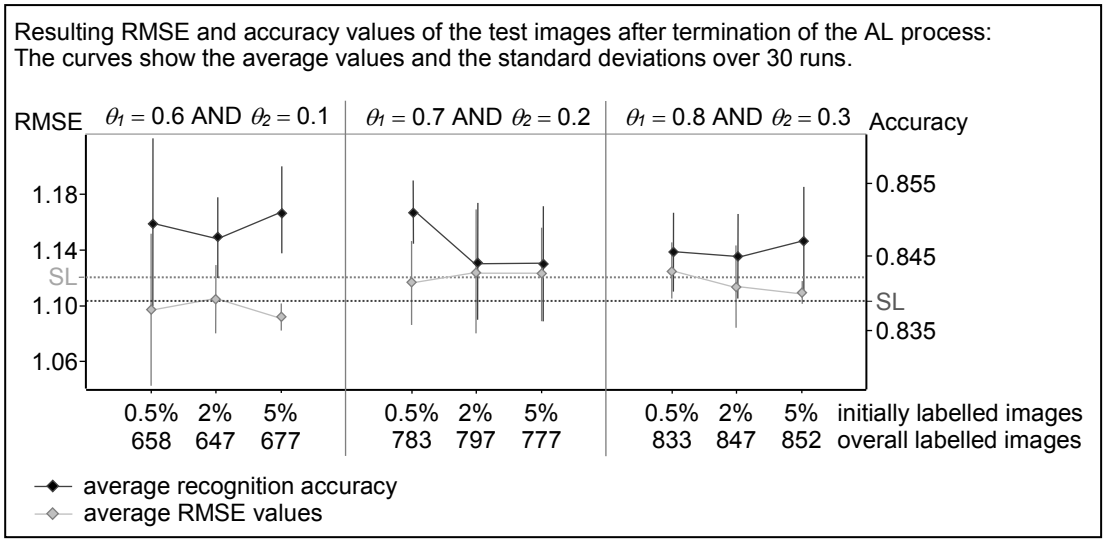


Figure 209: AL, PC: results for the distortion dataset, θ_1 AND θ_2, part I

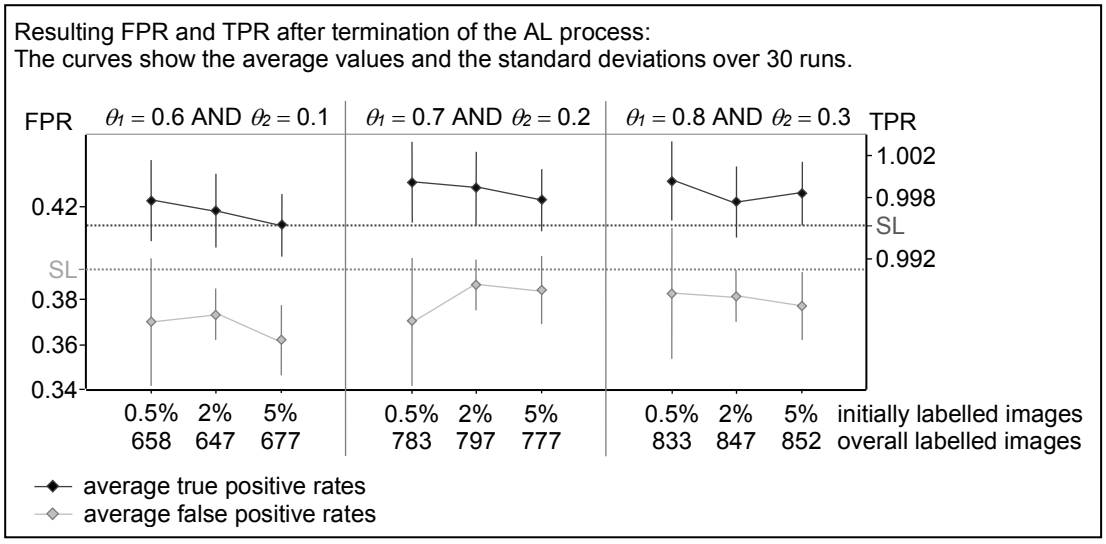


Figure 210: AL, PC: results for the distortion dataset, θ_1 AND θ_2, part II

A.26 AL, PC: learning curves for the distortion dataset

First, a training image is labelled by the Oracle if the maximum probability that the image belongs to the corresponding rating class is lower than the threshold θ_1. Exemplarily, the active learning procedure is carried out for 2% initially labelled training images and a threshold of 0.7, as shown in Figure 211.

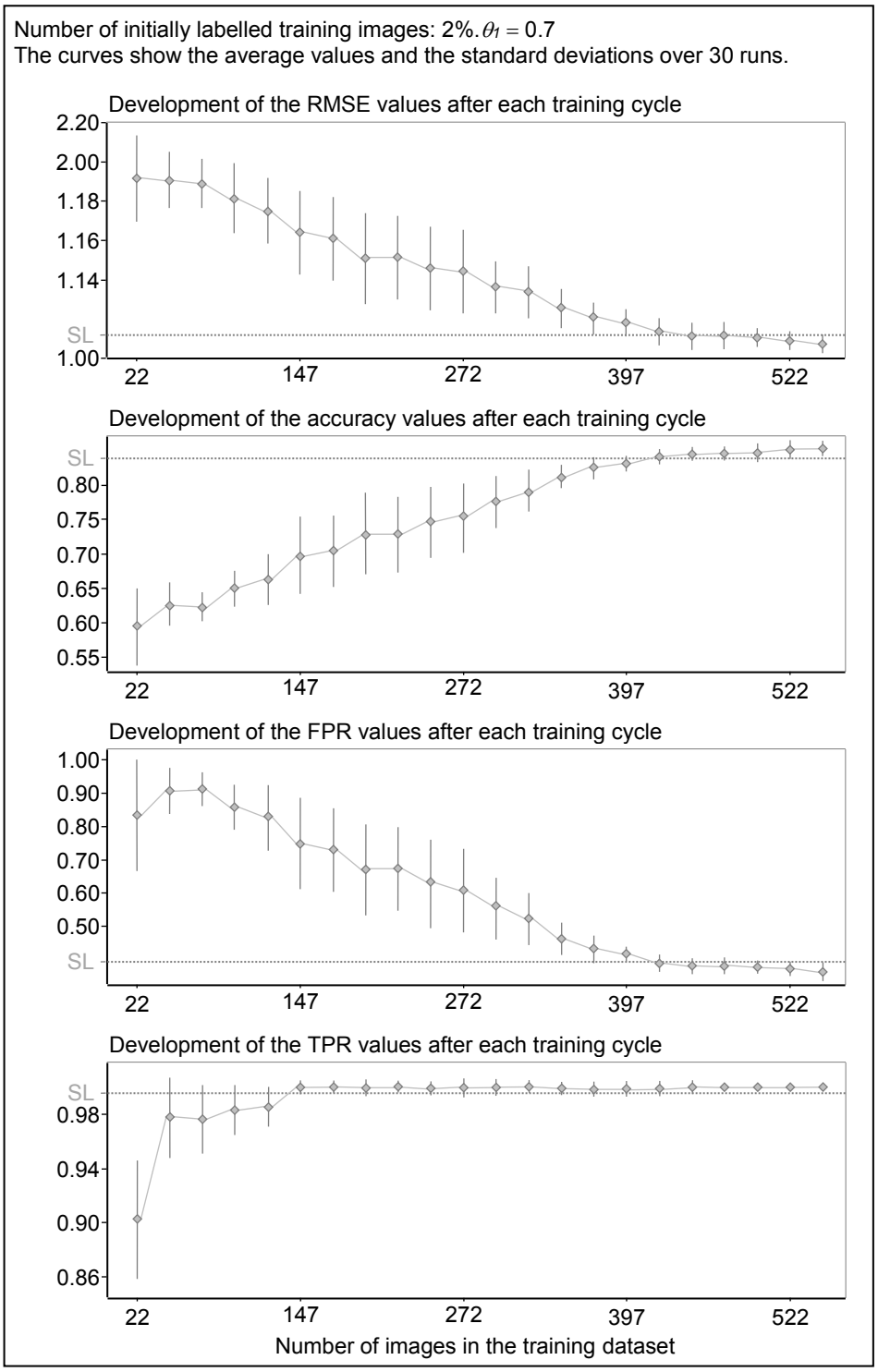


Figure 211: AL, PC: learning curves for the distortion dataset, θ_1

In the second step, a training image is labelled by the Oracle if the difference between the largest and the second largest class probability is lower than the threshold θ_2, which is set to 0.1. Exemplarily, the corresponding learning curves for 5% manually labelled images are shown in Figure 212.

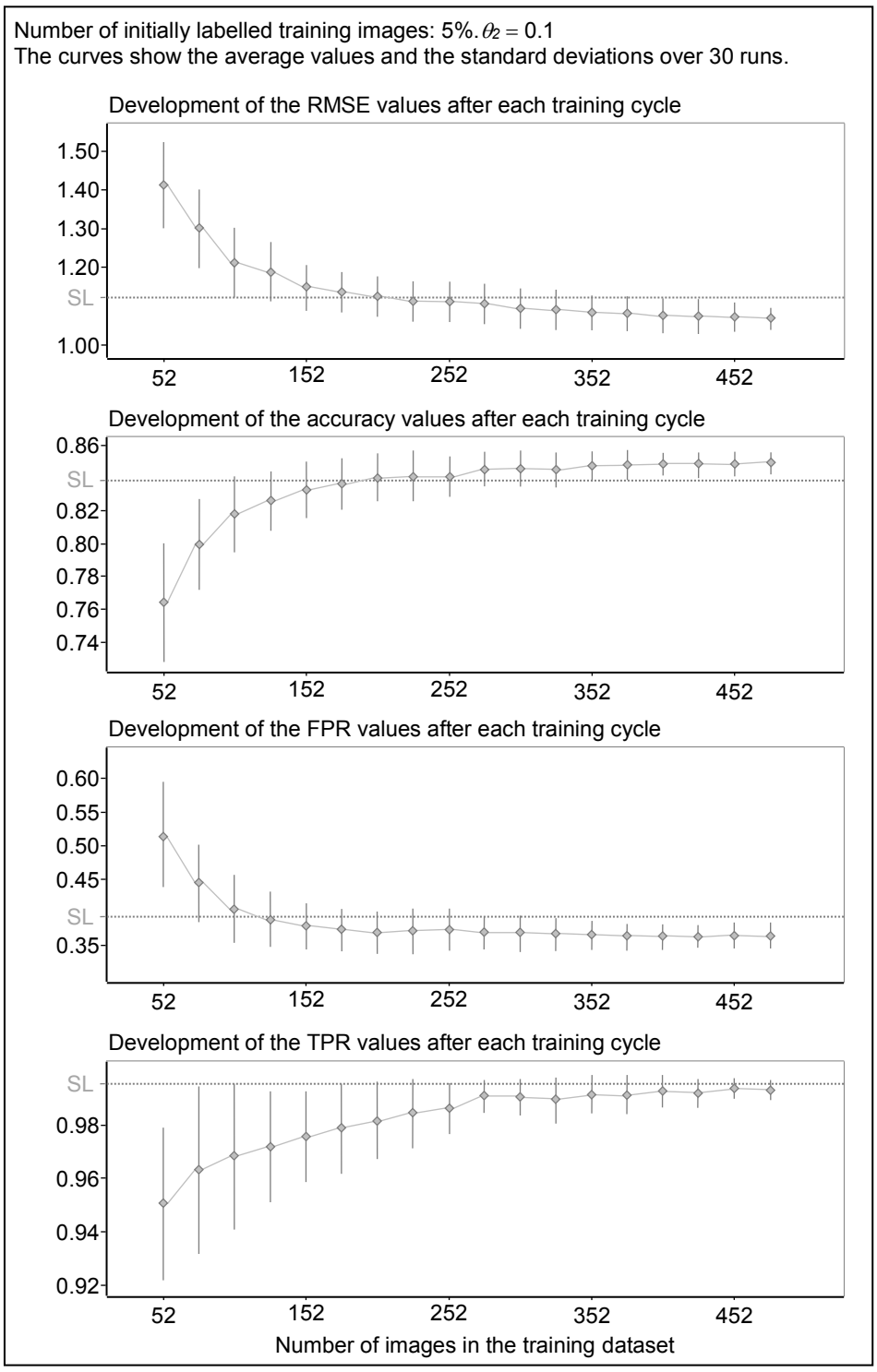


Figure 212: AL, PC: learning curves for the distortion dataset, θ_2

Finally, a training image is labelled by the Oracle if the maximum probability that the image belongs to the corresponding rating class is lower than the threshold θ_1 and if the difference between the largest and the second largest class probability is lower than the threshold θ_2. Exemplarily, the resulting learning curves for 2% manually labelled images and $\theta_1 = 0.6$ and $\theta_2 = 0.1$ are shown in Figure 213.

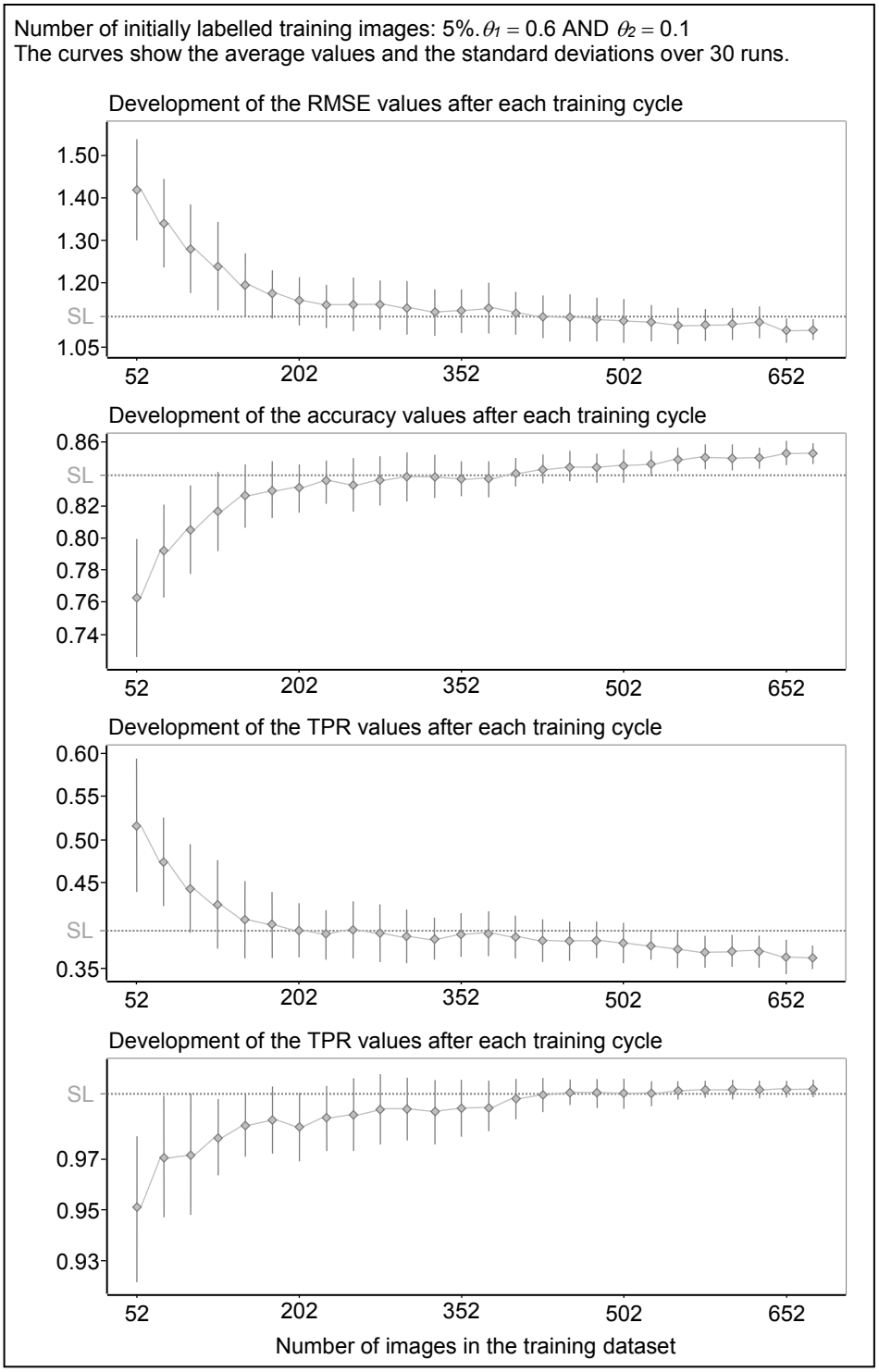


Figure 213: AL, PC: learning curves for the distortion dataset, θ_1 AND θ_2

A.27 AL, kNN: results for the distortion dataset

The active learning algorithm is able to query interactively for the correct label if the distance to the next training sample is greater than the threshold distance. The threshold distances are already determined in Table 33. The size of the initially labelled training set is set to 0.5%, 2%, and 5% of all training images from each rating class; see Table 41. The classification results after the termination of the AL process are summarised in Figure 214 and Figure 215. Since the selection of the labelled training data is done randomly, the average values and the standard deviations over 30 runs are calculated. The x-axes represent the number of initially labelled images in the training dataset and the y-axes the classification results.

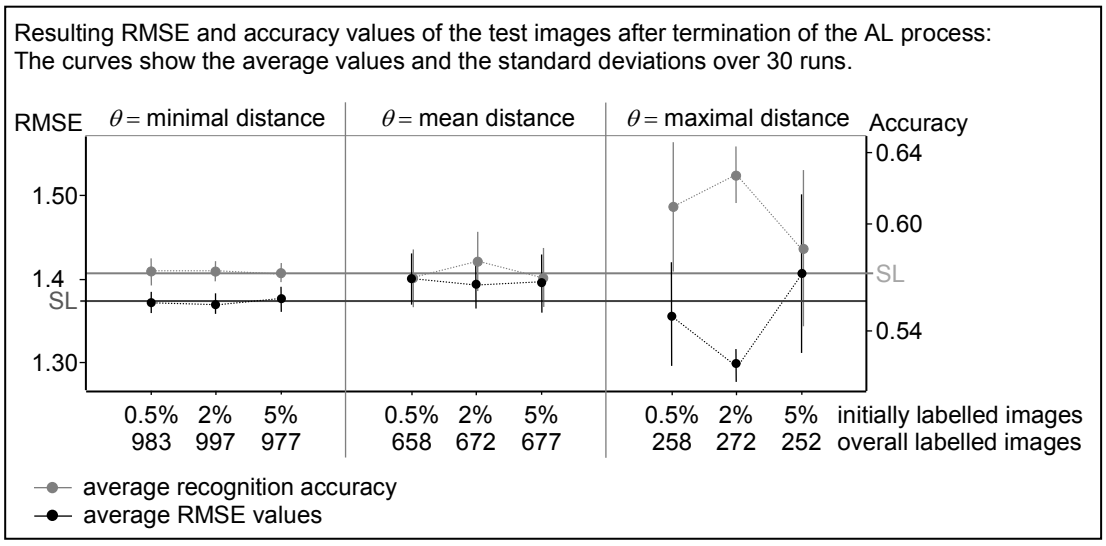


Figure 214: AL, kNN: results for the distortion dataset, part I

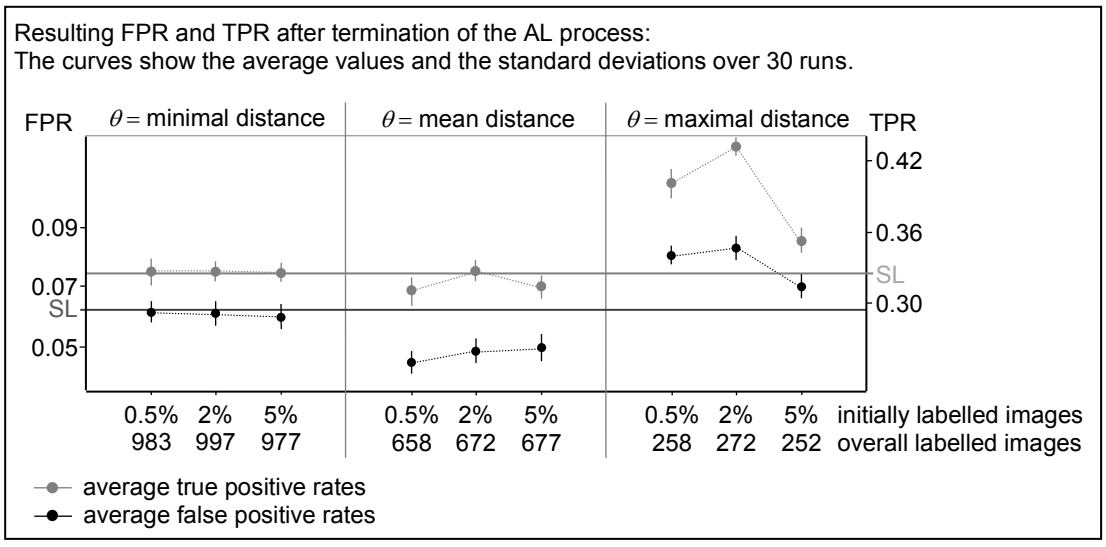


Figure 215: AL, kNN: results for the distortion dataset, part II

A.28 AL, kNN: learning curves for the distortion dataset

The learning curves for the distortion dataset for 2% manually labelled images are shown in Figure 216. If the distance to the next training sample is greater than the threshold distance, the class assignment is uncertain and the image is labelled by the Oracle and transferred into the training dataset.

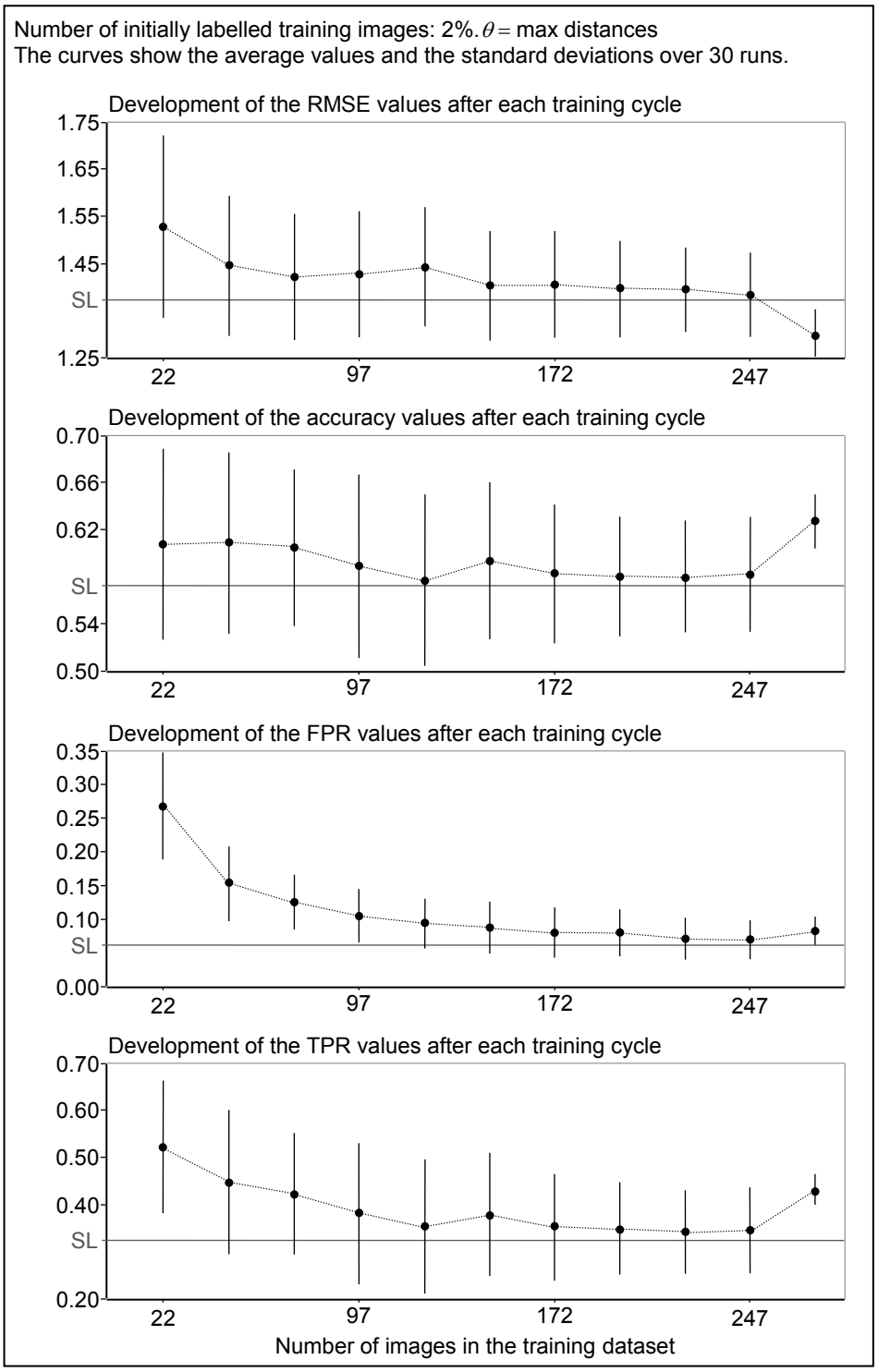


Figure 216: AL, kNN: learning curves for the distortion dataset

A.29 AL, LVQ: results for the distortion dataset

The active learning algorithm is able to query interactively for the correct label if the distance to the nearest prototype is greater than the threshold distance. The threshold distances are already determined in Table 34. The size of the initially labelled training set is set to 0.5%, 2%, and 5% of all training images from each rating class; see Table 41. The classification results after the termination of the AL process are summarised in Figure 217 and Figure 218. Since the selection of the labelled training data is done randomly, the average values and the standard deviations over 30 runs are calculated. The x-axes represent the number of initially labelled images in the training dataset and the y-axes the classification results.

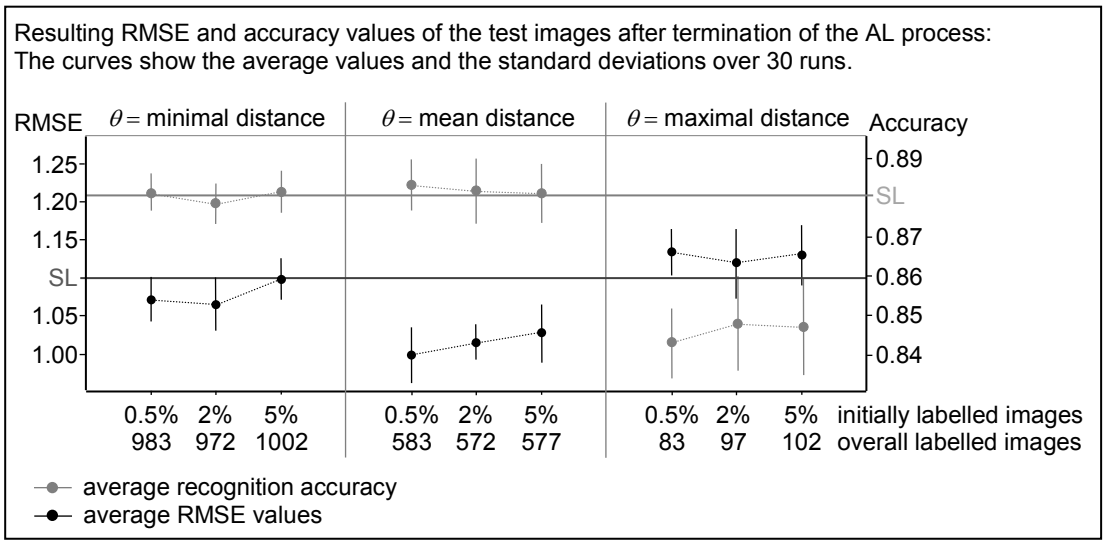


Figure 217: AL, LVQ: results for the distortion dataset, part I

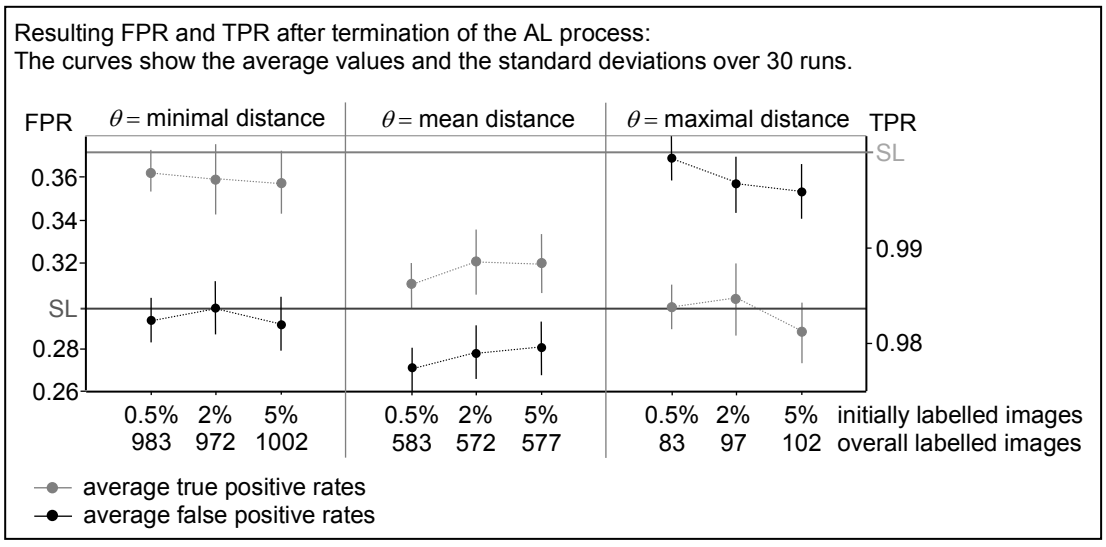


Figure 218: AL, LVQ: results for the distortion dataset, part II

A.30 AL, LVQ: learning curves for the distortion dataset

The learning curves for the distortion dataset for 0.5% manually labelled images are shown in Figure 219. An image is labelled by the Oracle if the distance to the next training sample is greater than the threshold distance.

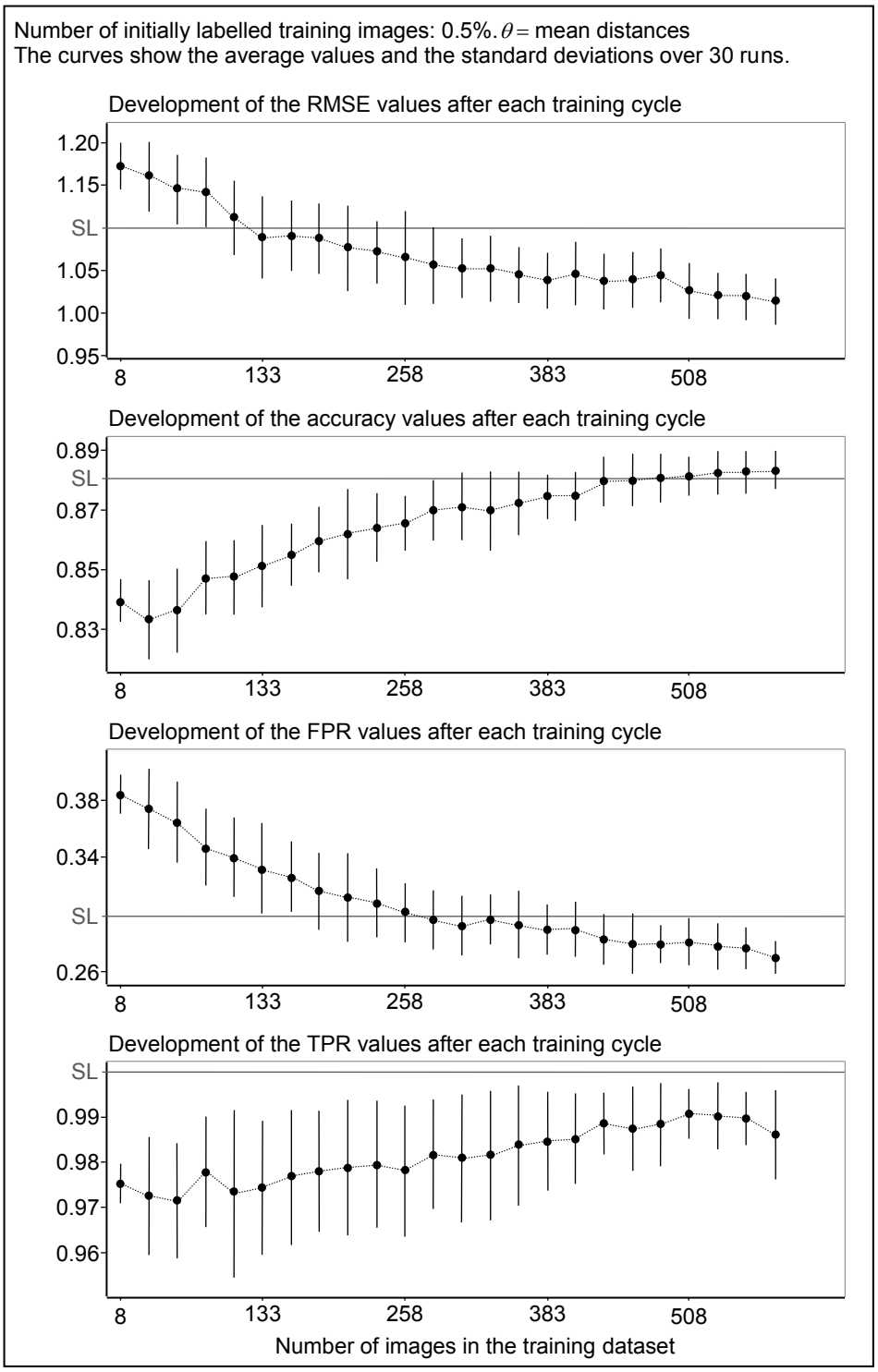


Figure 219: AL, LVQ: learning curves for the distortion dataset

A.31 AL, PC: results for the double image dataset

First, a training image is labelled by the Oracle if the maximum probability that the image belongs to the corresponding rating class is lower than the threshold θ_1. Here, θ_1 is set to 0.6, 0.7, and 0.8. The size of the initial training set is to 0.5%, 2%, and 5% of all training images from each rating class; see Table 42. The classification results after the termination of the AL process are summarised in Figure 220 and Figure 221. Since the selection of the labelled training data is done randomly, the average values and the standard deviations over 30 runs are calculated. The x-axes represent the number of initially labelled images in the training dataset and the y-axes the classification results.

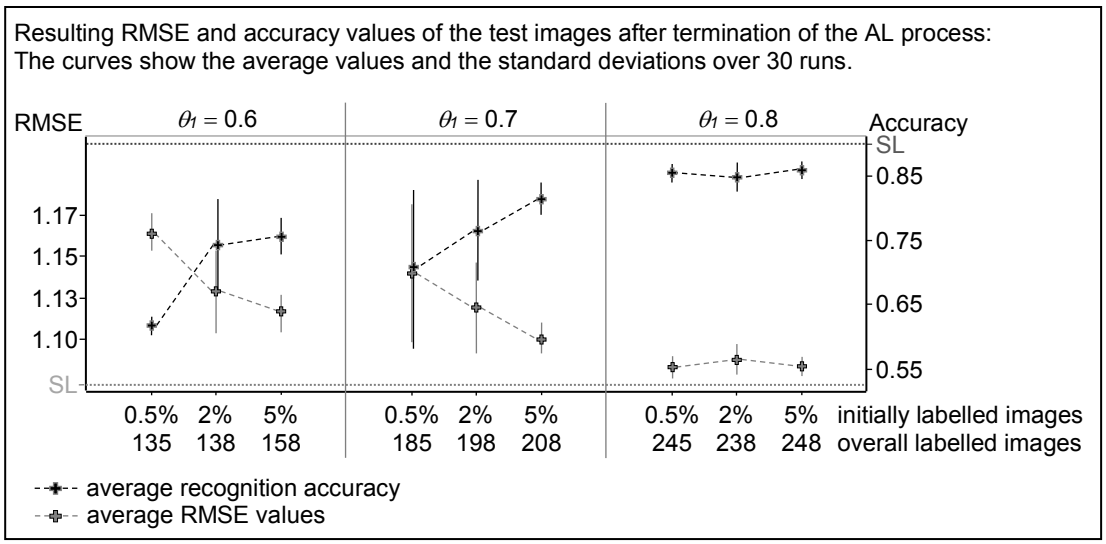


Figure 220: AL, PC: results for the double image dataset, θ_1, part I

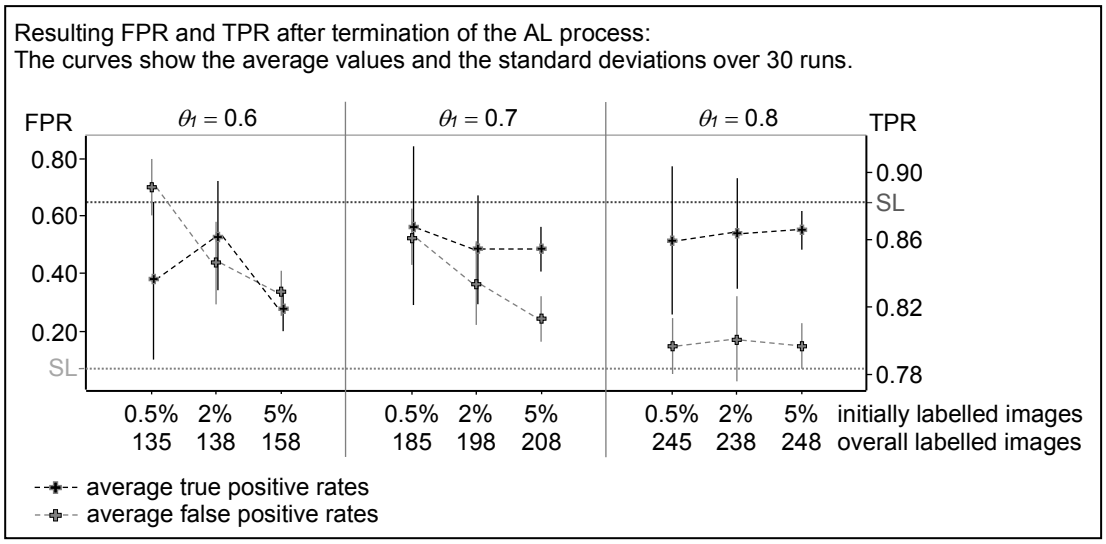


Figure 221: AL, PC: results for the double image dataset, θ_1, part II

In the second step, the Oracle is asked for the right label if the difference between the largest and the second largest class probability is less than a threshold θ_2, which is set to 0.1, 0.2, and 0.3. The size of the initial training set is to 0.5%, 2%, and 5% of all training images from each rating class; see Table 42. The classification results after the termination of the AL process are summarised in Figure 222 and Figure 223. Since the selection of the labelled training data is done randomly, the average values and the standard deviations over 30 runs are calculated. The x-axes represent the number of initially labelled images in the training dataset and the y-axes the classification results.

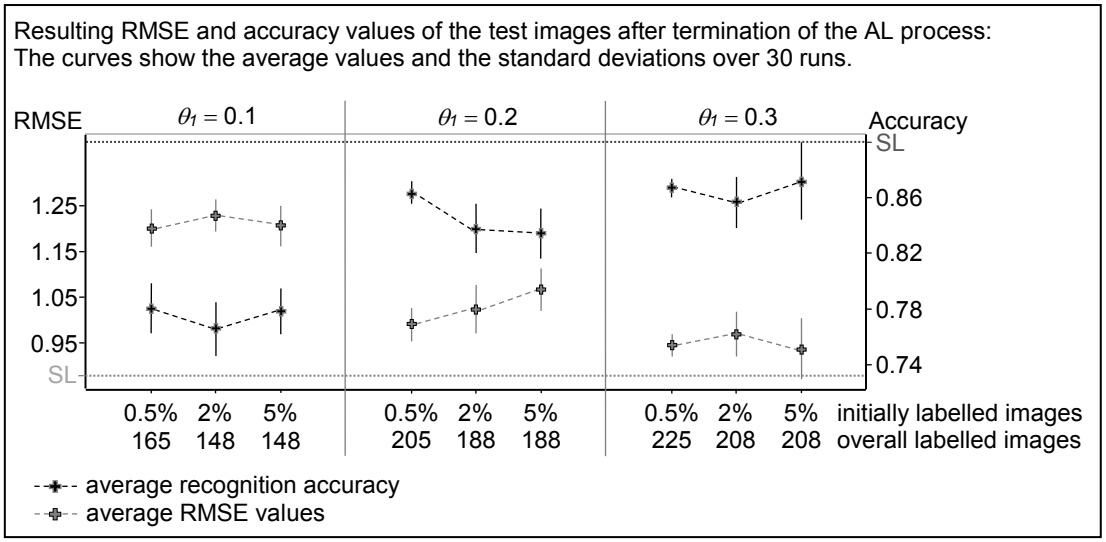


Figure 222: AL, PC: results for the double image dataset, θ_2, part I

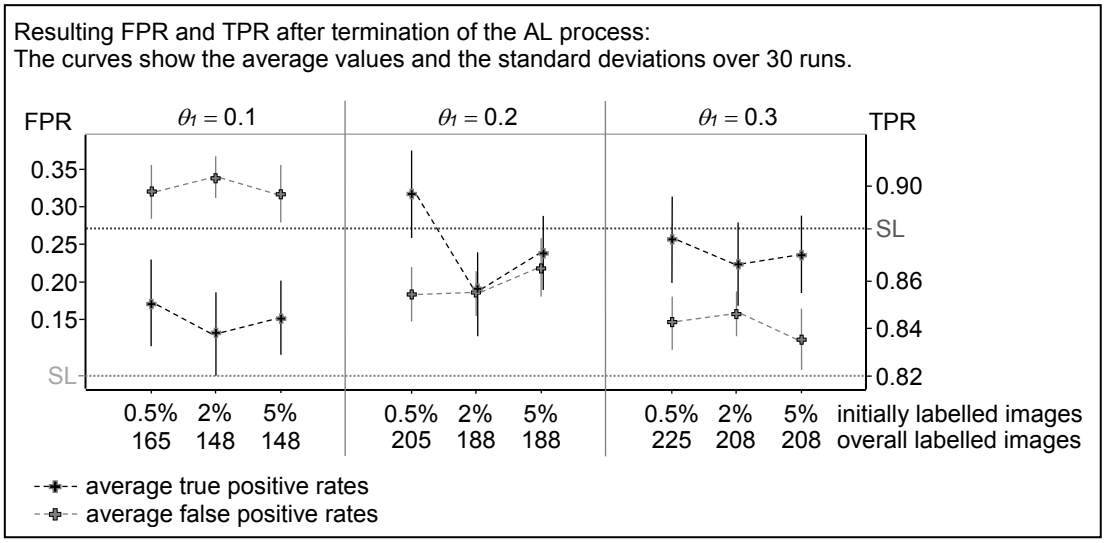


Figure 223: AL, PC: results for the double image dataset, θ_2, part II

Finally, the 2 selection criteria, which determine if the training image needs to the labelled by the Oracle, are combined. A training image needs to be labelled manually if the maximum probability that the image belongs to the corresponding rating class is lower than the threshold θ_1 and if the difference between the largest and the second largest class probability is less than a threshold θ_2. The size of the initial training set is to 0.5%, 2%, and 5% of all training images from each rating class; see Table 42. The classification results after the termination of the AL process are summarised in Figure 224 and Figure 225. Since the selection of the labelled training data is done randomly, the average values and the standard deviations over 30 runs are calculated. The x-axes represent the number of initially labelled images in the training dataset and the y-axes the classification results.

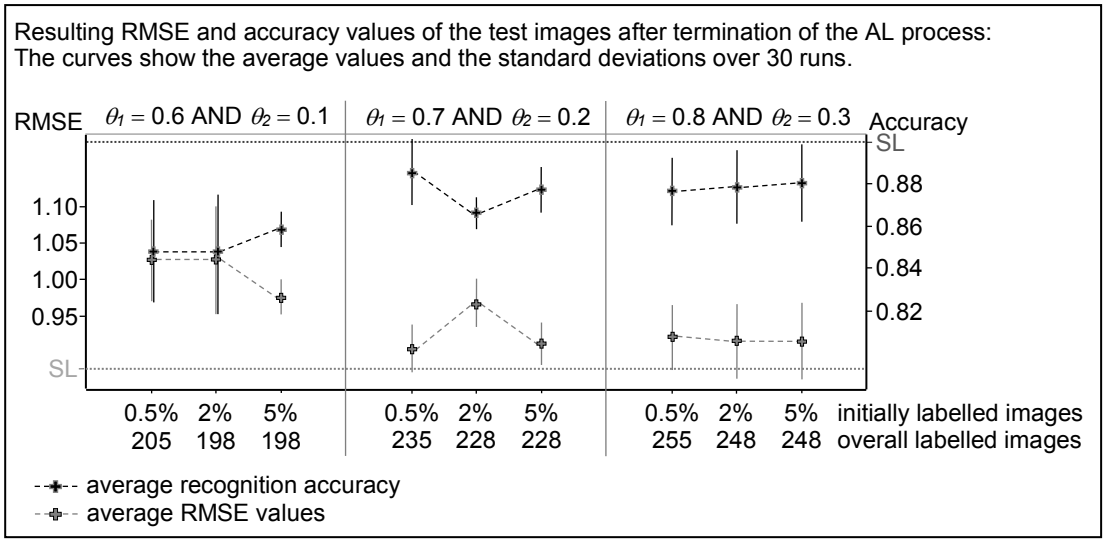


Figure 224: AL, PC: results for the double image dataset, θ_1 AND θ_2, part I

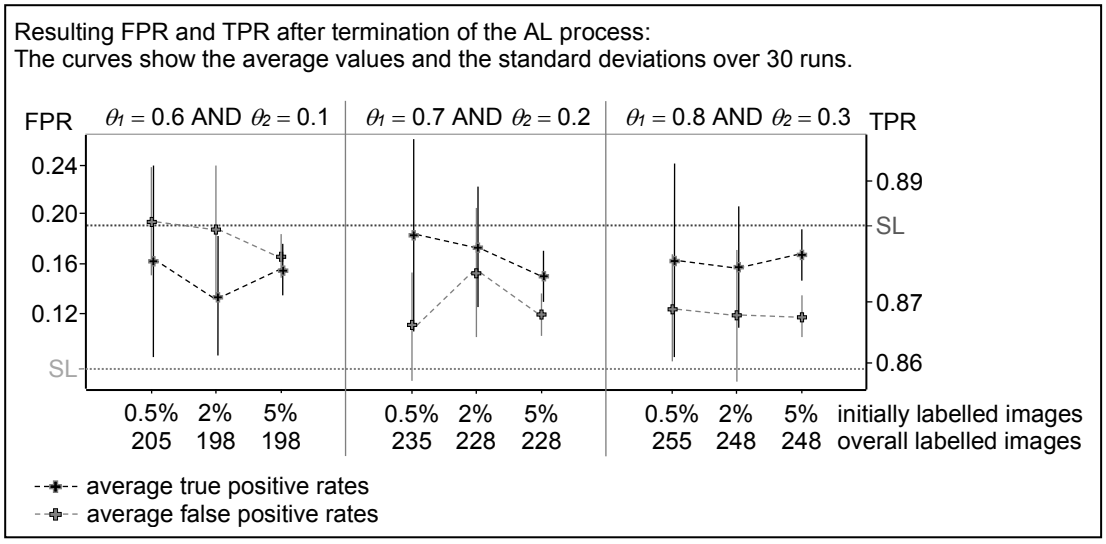


Figure 225: AL, PC: results for the double image dataset, θ_1 AND θ_2, part II

A.32 AL, PC: learning curves for the double image dataset

First, a training image is labelled by the Oracle if the maximum probability that the image belongs to the corresponding rating class is lower than the threshold θ_1. Exemplarily, the active learning procedure is carried out for 5% initially labelled training images and a threshold of 0.8, as shown in Figure 226.

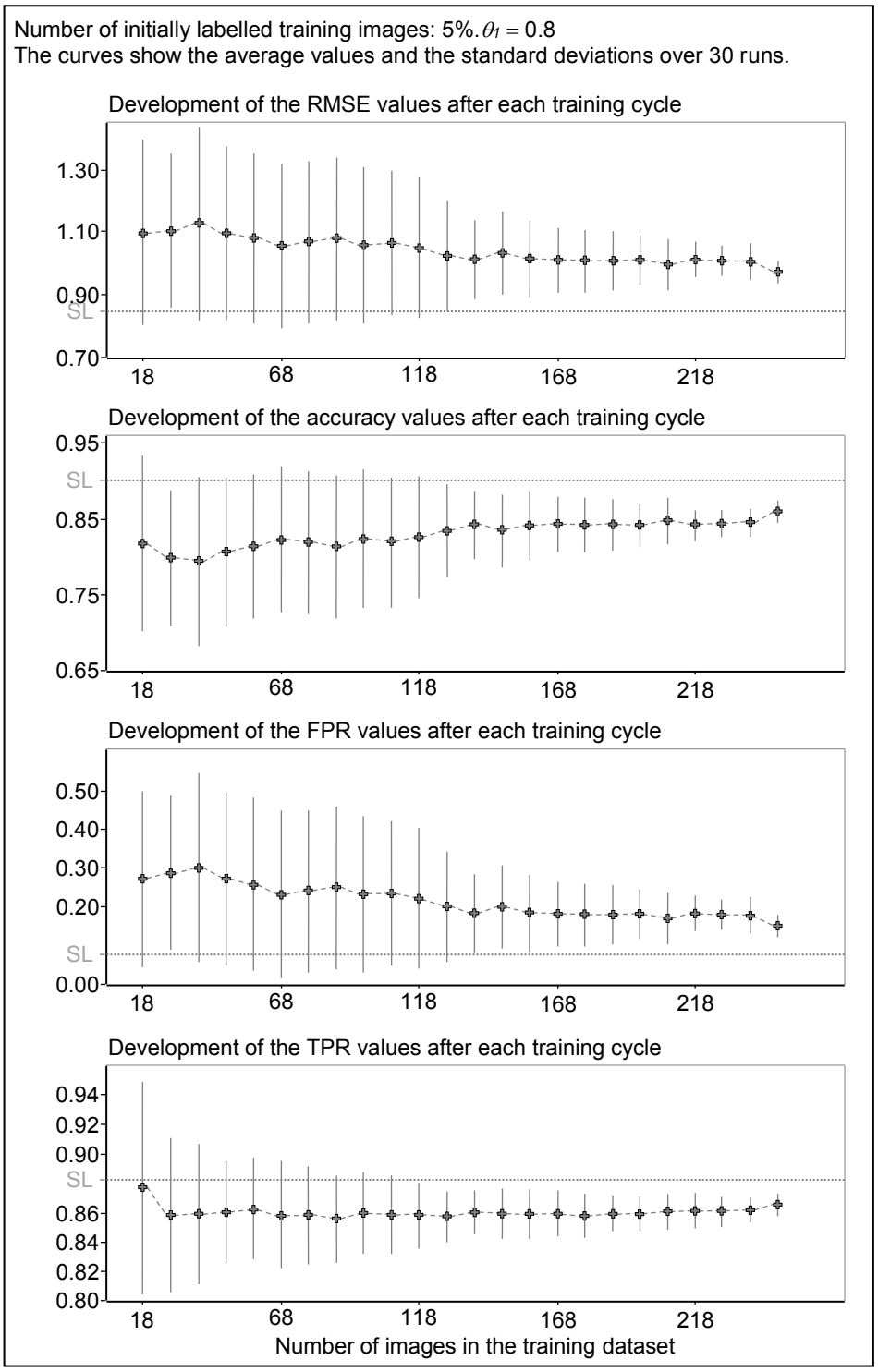


Figure 226: AL, PC: learning curves for the double image dataset, θ_1

In the second step, a training image is labelled by the Oracle if the difference between the largest and the second largest class probability is lower than the threshold θ_2, which is set to 0.3. The corresponding learning curves for 5% manually labelled images are shown in Figure 227.

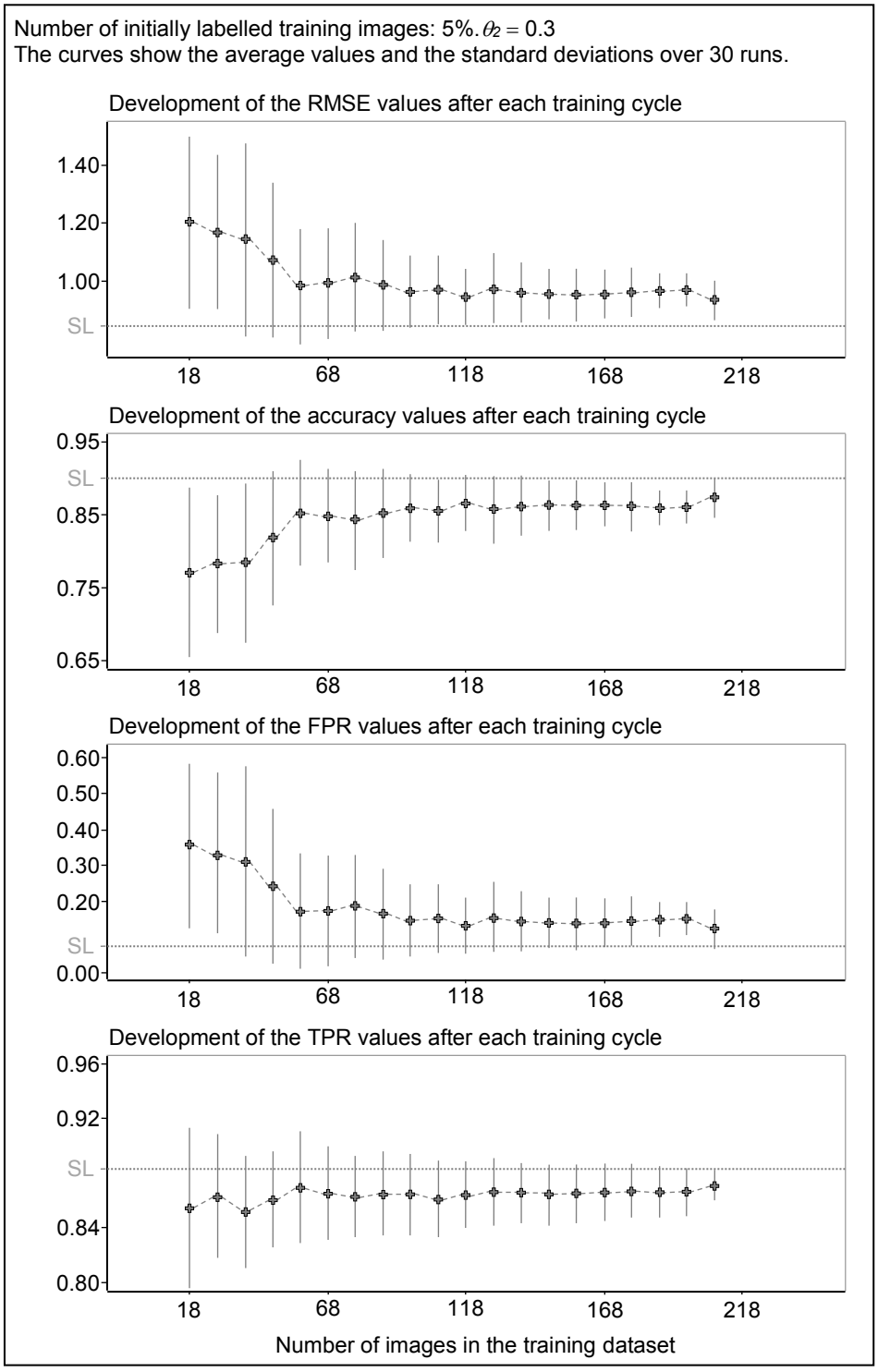


Figure 227: AL, PC: learning curves for the double image dataset, θ_2

Finally, a training image is labelled by the Oracle if the maximum probability that the image belongs to the corresponding rating class is lower than the threshold θ_1 and if the difference between the largest and the second largest class probability is lower than the threshold θ_2. The resulting learning curves for 0.5% manually labelled images and $\theta_1 = 0.7$ and $\theta_2 = 0.2$ are shown in Figure 228.

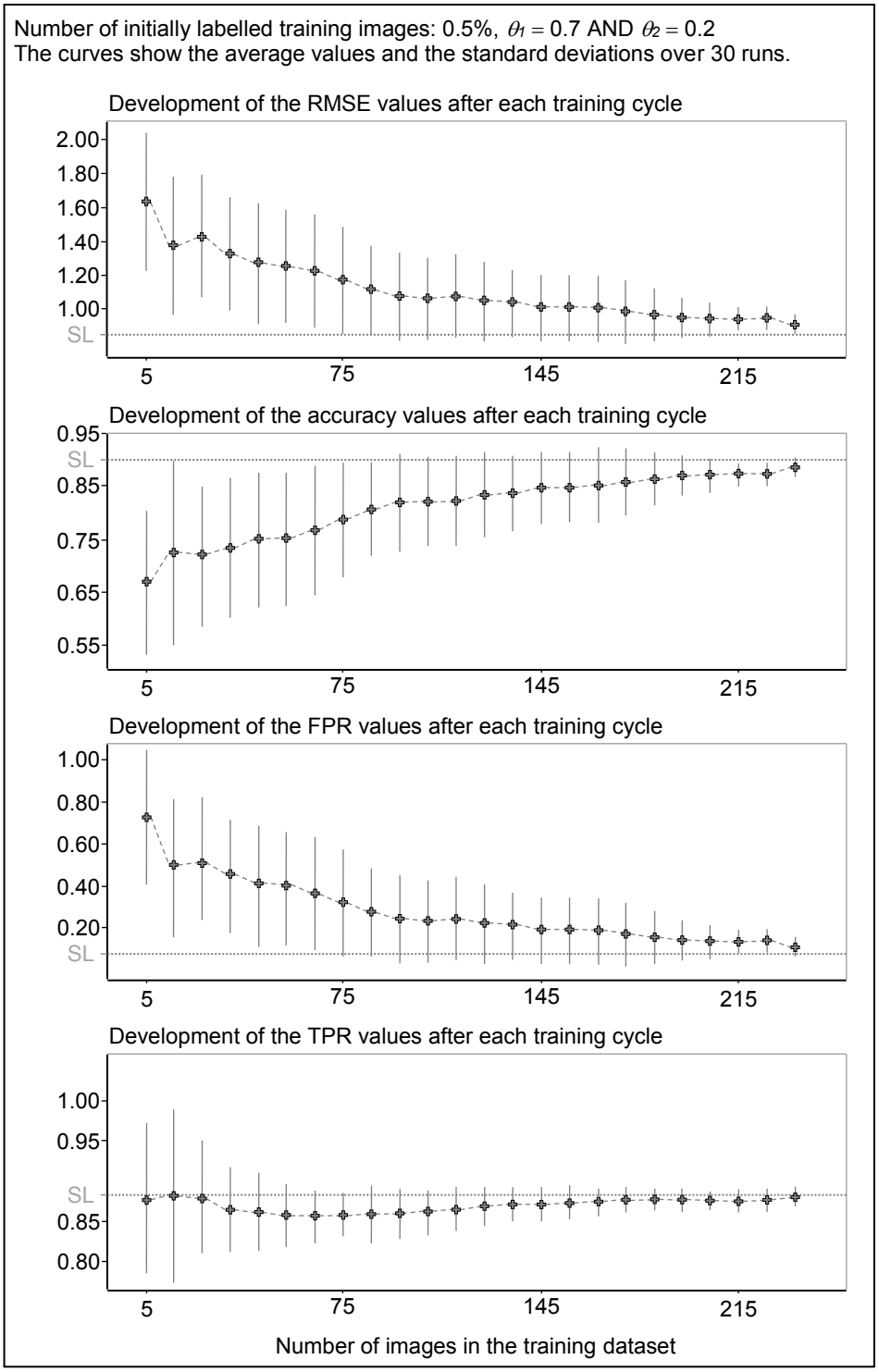


Figure 228: AL, PC: learning curves for the double image dataset, θ_1 AND θ_2

A.33 AL, kNN: results for the double image dataset

The active learning algorithm is able to query interactively for the correct label if the distance to the next training sample is greater than the threshold distance. The threshold distances are already determined in Table 36. The size of the initially labelled training set is set to 0.5%, 2%, and 5% of all training images from each rating class; see Table 42. The classification results after the termination of the AL process are summarised in Figure 229 and Figure 230. Since the selection of the labelled training data is done randomly, the average values and the standard deviations over 30 runs are calculated. The x-axes represent the number of initially labelled images in the training dataset and the y-axes the classification results.

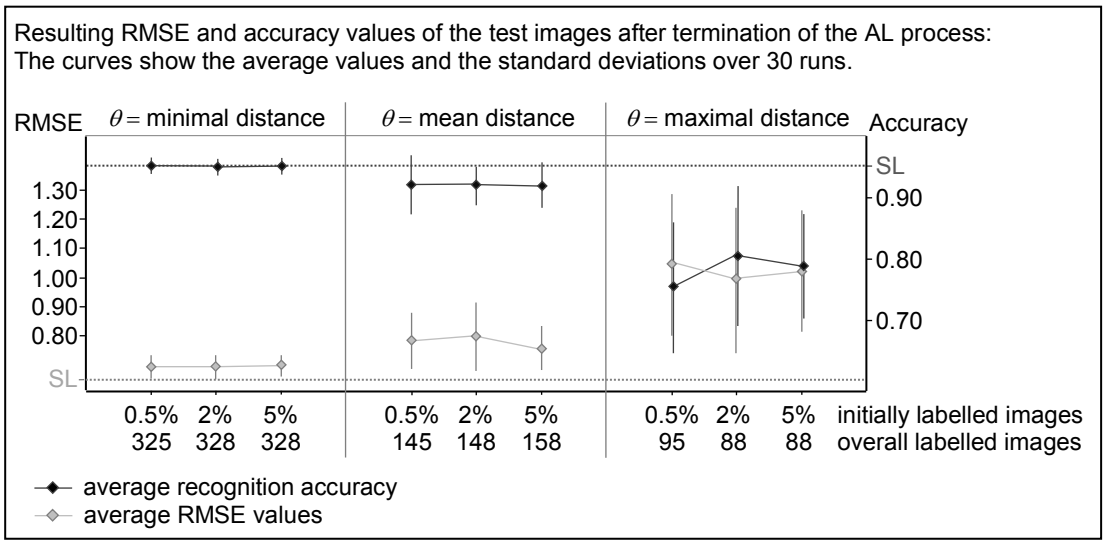


Figure 229: AL, kNN: results for the double image dataset, part I

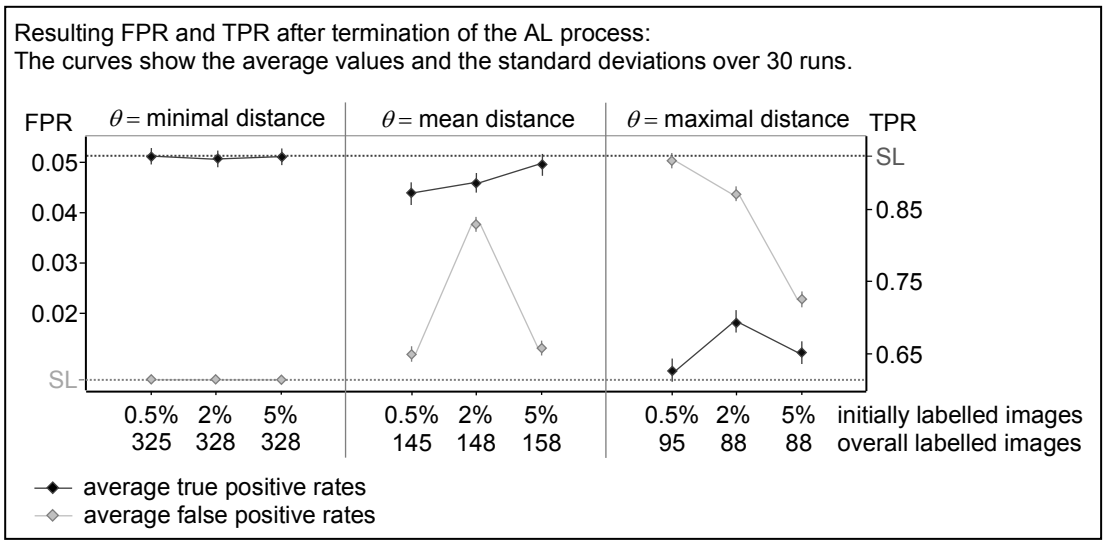


Figure 230: AL, kNN: results for the double image dataset, part II

A.34 AL, kNN: learning curves for the double image dataset

The learning curves for the distortion dataset for 5% manually labelled images are shown in Figure 231. If the distance to the next training sample is greater than the threshold distance, the class assignment is uncertain and the image is labelled by the Oracle and transferred into the training dataset.

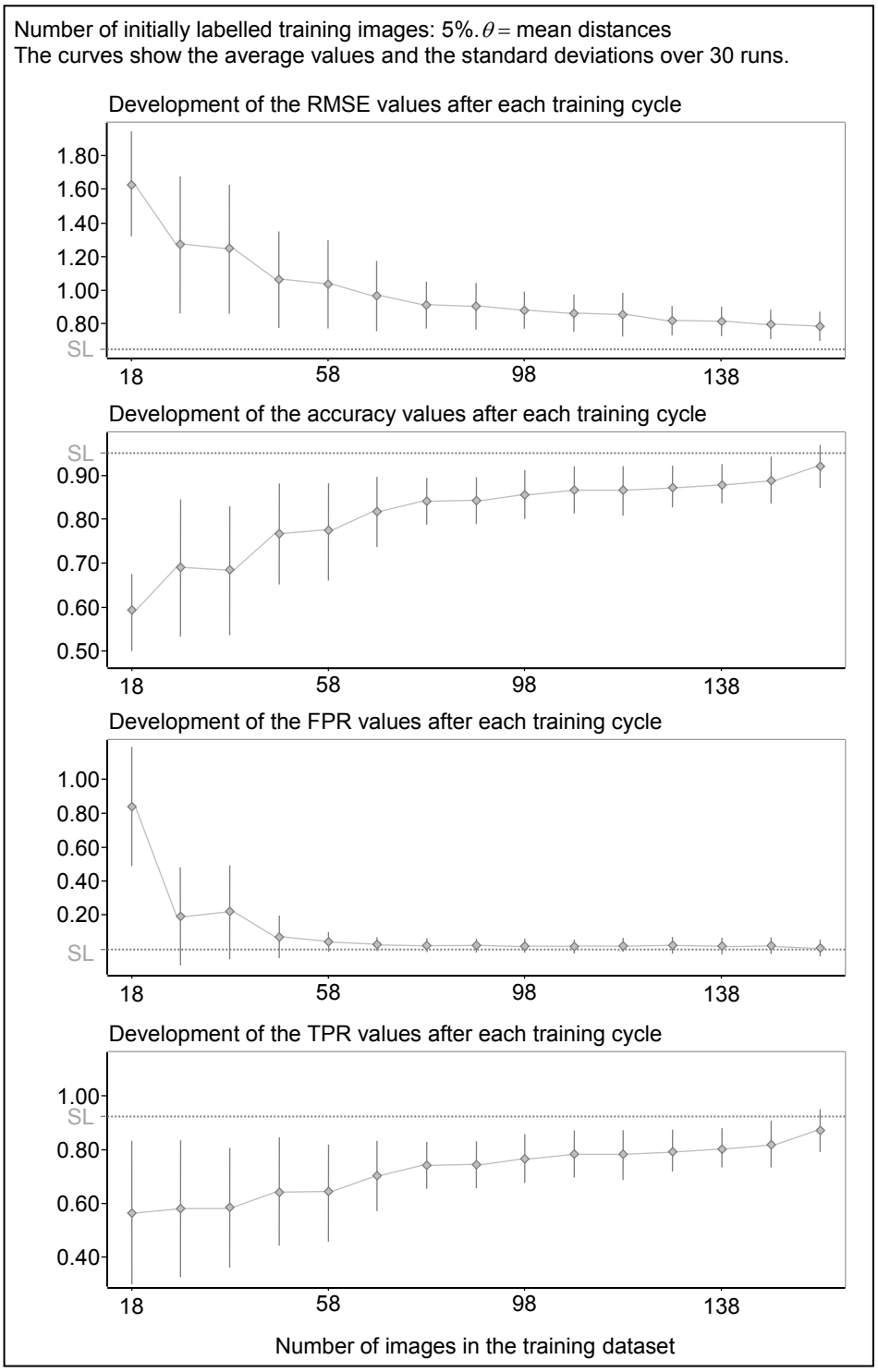


Figure 231: AL, kNN: learning curves for the double image dataset

A.35 AL, LVQ: results for the double image dataset

The active learning algorithm is able to query interactively for the correct label if the distance to the nearest prototype is greater than the threshold distance. The threshold distances are already determined in Table 37. The size of the initially labelled training set is set to 0.5%, 2%, and 5% of all training images from each rating class see; Table 42. The classification results after the termination of the AL process are summarised in Figure 232 and Figure 233. Since the selection of the labelled training data is done randomly, the average values and the standard deviations over 30 runs are calculated. The x-axes represent the number of initially labelled images in the training dataset and the y-axes the classification results.

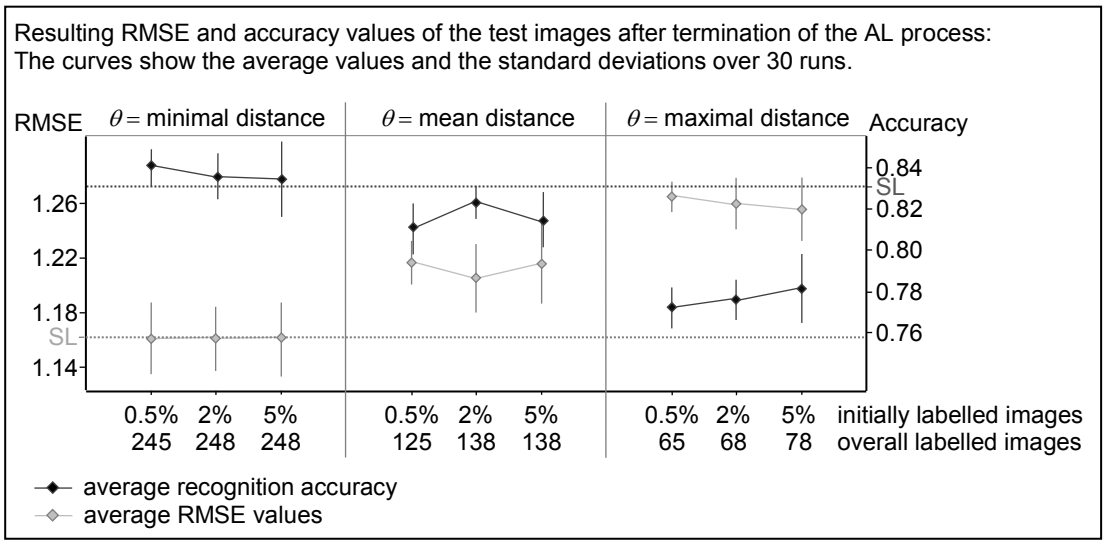


Figure 232: AL, LVQ: results for the double image dataset, part I

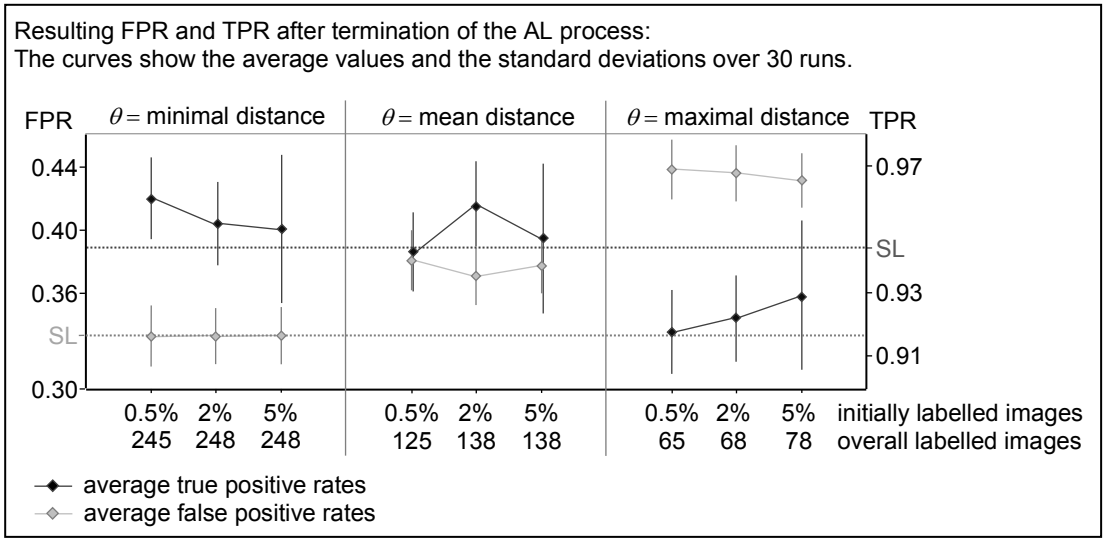


Figure 233: AL, LVQ: results for the double image dataset, part II

A.36 AL, LVQ: learning curves for the double image dataset

The learning curves for the distortion dataset for 0.5% manually labelled images are shown in Figure 234. An image is labelled by the Oracle if the distance to the next training sample is greater than the threshold distance.

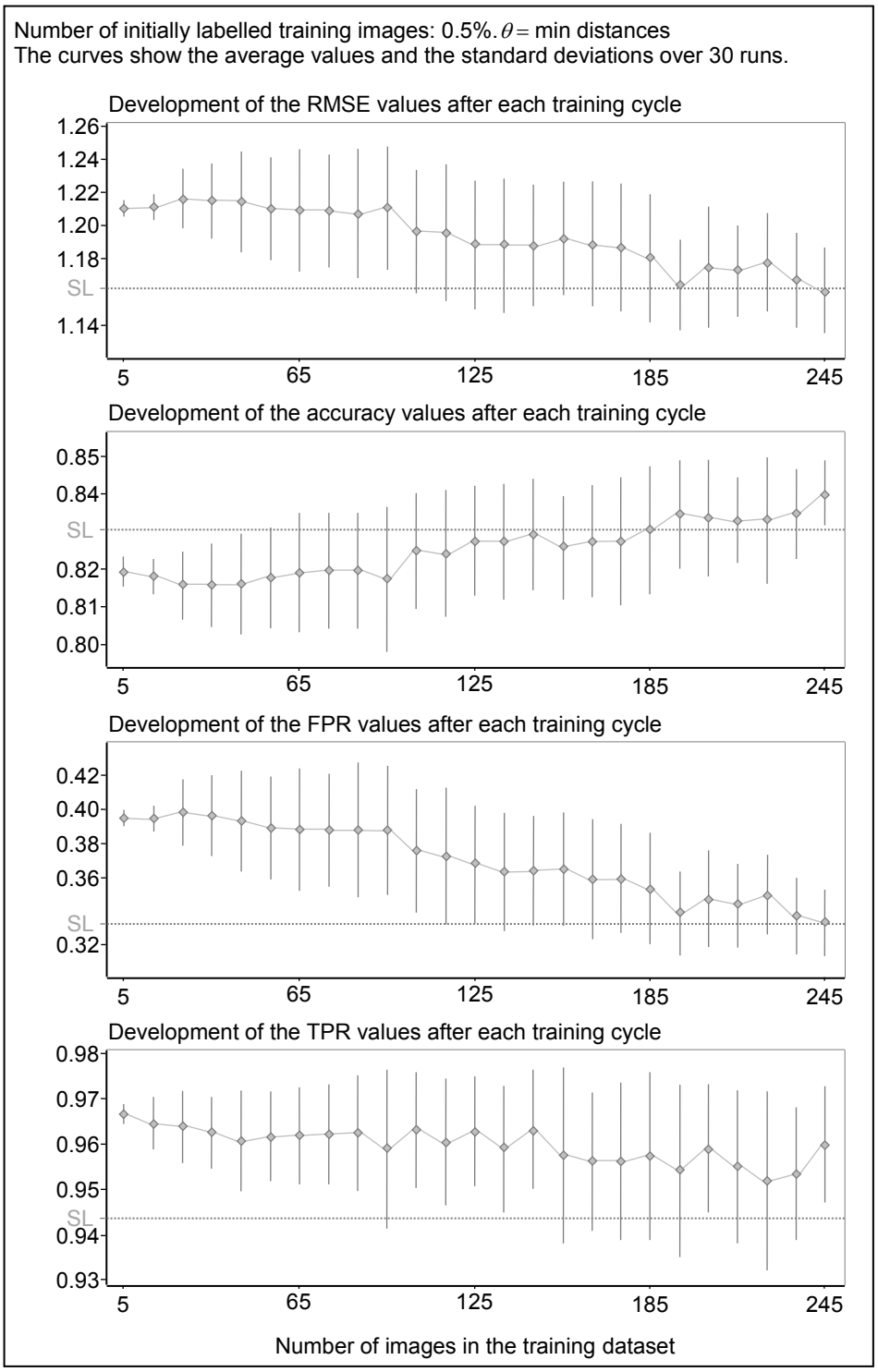


Figure 234: AL, LVQ: learning curves for the double image dataset

A.37 AL, PC: results for the distortion and double image dataset

First, a training image is labelled by the Oracle if the maximum probability that the image belongs to the corresponding rating class is lower than the threshold θ_1. Here, θ_1 is set to 0.6, 0.7, and 0.8. The size of the initial training set is to 0.5%, 2%, and 5% of all training images from each rating class; see Table 43. The classification results after the termination of the AL process are summarised in Figure 235 and Figure 236. Since the selection of the labelled training data is done randomly, the average values and the standard deviations over 30 runs are calculated. The x-axes represent the number of initially labelled images in the training dataset and the y-axes the classification results.

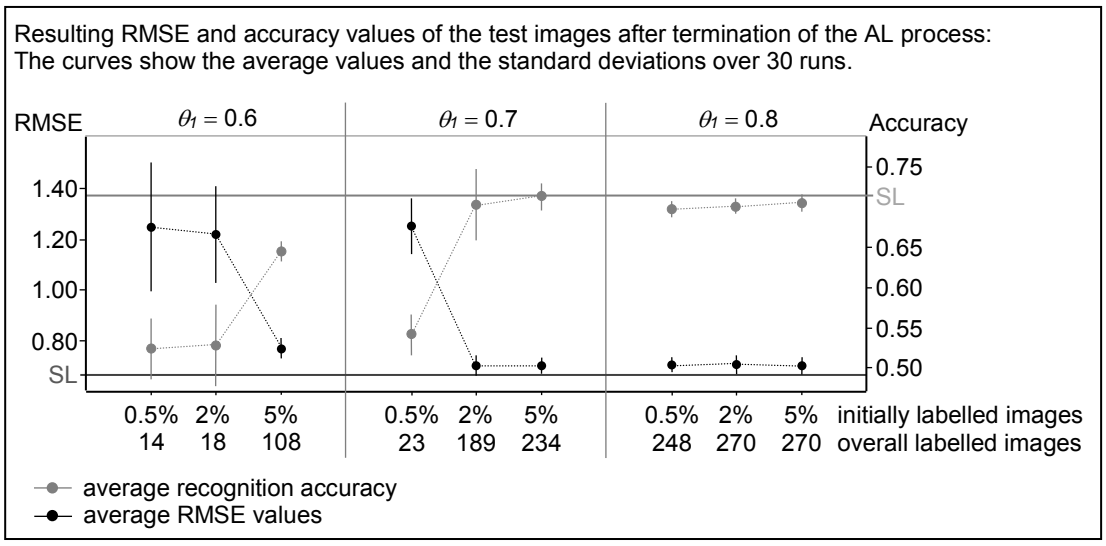


Figure 235: AL, PC: results for the distortion and double image dataset θ_1, part I

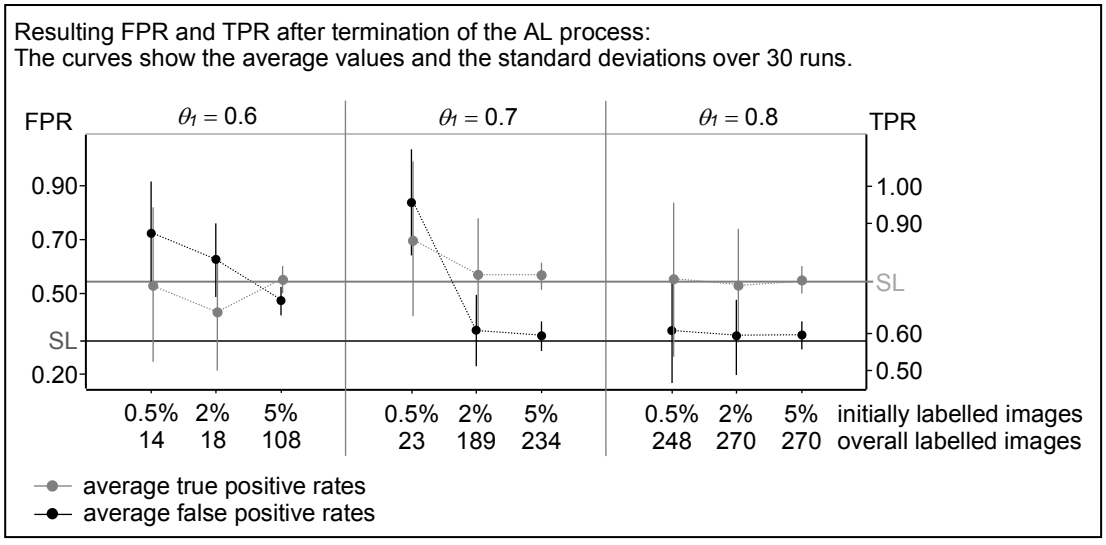


Figure 236: AL, PC: results for the distortion and double image dataset θ_1, part II

In the second step, the Oracle is asked for the right label if the difference between the largest and the second largest class probability is less than a threshold θ_2, which is set to 0.1, 0.2, and 0.3. The size of the initial training set is to 0.5%, 2%, and 5% of all training images from each rating class; see Table 43. The classification results after the termination of the AL process are summarised in Figure 237 and Figure 238. Since the selection of the labelled training data is done randomly, the average values and the standard deviations over 30 runs are calculated. The x-axes represent the number of initially labelled images in the training dataset and the y-axes the classification results.

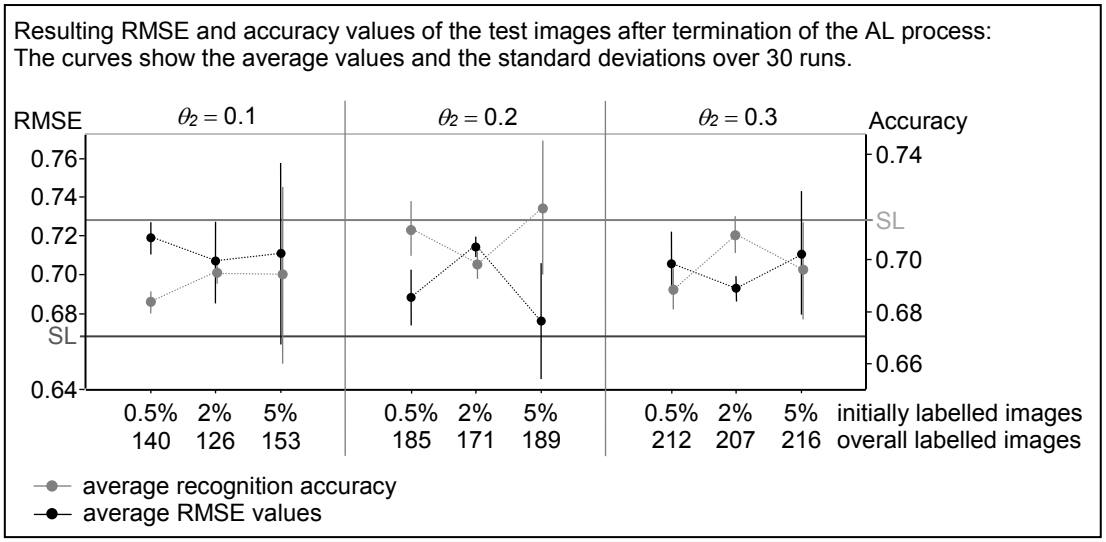


Figure 237: AL, PC: results for the distortion and double image dataset θ_2, part I

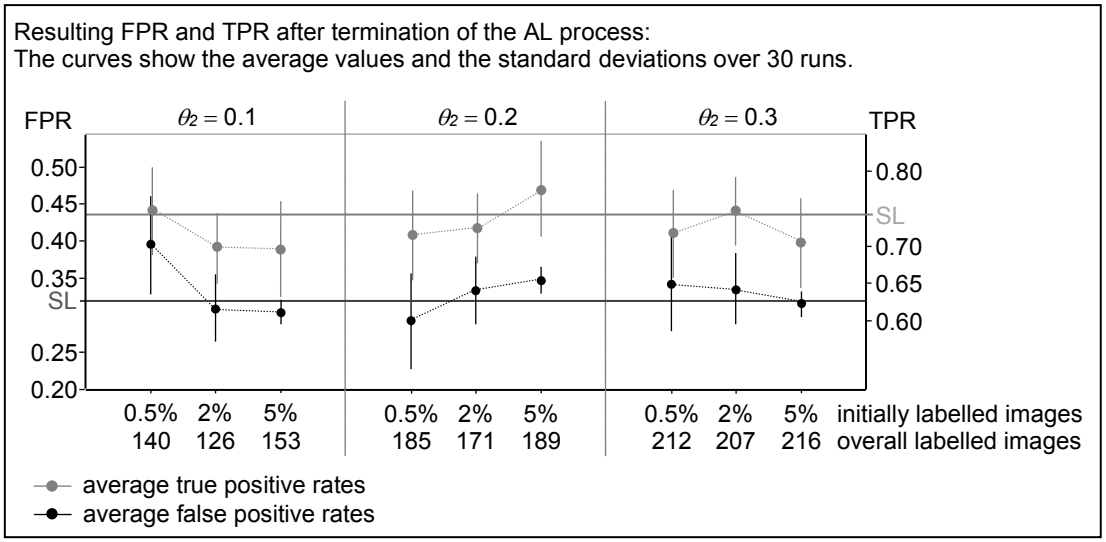


Figure 238: AL, PC: results for the distortion and double image dataset θ_2, part II

Finally, the 2 selection criteria, which determine if the training image needs to the labelled by the Oracle, are combined. A training image needs to be labelled manually if the maximum probability that the image belongs to the corresponding rating class is lower than the threshold θ_1 and if the difference between the largest and the second largest class probability is less than a threshold θ_2. The size of the initial training set is to 0.5%, 2%, and 5% of all training images from each rating class; see Table 43. The classification results after the termination of the AL process are summarised in Figure 239 and Figure 240. Since the selection of the labelled training data is done randomly, the average values and the standard deviations over 30 runs are calculated. The x-axes represent the number of initially labelled images in the training dataset and the y-axes the classification results.

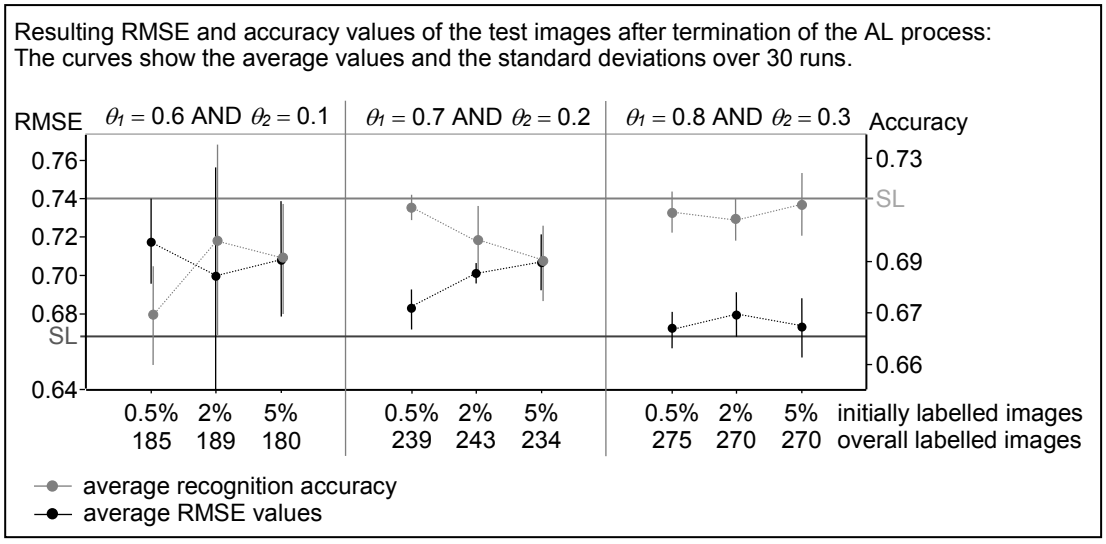


Figure 239: AL, PC: results for the distortion and double image dataset θ_1 AND θ_2, part I

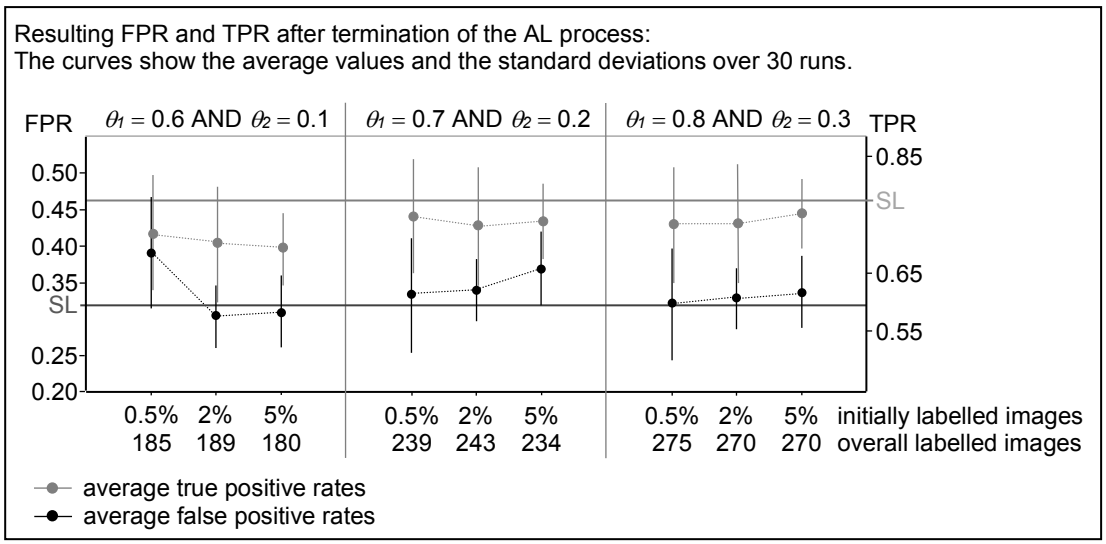


Figure 240: AL, PC: results for the distortion and double image dataset θ_1 AND θ_2, part II

A.38 AL, PC: learning curves for the distortion and double image dataset

A training image is labelled by the Oracle if the maximum probability that the image belongs to the corresponding rating class is lower than the threshold θ_1, which is set to 0.7. The learning curves for 2% manually labelled images are shown in Figure 241.

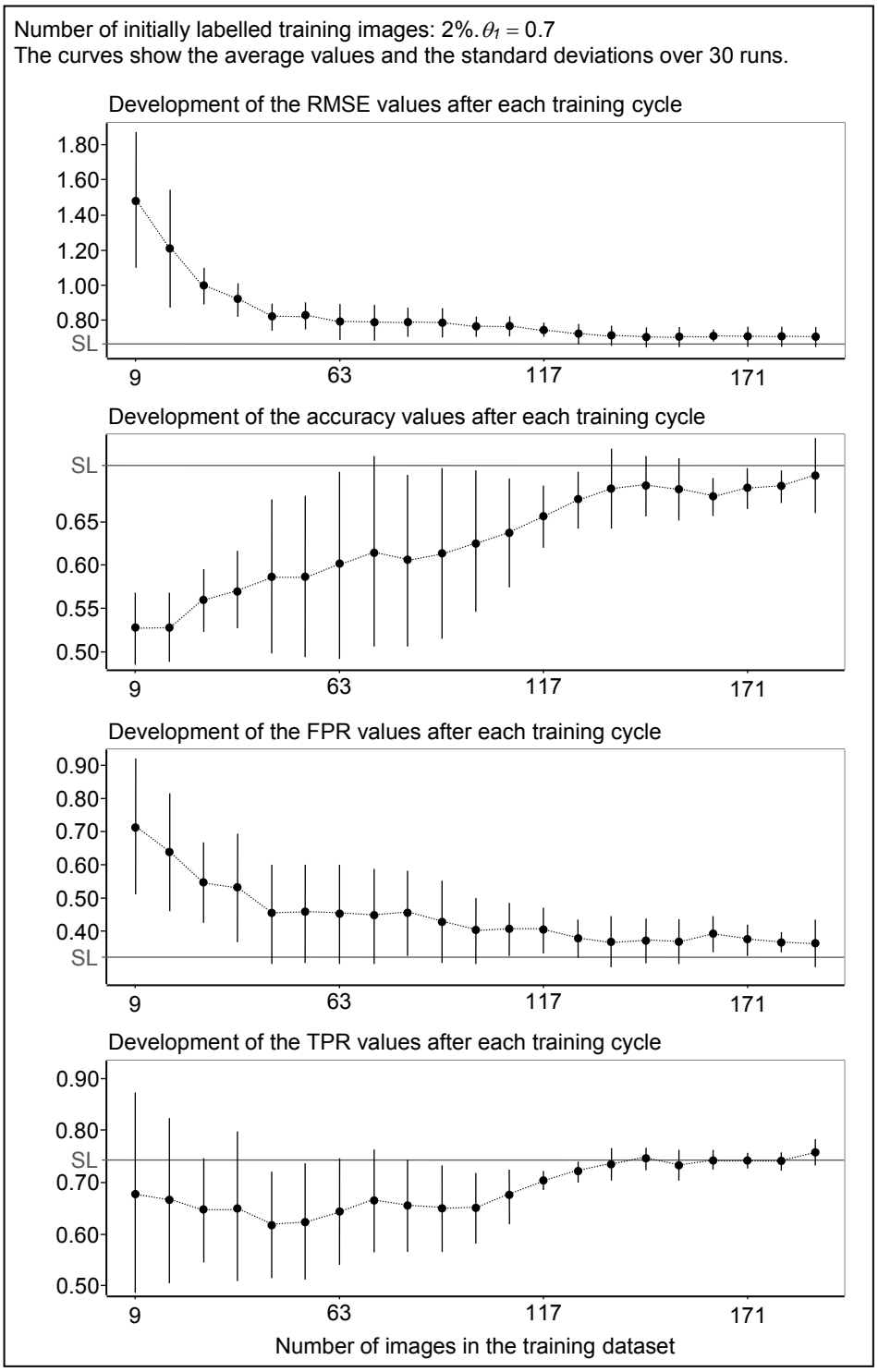


Figure 241: AL, PC: learning curves for the distortion and double image dataset, θ_1

In the second step, a training image is labelled by the Oracle if the difference between the largest and the second largest class probability is lower than the threshold θ_2, which is set to 0.2. The corresponding learning curves for 2% manually labelled images are shown in Figure 242.

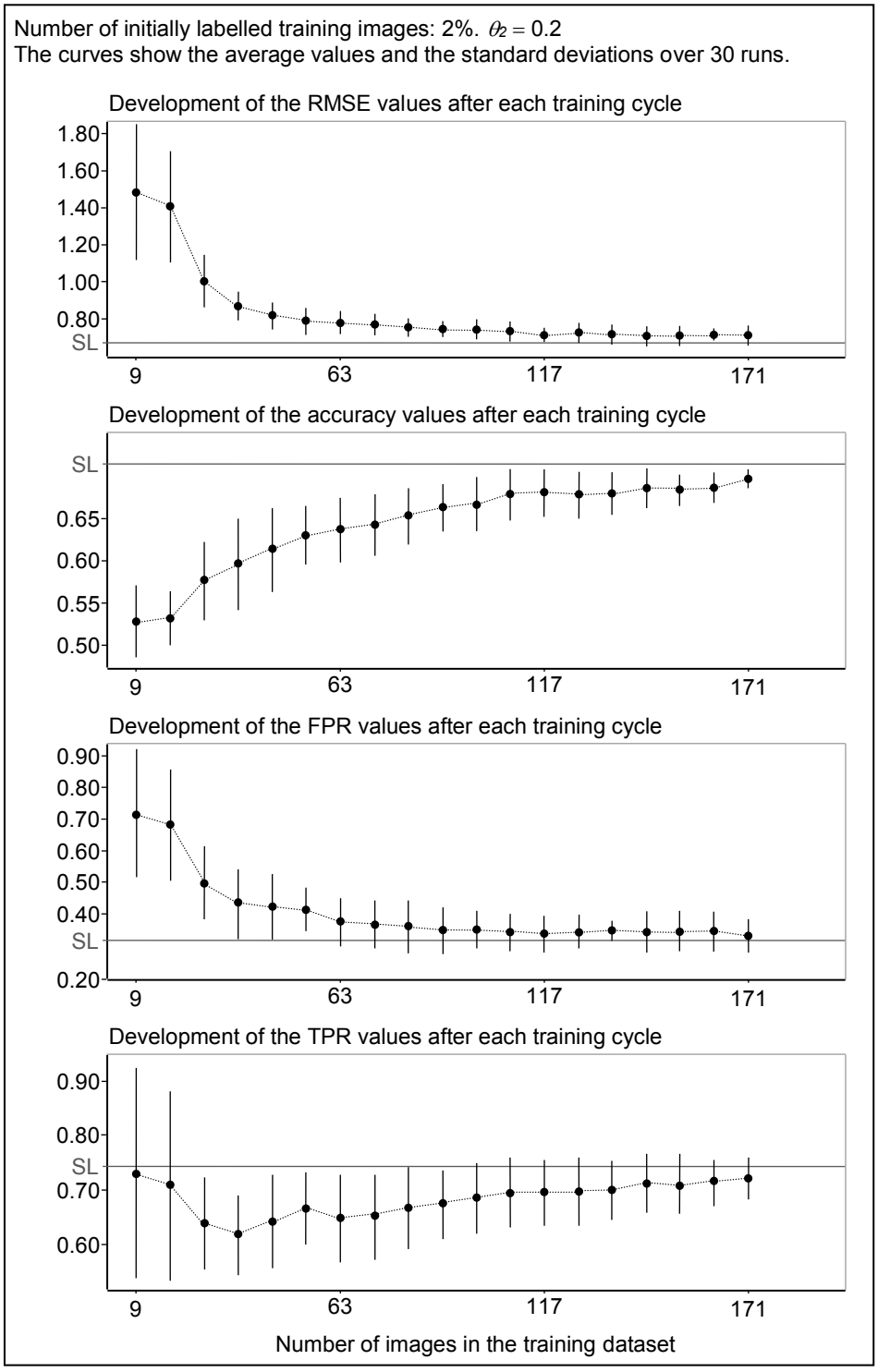


Figure 242: AL, PC: learning curves for the distortion and double image dataset, θ_2

Finally, a training image is labelled by the Oracle if the maximum probability that the image belongs to the corresponding rating class is lower than the threshold θ_1 and if the difference between the largest and the second largest class probability is lower than the threshold θ_2. The resulting learning curves for 2% manually labelled images and $\theta_1 = 0.6$ and $\theta_2 = 0.1$ are shown in Figure 243.

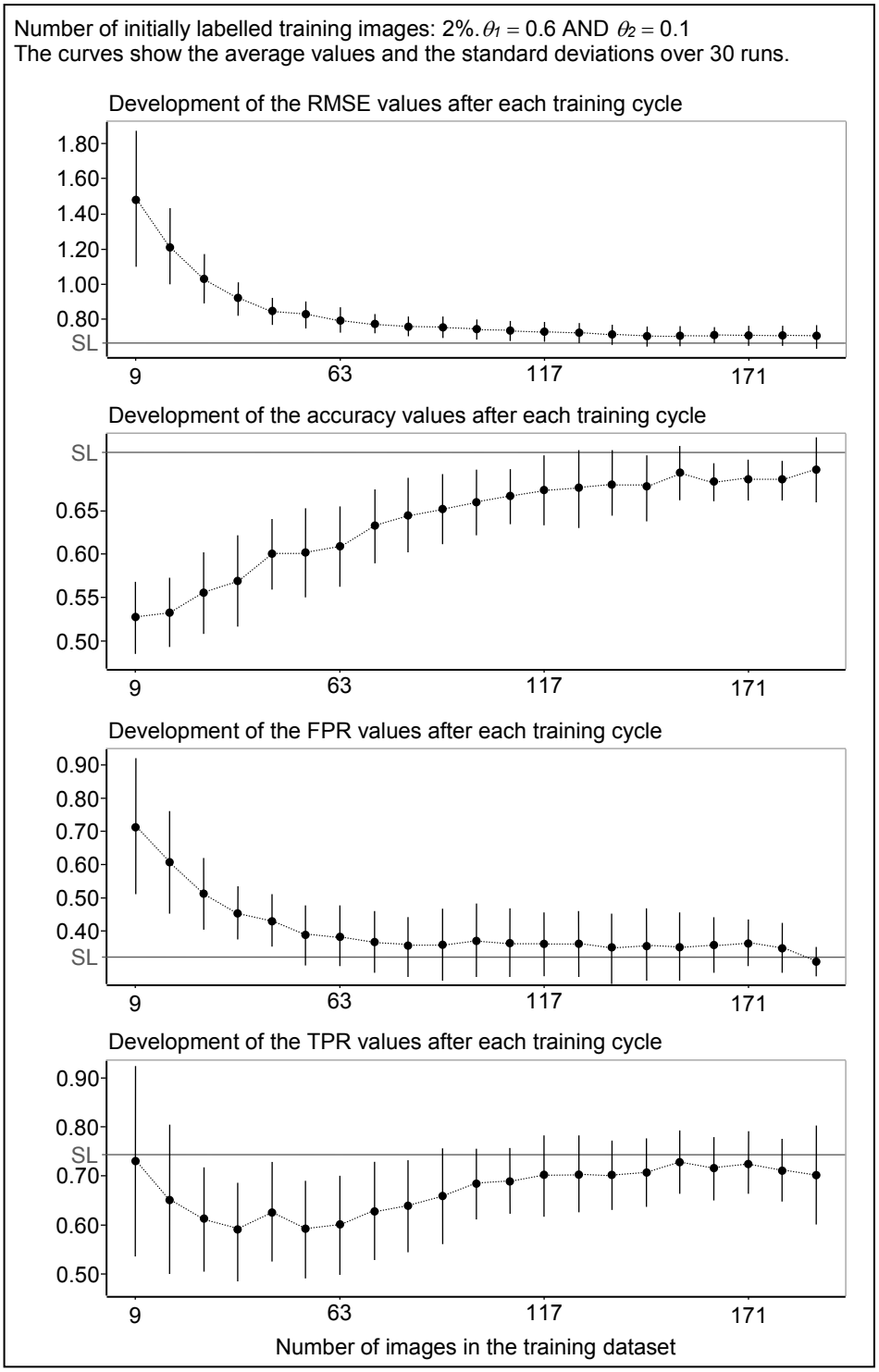


Figure 243: AL, PC: learning curves for the distortion and double image dataset θ_1 AND θ_2

A.39 AL, kNN: results for the distortion and double image dataset

The active learning algorithm is able to query interactively for the correct label if the distance to the next training sample is greater than the threshold distance. The threshold distances are already determined in Table 39. The size of the initially labelled training set is set to 0.5%, 2%, and 5% of all training images from each rating class; see Table 43. The classification results after the termination of the AL process are summarised in Figure 244 and Figure 245. Since the selection of the labelled training data is done randomly, the average values and the standard deviations over 30 runs are calculated. The x-axes represent the number of initially labelled images in the training dataset and the y-axes the classification results.

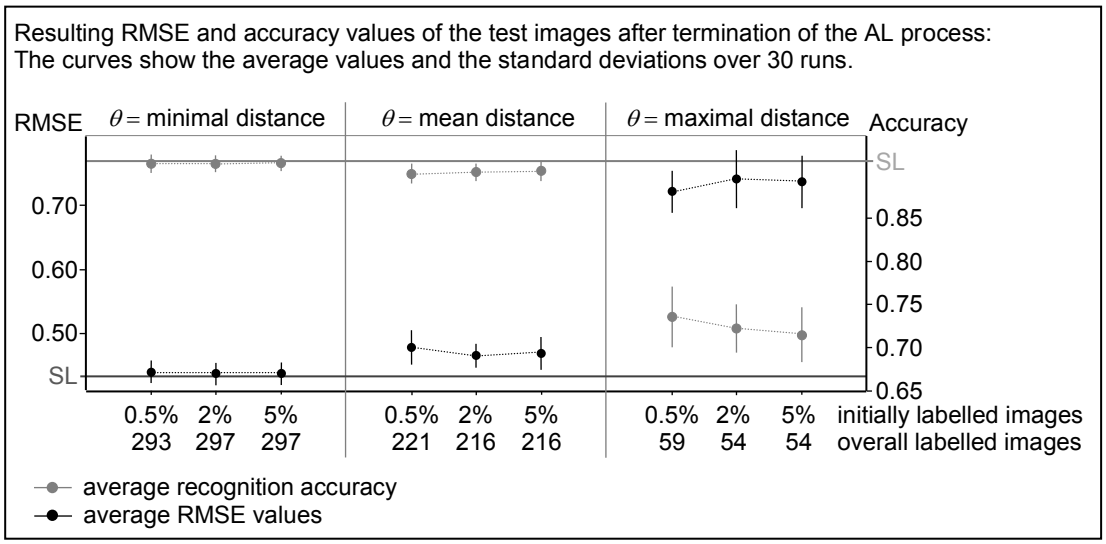


Figure 244: AL, kNN: results for the distortion and double image dataset, part I

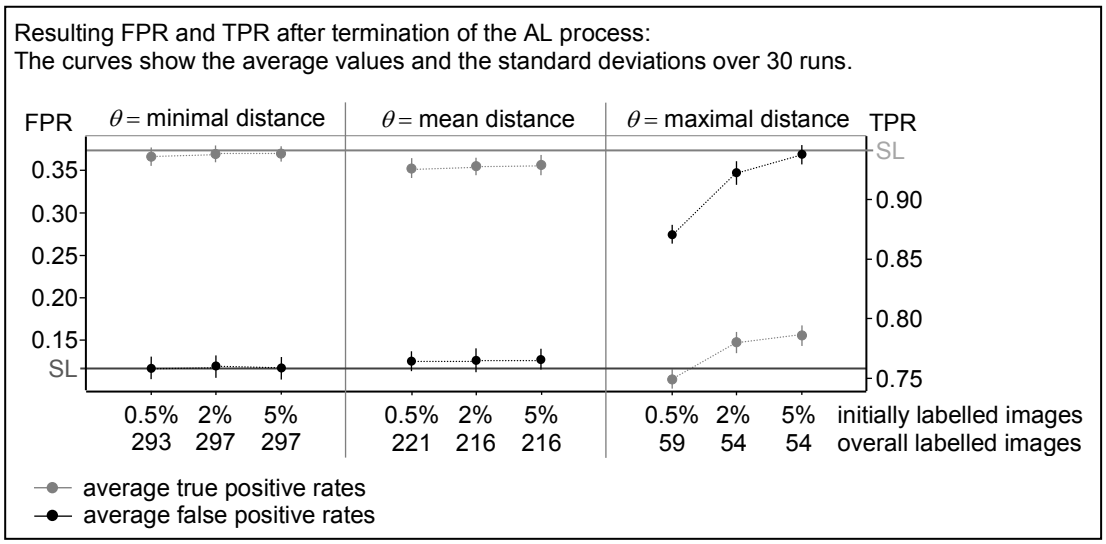


Figure 245: AL, kNN: results for the distortion and double image dataset, part II

A.40 AL, kNN: learning curves for the distortion and double image dataset

The learning curves for the distortion dataset for 5% manually labelled images are shown in Figure 246. If the distance to the next training sample is greater than the threshold distance, the image is labelled by the Oracle and transferred into the training dataset.

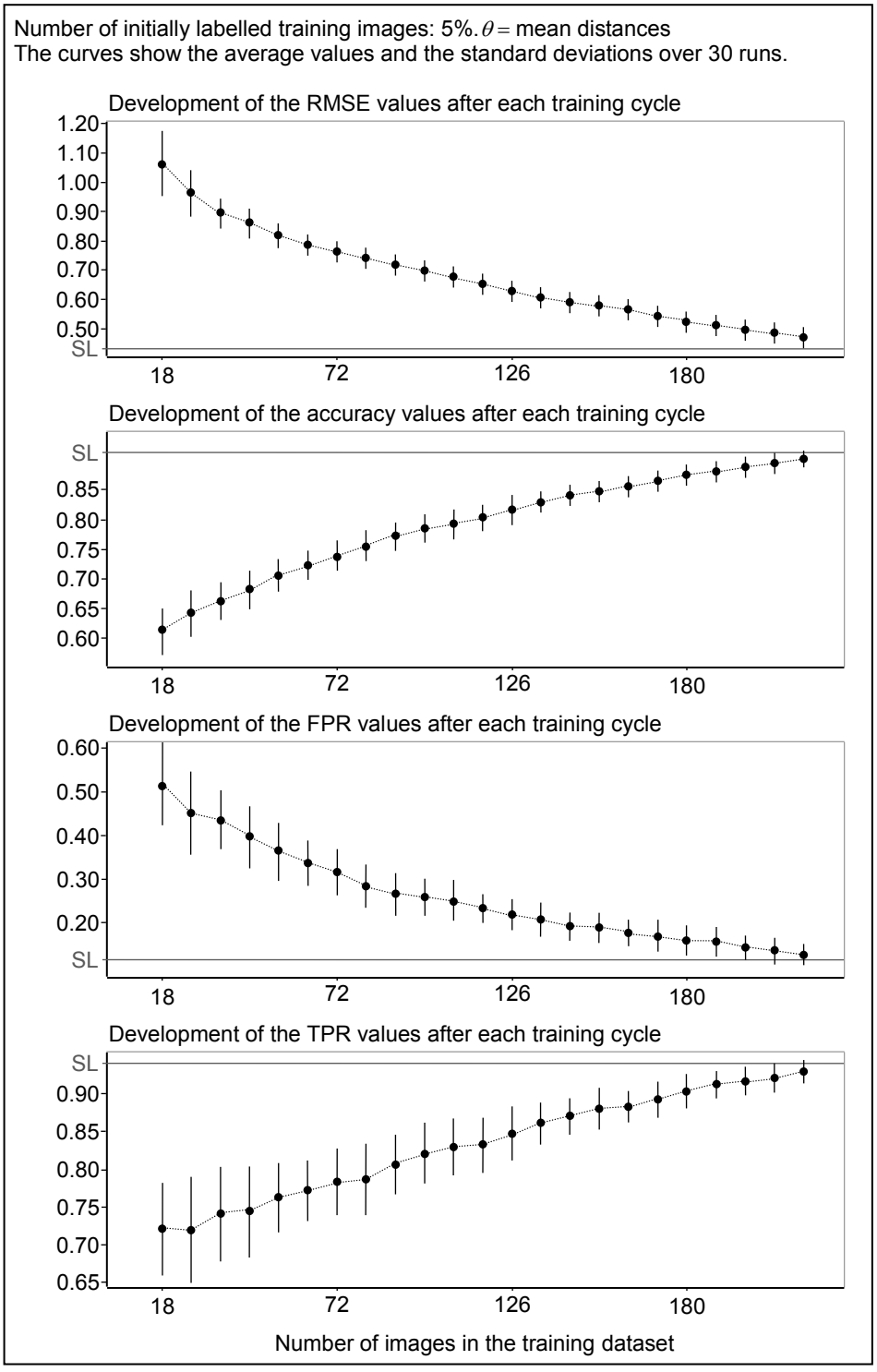


Figure 246: AL, kNN: learning curves for the distortion and double image dataset

A.41 AL, LVQ: results for the distortion and double image dataset

The active learning algorithm is able to query interactively for the correct label if the distance to the next training sample is greater than the threshold distance. The threshold distances are already determined in Table 40. The size of the initially labelled training set is set to 0.5%, 2%, and 5% of all training images from each rating class; see Table 43. The classification results after the termination of the AL process are summarised in Figure 247 and Figure 248. Since the selection of the labelled training data is done randomly, the average values and the standard deviations over 30 runs are calculated. The x-axes represent the number of initially labelled images in the training dataset and the y-axes the classification results.

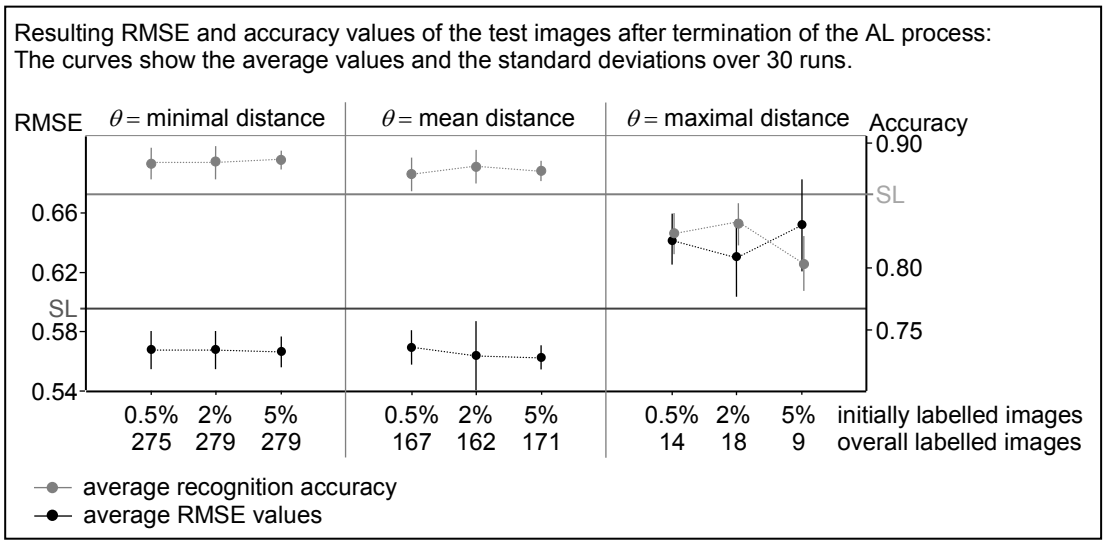


Figure 247: AL, LVQ: results for the distortion and double image dataset, part I

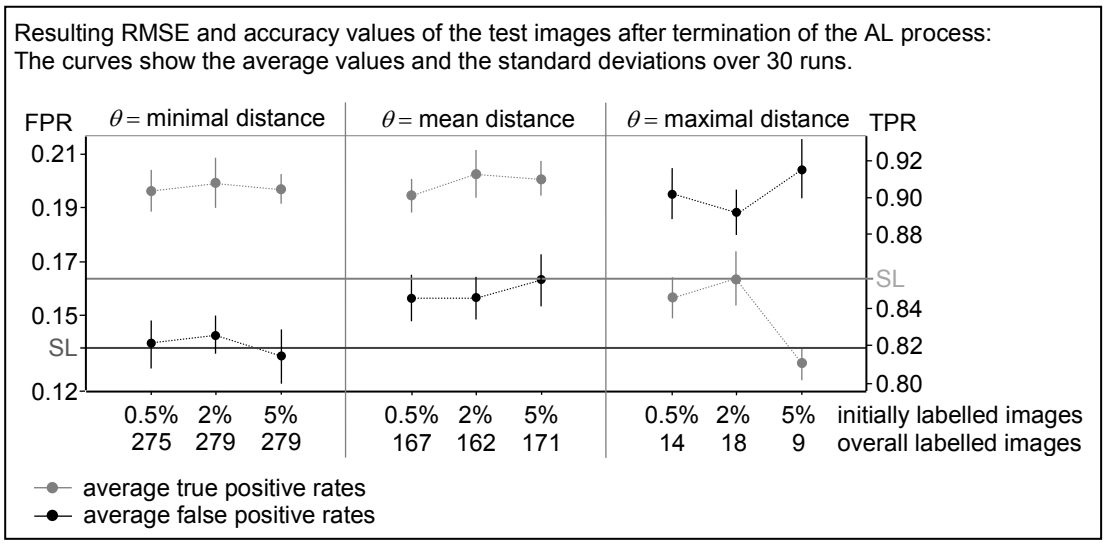


Figure 248: AL, LVQ: results for the distortion and double image dataset, part II

A.42 AL, LVQ: learning curves for the distortion and double image dataset

The learning curves for the distortion dataset for 2% manually labelled images are shown in Figure 249. If the distance to the next training sample is greater than the threshold distance, the image is labelled by the Oracle.

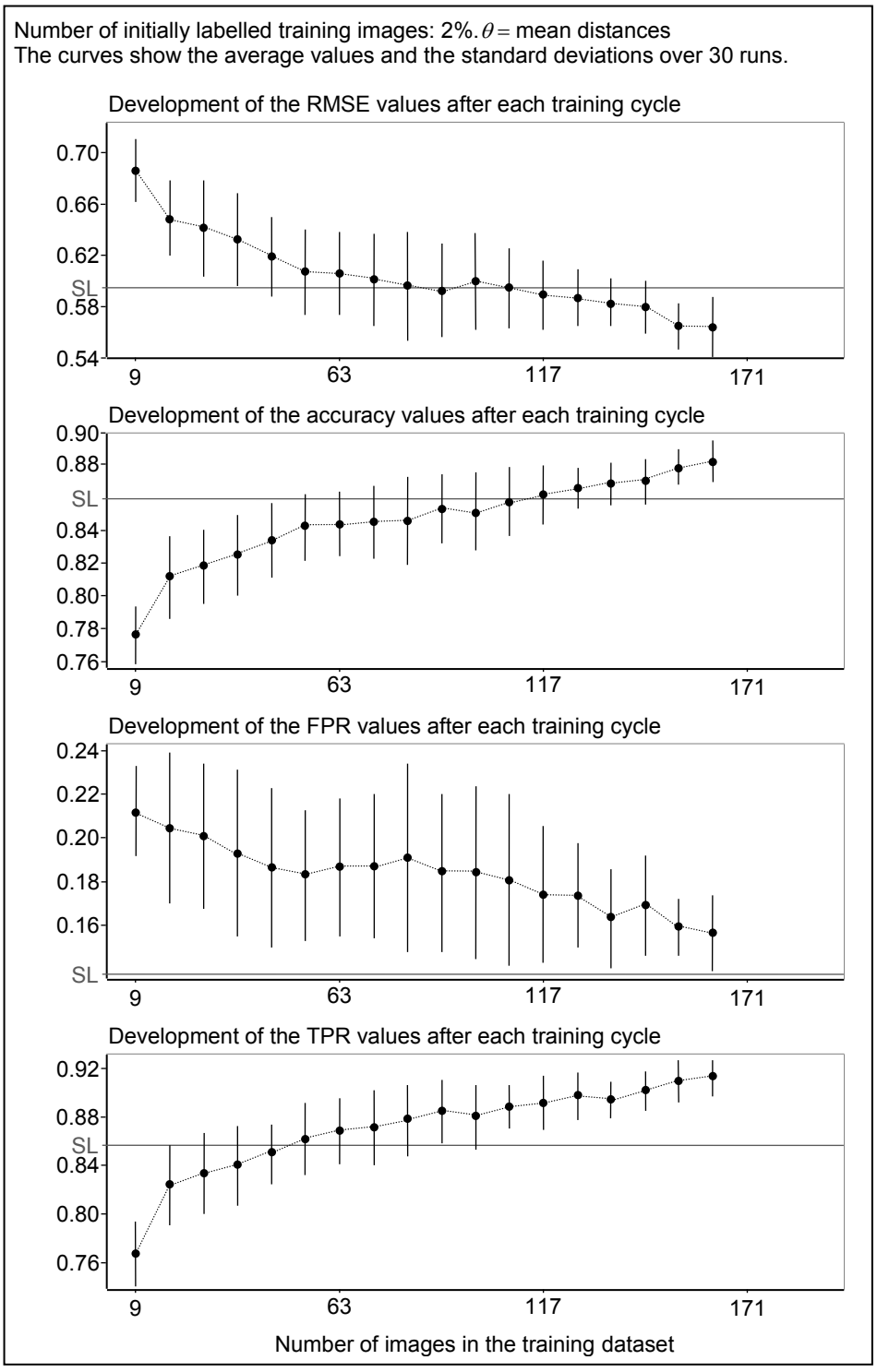


Figure 249: AL, LVQ: learning curves for the distortion and double image dataset

A.43 List of figures

A.44 List of tables

A.45 List of equations

A.46 CV / Publications

Name: ..... Sonja Maria Köppl
Date of birth:..... 08 February 1985
Nationality: ..... German
E-mail address: ..... sonja.koeppl@tu-dortmund.de

Academic studies: ..... Electrical Engineering and Informatics

Engineering Degree (Dipl.Ing.)
University of Applied Science, Kempten, Germany

Master of Engineering (M.Eng.)
University of Applied Science, Augsburg, Germany

Master of Engineering (M.Eng.)
University of Ulster, Belfast, Northern Ireland

Ph. D. studies:..... Realised at the research centre of the Daimler AG, German
In cooperation with the University of Dortmund, Germany.

Supervised theses:..... Master thesis of M. Marcel.
Objektive Bewertung der wahrnehmbaren Qualität von Head-Up Displays mit Fokus auf binokulare Disparität. Hochschule Ulm, 2014.

Paper: ..... S. Köppl, M. Hellmann, K. Jostschulte, C. Wöhler.
Evaluation of the individually perceived quality from head-up-display images relating to distortions.
In: A. F. X. Wilhelm, H. A. Kestler (eds.), Analysis of Large and Complex Data. Studies in Classification, Data Analysis, and Knowledge Organization. Springer-Verlag, 2016.

Presentations: ..... S. Köppl, M. Hellmann, K. Jostschulte, C. Wöhler.
Apply classification methods to predict the perceived quality from head-up display images.
Second European Conference on Data Analysis, Bremen, Germany, 2014.

S. Köppl, M. Hellmann, K. Jostschulte, C. Wöhler.
Concretion of subjective vehicle impressions – Objectification of the individually perceived quality from head-up display images.
Second European Conference on Data Analysis, Bremen, Germany, 2014.