Bundesamt
für Sicherheit in der
Informationstechnik

BioFace

Comparative Study of Facial Recognition Systems

Public Final Report
BioFace I & II

June 2003

Version 2.1

Bundesamt für Sicherheit in der Informationstechnik
Godesberger Allee 185-189, 53175 Bonn ⬩ Postfach 200363, 53133 Bonn
Tel.: + 49 (0) 1888 9582-0 ⬩ Fax: + 49 (0) 1888 9582-400 ⬩ Internet: www.bsi.bund.de

Table of Contents

Preface

Since the terrorist attacks of 11 September 2001 in the USA interest in biometric identification systems has once again risen sharply among law enforcement agencies. In connection with the Federal Ministry of the Interior's package of anti-terrorism measures, the use of biometric identification to authenticate immigrants and visa applicants (Immigration Act, Prevention of Terrorism Act) is under consideration. Authentication in this context refers to the unmistakable link between an individual and his or her identity. This identity can be stored on a document available for production, whose authenticity can be verified, or it can be in a central repository. Where it is stored is immaterial. The link can be evidenced by distinguishing physiological or behavioural characteristics of the person – biometric features – that are captured and verified. The characteristics to be captured can be of a static nature, for example, fingerprints, hand geometry etc., or they can describe dynamic attributes such as the voice.

In the sub-projects BioFace I and BioFace II, a comparative study of the recognition performance of facial recognition systems was carried out. The studies were conducted firstly at the level of pure algorithm tests (laboratory tests) in the area of verification (1:1 comparison) and identification (1:n comparison) and secondly at the level of a test under realistic conditions of use in the area of identification (practical tests/system test). The primary aim was to analyse the capability of the systems with large volumes of data and the influence of noise factors. The present report documents the framework conditions, the data material used and the procedure and results of the studies themselves.

BioFace I and BioFace II constituted a joint project involving both the Federal Office for Information Security (FOIS) in Bonn and the Federal Office of Criminal Investigation (BKA) in Wiesbaden, under the overall control of the FOIS. Additional assistance was provided by the Fraunhofer Institute for Computer Graphics Research (IGD).

We would like to thank all those involved in the project for their commitment and especially the Fraunhofer Institute for Computer Graphics Research for carrying out the study so competently.

Bonn, Wiesbaden
June 2003

Naming conventions for the BioFace project series

The BioFace project series as carried out to date has been divided into two parts, each identified by Roman numerals. BioFace I entailed the assembly of images from various sources, development of a classification scheme and classification of the images according to this scheme (see section 4). In BioFace II, two different tests, an algorithm test and a system test, were then carried out using the data generated during BioFace I. As these were originally planned to follow a chronological sequence, the algorithm test is referred to as BioFace II Phase 1 (section 5) and the system test as BioFace II Phase 2 (section 5.7.5.6). The naming conventions adopted in the individual studies in the algorithm test are defined in section 5.4.1. The system test was carried out in two parts: Part 1 comprised enrolment under ideal conditions, while in Part 2 enrolment was carried out by means of digital photography.

BioFace I : data capture and classification
BioFace II 1: algorithm test
BioFace II 2.1: system test with enrolment under ideal conditions
BioFace II 2.2: system test with enrolment of digital images

Table 1: naming conventions used during the BioFace projects

1 Summary

1.1 Conclusions of the algorithm tests

1.1.1 Facial recognition systems as an aid to verification

The investigations for the verification scenario were based in each case on one image per person in a search database and a second image of this person in the reference database. In each case the two images were compared with each other ("cross-over test") and the comparison results (matching scores) were recorded. Databases containing images of 5,000, 10,000, 20,000 and 50,000 persons were investigated.

In the verification scenario, the size of the database against which verification was carried out (reference database) had no significant effect on the matching scores. Although there was a shift in the frequencies of matching scores for images of different persons (non-matches) in the direction of higher scores, this trend was so small that it can be ignored.

This suggests that the extracted biometric features (templates) within the reference database were independent of each other as far as the algorithms studied were concerned.

Age differences between images of a person (tests were conducted using images captured up to ten years apart) did not pose any serious problems to the systems. Although the matching scores declined the further apart the dates of origin of the images compared, nevertheless, the difference was not so significant as to throw into question the verification results. However, taken together with the ability to separate between matches and non-matches (see next paragraph), complications could occur when the matches fall below a certain value due to age differences.

More sensational, on the other hand, is the observation that the range of values for comparisons of the image of a person with other images of the same person (matches) and comparisons of the same image with images of other people (non-matches) strongly overlapped. Moreover, this overlapping increased as a result of the upward shift in matching scores described above, as reference database size increases. While this did not indicate the existence of "biometric twins", as the templates did not necessarily have to be identical, many templates were extremely similar, so that one could speak of "almost biometric twins". This means that the systems were unable to distinguish clearly between matches and non-matches. Consequently it was very difficult to define a threshold from which a system was totally certain that the person to be verified was the same as the person on the reference image. One could of course define thresholds, but with all the systems these would be so high that too many matches would incorrectly be classified as "not identical", i.e. the false rejection rate (FRR) would be very high.

However, it should be noted that the quality of the image material in the reference database had a significant effect on the scoring of non-matches. This suggests that the excessively "good"/high scores obtained for non-matches in BioFace II were not just due to the quality of the algorithms. This fact will be borne in mind in a follow-up project, in which the images in the reference database will have to satisfy high quality requirements, so that they do not "confuse" the algorithms through poor quality. However, one should also bear in mind that when such systems are used in the real world, optimal image material will not necessarily be available.

Given the aforementioned limitations, the suitability of facial recognition systems as (supporting) verification systems is neither proved nor disproved by BioFace II. The stability of the scoring of matches proves that the systems possess the reliability that is necessary. However, the systems do not provide the reliable differentiation between "biometric twins" that is necessary for their use in practice.

1.1.2 Facial recognition systems as an aid to identification

For the investigations relating to the identification scenario, searches were carried out for images of 116 persons in differently sized reference databases containing 305 different images of precisely those persons. The reference databases contained 1,000, 5,000, or 50,000 images of other persons (filler images). The identification outcome was a list of the ten best matches per identification run.

In the identification scenario, the size of the database turned out to have a clear influence on the recognition performance of the systems. During the investigations, the position of the matches within the ten best matches was considered. It emerged that the larger the reference database, the more non-matches displaced matches in the first ten positions.

In other words, the systems made more and more mistakes, which can be explained in terms of the results of the verification scenario (see above): the larger the reference database was, the higher the average scores of the non-matches.

The same reservations apply to this interpretation of the study results as for the verification scenario.

Given the aforementioned limitations, the suitability of facial recognition systems as (supporting) identification systems is neither proved nor disproved by BioFace II. However, in the identification scenario there is less room for compensating for the weaknesses of the systems as regards separating matches from non-matches than in the verification scenario, so that in this case further improvements to the algorithms are imperative before the systems are suitable for use.

1.2 Conclusions of the system test

In the system test, 20 subjects passed by the cameras of four facial recognition systems as they entered a defined area, during which they consciously looked at the camera. The systems had to identify the person in a database which contained 500 images of other persons as well as one image of each subject. A person was deemed to have been successfully identified when that person’s image was included in the five best matches.

The system test was carried out in two phases: during phase I, each subject stood in front of the cameras and the images taken in this way were used as reference images. In phase II, the subjects were photographed separately using a digital camera and the digitised images were then enrolled into the systems.

It must be stated at the outset that all the systems performed better using the reference images generated during Phase I than with the reference images from Phase II. This leads to the following conclusions:

1. The more similar the environment of the images to be compared (background, lighting conditions, distance/size of the head), the better the facial recognition performance.
2. The more different the optical characteristics of the camera used for the enrolling process and for photographing the comparison image (light intensity, focal length, colour spectrum etc.), the worse the facial recognition performance.

All in all, two out of the four systems tested had a false rejection rate (FRR) of 64% and 68% respectively in Phase I and 75% and 73% in Phase II. The two other systems, with FRRs of 90% and 98% (Phase I) and 99% and 99.7% (Phase II), hardly recognised any of the subjects, and the weaker of these two systems was in fact so unreliable that it was sonly available for use on a few of the days.

Recognition performance was not nearly as good as the promotional material of the system vendors would lead one to believe. A recognition performance of just under 50% may be adequate in an automated surveillance scenario (in which, despite the shortcomings, almost one-half of the population examined are detected without human intervention), but for automated access control it is totally unacceptable as over half of those entitled to pass would be denied access.

Another important aspect of testing the practicability of the system is the availability of the systems and the quality of the support provided by the vendor or distributor of the system. Here it was evident that some vendors/distributors were unable or unwilling to provide all-round support. In some cases they were not even able to answer technical questions themselves, or experts had to be flown in from faraway to get the systems up and running and carry out any necessary troubleshooting on-site. This latter requirement alone effectively rules out the use of such systems in German federal authorities simply from a security point of view.

However, when evaluating the conclusions, especially those regarding the technical performance of the systems, it should be borne in mind that the sample (i.e. the number of subjects) was not large enough to produce statistically resilient figures. Due to absences, on average only 10 to 15 of the 20 subjects actually passed the access point per day, so that the usable sample size was reduced still further. Moreover, the subjects selected did not have to meet any requirements as regards well differentiated facial features, so that “problem cases” significantly depressed the relative recognition performance downwards.

To achieve adequate results for a surveillance or access scenario, the system test would have to be repeated with a significantly larger number of subjects.

2 Management Summary

2.1 Conclusions of the algorithm tests

2.1.1 Facial recognition systems as an aid to verification

The suitability of facial recognition systems as (supporting) verification systems is neither proved nor disproved by BioFace II. The stability of the scoring of matches proves that the systems possess the reliability that is necessary. However, the systems do not provide the reliable differentiation between "biometric twins" that is necessary for their use in practice.

The following limitation applies: the quality of the image material of the reference database has a material effect on the scoring of non-matches (comparison of images of two different people). This suggests that the excessively "good"/high scores obtained for non-matches in BioFace II were not just due to the quality of the algorithms. This fact will be borne in mind in a follow-up project, in which the images in the reference database will have to satisfy high quality requirements, so that they do not "confuse" the algorithms through poor quality. However, one should also bear in mind that when such systems are used in the real world, optimal image material will not necessarily be available.

2.1.2 Facial recognition systems as an aid to identification

Given the aforementioned limitations, the suitability of facial recognition systems as (supporting) identification systems is neither proved nor disproved by BioFace II. However, in the identification scenario there is less room for compensating for the weaknesses of the systems as regards separating matches (comparison of two images of the same person) from non-matches (comparison of the images of two different people) than in the verification scenario, so that in this case further improvements to the algorithms are imperative before the systems are suitable for use. However, the same limitation applies as for the verification scenario.

2.2 Conclusions of the system test

The recognition performance achieved of just under 50% may be adequate for an automated surveillance scenario, but is unacceptable for automated access control.

With regard to system availability and the quality of support available from the vendor or distributor of the system, it was evident that some vendors/distributors were unable or unwilling to provide all-round support. On some occasions it was necessary for technicians to be flown in from faraway for troubleshooting purposes. This latter requirement alone effectively rules out the use of such systems in German federal authorities simply from a security point of view.

However, it should be noted that the sample (i.e. the number of subjects) was not large enough to produce statistically resilient figures, i.e. the system test would have to be repeated with a significantly larger number of subjects.

3 Introduction

A biometric system can be used either to verify a predefined identity (verification) or to search for one identity which matches a person from a set of possible identities (identification).

In a verification system, the user data to be compared can be held, for example, on an authentic document (identity card etc.). At the time of verification, a comparison is then carried out with the stored biometric data (1:1 match). In an identification system on the other hand, the template created from the user is compared with n possible, registered templates and the best match is ascertained. However, the template presented and the best match have to achieve a defined minimum similarity. This means that to authenticate the person and hence reliably attribute to that person the identity associated with the template, the degree of similarity has to exceed a threshold.

The aim of the BioFace project was to assess the effectiveness of current facial recognition technology. Compared with previous laboratory and field studies in Europe, which have concentrated on the measurement of performance within small and medium-sized groups, one of the distinguishing features of BioFace was that a large number of enrolled persons was used.

The project was divided into two phases. In Phase I, the algorithms of different vendors were tested under the same conditions, including with different databases of photographs increasing in size. In Phase II the system was installed at the Federal Office of Criminal Investigation, and subjects passing an access control point were compared with images held in a database and identified on that basis.

4 BioFace I

The first part of the project served to prepare the way for succeeding parts of the project. Checking the usability of the images was the most important activity here.

4.1 Preliminary studies and preparatory work

To exclude the possibility of defective or non-JPEG standard compliant images influencing the investigation later on, all the images were first of all examined using the ImageMagick[1], image processing software, which was available free of charge. During this process the following parameters were collected and stored in the image database:

- pixel size in the X direction
- pixel size in the Y direction
- colour depth in bits
- number of colours used
- colour or greyscale image
- file size in bytes

Any defective files which could not be opened or which produced an error on opening were weeded out. To find any bit-identical files, the MD5 value of every image was worked out and stored in the database. Duplicate images were weeded out by this means.

The remaining stock of images were classified according to different noise factors. The noise factors defined and used are explained below.

4.2 Noise factors

The next stages of the project were intended to examine the extent to which the behaviour of facial recognition software is influenced by noise factors. As no experience yet existed in this area, the question of which parameters might influence the quality and usability of an image for facial recognition was first of all debated from a theoretical point of view. These noise factors were grouped according to their causes (as far as possible) and combined in a classification scheme that enabled the available images to be checked for suitability in advance, independently of any facial recognition software.

The resulting groups are as follows:

- noise factors that are attributable to the person who is the subject of the image
- photographic noise factors (both at the time of taking the image and also during subsequent digital processing)
- errors during recording

4.2.1 Optimal photographs

For a facial recognition system that works with two-dimensional photographs, the best angle to take the photographs from is directly in front. Although photographs in which the head is turned or inclined at an angle can be corrected using various image transformations, the part concealed by the head cannot be reconstructed.

Considered from the point of view of radiometric (photographic) parameters, in the ideal photograph, the subject is evenly and well illuminated, the face fills the image and there are no shadows or reflections on the face. During subsequent digital processing, the resolution

[1] ImageMagick: http://www.imagemagick.org/

chosen must be adequate to allow all the details of the face to be well reproduced. With lossy compression formats, the compression factor must be chosen in such a way that no material changes (artefacts) occur in the image material. It should be borne in mind here that modern formats are significantly less vulnerable to such compression-induced changes due to wavelet compression (e.g. LuraWave or JPEG 2000).

It is obvious that as great an area of the face should be identifiable as possible. Accessories such as sunglasses, hats or caps etc. naturally constitute noise. Even normal spectacles can confuse the systems; however, in a civilian application the requirement that spectacles should be removed during scrutiny by the facial recognition system could not be implemented. Acceptance would be significantly worse, not just because of the higher effort associated with use, but it must also be borne in mind here that without any vision aids people with defective vision have only a limited field of view.

4.2.2 Noise factors that are attributable to the subjects themselves

Noise factors which impede identification due to the condition or behaviour of the person being photographed:

- facial expression
- ageing
- illness-induced changes, wounds
- the wearing of clothing or objects that cover part of the head, e.g. spectacles, beards etc.

4.2.3 Photographic noise factors

With these noise factors, identification is impeded by the manner of recording or object mapping:

- too much or too little light falling on the subject
- non-standard recording angle
- lack of contrast
- low resolution
- fuzziness

Identification can also impeded through properties of the photographic media:

- photographs printed on normal paper
- transparency stuck over the photograph (foreign passports)
- scanned photograph on plastic background (German identity card)

4.2.4 Errors during recording

The photograph does not comply with the requirements for a photograph from the front, for the following reasons:

- the image is in profile/semi-profile
- the head does not fill the image; the upper body or even the entire person is shown
- images of other body parts (e.g. arm with tattooing)

4.2.5 Classification

To classify the photographs according to these noise factors, a Java interface was developed which always presented the user with six images simultaneously. Next to each image the noise factors were displayed in the form of buttons, permitting the user to classify the image quickly and simply. The classification tool would then write a log file that could be further analysed afterwards.

4.3 The test images

4.3.1 Resolutions

The images were available in different geometric resolutions. The minimum, maximum and average resolutions of the digital images are shown in Table 2:.

Minimum resolution X	Minimum resolution Y	Maximum resolution X	Maximum resolution Y	Average resolution X	Average resolution Y
66	80	1536	2048	476	642

Table 2: geometric resolution of the test images

4.3.2 Images per person

Table 3 shows how many images were available per person. For the overwhelming majority of persons, only one usable image (taken from the front) was available. However, for an appreciable number of persons there was more than one image.

Images per person	Number
1	177,181
2	7,806
3	1,454
4	326
5	95
6	22
7	12
8	3
9	2
11	2
12	1
15	1
17	1

Table 3: number of usable images per person

4.4 Call for participation

As well as preparing the image data, a selection of participating systems was also carried out for both the algorithm and system tests for BioFace II. For the algorithm tests, vendors of facial recognition algorithms were invited to submit proposals for the algorithm test in an international competition. All applications received by 31 July 2002 were then considered. The selection of systems for the system test was run in a similar fashion, except that in this case the competition was restricted to German suppliers.

In the end, three vendors took part in the algorithm test. For the purposes of this report the algorithms have each been assigned a random number and are therefore known as algorithms 1-3.

Four vendors took part in the second part, the system test, the three vendors from the algorithm test plus one additional one. To avoid confusion, the systems have been called systems A to D, with numbering of algorithms and numbering of the systems independent of each other.

5 BioFace II Phase 1: algorithm test

Phase 1 examined the effects that various factors would have on recognition performance of the facial recognition algorithms under investigation. The investigation focussed on application scenarios with a large number of users.

5.1 The test environment

Two computers were prepared for the algorithm test, a database computer which provided the information on test sequences and recorded the results and a test computer on which the special recognition algorithms performed the investigations.

The database computer was a FreeBSD machine, which held an SQL database. The information about the images and also the test schema was stored on this in advance. Again, the analysis of the log files was added to this database.

In accordance with the participating companies' specifications, the test computer had a 2.2GHz Intel® Pentium® IV processor and 1 GB RAM. The operating system used was Microsoft® Windows® 2000 Professional. First of all the operating system and the necessary drivers were installed and then the software interface needed for the test. A backup copy of this basic installation was created which then served as the basis for the installation of all the algorithms.

Following the installation of an algorithm, one image was then created and this was re-input after every test run, so as to eliminate any possible side-effects due to changes in the operating system configuration or stability.

For every test run, a batch file was created with the aid of test definitions from the database. This file automatically enrolled the relevant images into the biometric system and, depending on the particular test definition, then performed either identification or selective verifications with other commands. The log files created were then transferred to the database computer and the algorithm image was reloaded.

5.2 The test database

An SQL database was used (PostgreSQL[2]) to store the test definitions, the relevant set of test images, the image information and in some cases the test results.

Figure 1 shows the database scheme that was used to store the data in the algorithm test.

[2] PostgreSQL: http://www.postgresql.org/

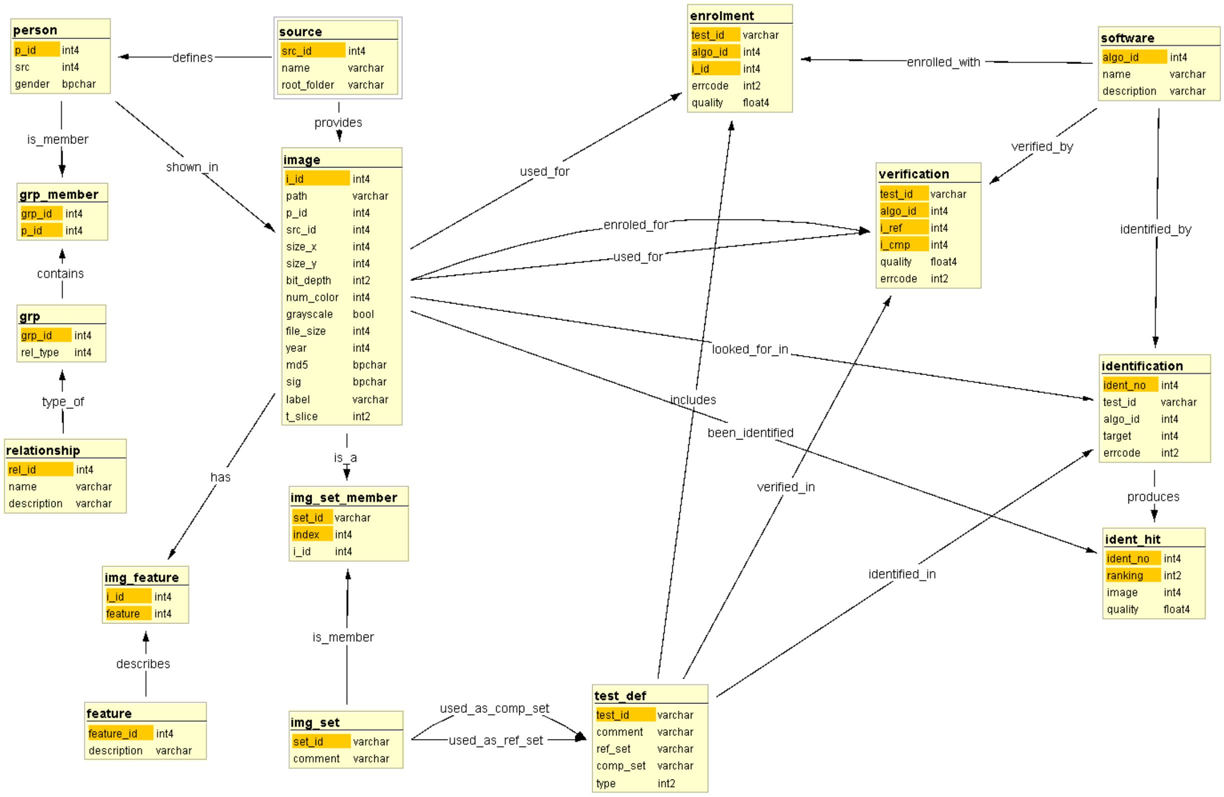


Figure 1: database structure BioFace II Phase 1

The database scheme was divided into two parts. The first part contained information on the images used, the test groups and test definitions, while the second part contained in some cases the results of the algorithm test for the scoring.

The central element of the information on the test images was the "image" table, which held the relevant technical information for every image. For each image, the origin was stored as a link to the "source" table. Moreover, images of the same person had to be ascertained through an identical reference to the element in the "person" table that corresponded to that person. The noise factors were collected in the "img_feature" table. One element from "image" could contain several noise factors, i.e. several elements from "img_feature" could point to the same image. The test definitions and also the reference and comparison sets were recorded as elements of the "test_def" table, with image sets being elements in "img_set". The images associated with the sets were entries in the "img_set_member" table, which in turn pointed both to the image in "image" and also to the associated set in "img_set".

The results section consisted of three tables, "enrolment", "verification" and "identification". For each enrolment, the following information was held in the "enrolment" table: test ID, algorithm used, image concerned as reference to "image" and results in the form of error code and matching score. The procedure was the same, whether enrolment was for a verification test or an identification test. In the case of verification, reference and comparison images were recorded in a line in "verification", which in turn pointed to the element in "image". Moreover, a pointer to the "software" table showed which algorithm was used to carry out a given comparison. Since several results, possibly differing in number, had to be stored during an identification, the data on the identification itself (algorithm, search image, test) had to be recorded in the "identification" table, and for each hit a "ident_hit" entry pointing to the associated identification in "identification" was generated.

5.3 Definition of terms

Term	Definition
Matching score	Value which stated the degree of agreement between two facial images.
Comparison	A comparison image was sent to the system, with details of the identification of one particular image within the reference set. The outcome of the comparison was the matching score.
Search	A search image was sent to the system without any additional information. The result was a number of hits and their matching schools.
Reference set	The set of images which are enrolled into the system.
Comparison set	The set of images that were sent to the system for comparison.
n	Number of images within the reference set used in the test.
m	Number of images within the comparison set used in the test.
x	Any image from the reference set
y	Any image in the reference set that was not x.
x'	The image from the comparison set that was taken from the reference set and in the reference set was image x.
y'	The image from the comparison set that was taken from the reference set and in the reference set was image y.

Table 4: definition of terms

5.4 Test identification

5.4.1 Test names

The tests were numbered using the following schema:

The code number was divided into three fields, each of which was separated by a hyphen.

1. Test group indicator field: this value specified the group to which that test belonged. The possible values were:
   1 - Noise-free comparison
   2 - Noisy search with good reference material
   3 - Noisy search with good comparison material
   4 - Hunt
   5 - Long-term comparison
   6 - Long-term comparison in large reference set
   7 - Other tests
2. Unique identification number within the test group
3. Algorithm designation: if this field did not exist, then tests on all the products were meant.

5.4.2 Set numbering

To differentiate between the various sets of images that were used within the tests, they were labelled according to the following schema. The code number began with a letter which classified the use (R = reference, C = comparison, L = long-term), whereby on many of the tests the reference set was also used as the comparison set. The identifying letter was followed by an identification number, which together with the letter constituted a unique identifier.

5.5 Sets

The tables below show the sets used in the tests. Unless otherwise stated, the sets contained only images that had been classified as noise-free in the first part of the project (BioFace I) so as to prevent the results being influenced by uncontrollable noise factors.

Code number	Number of images	Comments
R1	5,000	
R2	10,000	Contained R1
R3	20,000	Contained R2
R4	50,000	Contained R3
R5	40	
R6	500	Contained "noisy" images
R7	500-x	R6, adjusted to exclude poor quality images (x part)
R8	305	Reference images of different persons
R9	1,305	R8 with 1,000 noise-free images
R10	5,305	R8 with 5,000 noise-free images
R11	50,305	R8 with 50,000 noise-free images

Table 5: reference sets

Code number	Number of images	Comments
C1	500	Subset of R1, contained noisy images
C2	500	Contained the noise-free original images for C1
C3	500	Contained only noise-free images
C4	500	Contained only noise-free images
C5	500	Contained only noise-free images
C6	500	Contained only noise-free images
C7	500	Contained only noise-free images
C8	116	Search images for the persons reproduced in R8

Table 6: comparison sets

Code number	Number of images	Comments
L1	213	
L2	2	Contained different images for one person
L3	7	Contained different images for one person
L4	6	Contained different images for one person
L5	11	Contained different images for one person
L6	5	Contained different images for one person
L7	4	Contained different images for one person
L8	7	Contained different images for one person
L9	5	Contained different images for one person
L10	6	Contained different images for one person
L11	4	Contained different images for one person
L12	7	Contained different images for one person
L13	9	Contained different images for one person
L14	8	Contained different images for one person
L15	11	Contained different images for one person
L16	7	Contained different images for one person
L17	4	Contained different images for one person
L18	8	Contained different images for one person
L19	7	Contained different images for one person
L20	4	Contained different images for one person
L21	7	Contained different images for one person
L22	3	Contained different images for one person
L23	6	Contained different images for one person
L24	12	Contained different images for one person
L25	15	Contained different images for one person
L26	7	Contained different images for one person
L27	5	Contained different images for one person
L28	6	Contained different images for one person
L29	6	Contained different images for one person
L30	7	Contained different images for one person

Table 7: long-term sets

5.6 Tests

5.6.1 Test Group 1 noise-free comparison

The aim of this group of tests was to examine the influence of the number of images enrolled on the matching scores. Initially a stock of 5,000 images was used, following which further images were then added gradually. The incremental steps were 10,000, 20,000 and 50,000 images. The comparisons x x' (self-agreement) and x-y' y-x' (symmetry) were explicitly included here. All the resulting matching scores were recorded; however, for the analysis only the results of comparing images of different persons were considered, since on the basis of the image set used, these were clearly in the majority. Comparisons between images of the same person occurred only sporadically and the number of these was not sufficient to be able to draw any meaningful conclusions.

Tests	Reference set used	Size of reference set	Comparison set used	Size of comparison set
1-1	R1	5,000	R1	5,000
1-2	R2	10,000	R2	10,000
1-3	R3	20,000	R3	20,000
1-4	R4	50,000	R4	50,000

Table 8: tests for Test Group 1

5.6.2 Test Group 2 noisy search with good reference material, and Test Group 3, noisy search with good comparison material

In the course of creating and applying the classification schema for noise factors, it became apparent that a number of noise factors can only be systematically examined by creating appropriate photographs. Thus, the angle of rotation or inclination of the head was a significant influencing factor. Although the image material naturally included images with correspondingly rotated head, it was not possible to determine the angle of rotation, which at best could only be estimated. It would have been more sensible here to create corresponding photographs with clear increments of rotation angle in order to be able to find out where the limits of recognition performance of the facial recognition algorithms lie.

It was therefore agreed with the FOIS, as the sponsor, not to carry out Test Groups 2 and 3 as specified, but instead to examine them in a separate project[3].

5.6.3 Test Group 4 hunt

In this test group, a hunt scenario was simulated using images from the stocks of photographs. Several photographs of different people were digitised, with one photograph used as the search image and the rest as reference images.

The relevant reference set was enrolled into the algorithms and identification was carried out using the search images. The ten best hits were recorded.

[3]See section 7.2 on page 125.

Tests to be carried out	Reference set used	Size of reference set	Comparison set used	Size of comparison set
4-1	R8	305	C8	116
4-2	R9	1,305	C8	116
4-3	R10	5,305	C8	116
4-4	R11	50,305	C8	116

Table 9: tests for Test Group 4

5.6.4 Test Group 5 long-term comparison

The aim of this group of tests was to examine how the algorithms responded to images that had been captured over an extended period of time. In this case the reference set contained images of different persons who were photographed several times over an extended period (10 years). In this test group, the reference set and the comparison set were identical.

In Test 5-1 all the images were grouped into a single set, following which each image was removed sequentially from the set and compared with all the images in the set (n2 comparisons). The comparisons x x' (self-agreement) were explicitly included here.
During tests 5-2 to 5-30, for each person (a total of 29 persons) a separate set was generated, following which each image was removed sequentially from the set and compared with all the images in the set (n2 comparisons). The comparisons x x' (self-agreement) were explicitly included here.

Tests to be carried out	Reference set used	Size of reference set	Comparison set used	Size of comparison set
5-1	L1	213	L1	213
5-2 to 5-30	L2 to L30	Various	L2 to L30	Various

Table 10: tests for Test Group 5

5.6.5 Test Group 6 long-term comparison in large reference set

Carrying out this group of tests was optional and depended on the results of Test Group 5. If no adequate results were achieved with Test Group 5, then there appeared to be little point in carrying out Test Group 6.

This group of tests was not carried out as it would have meant examining two influencing factors simultaneously: ageing and database size. Both these factors had already been examined on their own in Test Groups 1, 4 and 5, so that combining them would not have yielded any additional useful results. Even if a deterioration had occurred in the results, it would have been impossible to determine the cause due to the combination of several sources of error (image quality, dataset).

5.6.6 Test Group 7 other tests

This group of tests comprised plausibility checks which served to generate information on the effects of the design of the data sets.

In Test 7-1, all the images were compared with each other by means of verification. The sets contained only images from the Fraunhofer Institute for Computer Graphics Research collection of images. The aim was to examine whether results broadly similar to those achieved in Test Group 1 would be obtained, so as to exclude any significant influence due to the use of the test images in the tests.

In Test 7-2, an unchanged database was then tested with itself, following which any images of poor quality were deleted (k Items) and compared once again in Test 7-3. The aim here was to examine the effects of poor quality images on the overall results.

Tests to be carried out	Reference set used	Size of reference set	Comparison set used	Size of comparison set
7-1	R5	40	R5	40
7-2	R6	500	R6	500
7-3	R7	500-k	R7	500-k

Table 11: tests for Test Group 7

5.7 Results

5.7.1 Peculiarities of the algorithms

The software for the algorithm test was installed by or made available for installation by the vendors at the Fraunhofer Institute for Computer Graphics Research in the summer of 2002. As an update of all the software or of software components would have meant that the results of tests carried out before the update were not comparable to tests carried out after the update, no software changes were allowed in the course of the algorithm test.

5.7.1.1 General

A complete run using the facial recognition systems proceeded as follows:

- Enrolment. One or more reference images were enrolled into the system, i.e. "introduced" to the system. First of all a face-finding algorithm localised the face on the images and extracted the face area from the image. From the resulting partial image, algorithm-specific features were then derived and stored in a data record corresponding to that image ("template", also known as "reference template" during the enrolment). During this operation, an identification value (ID number, name etc.) had to be given to the system so that it could be saved with the template. In the case of identification this value was subsequently returned, or in the case of verification the corresponding data record was searched out for comparison.
  If the search for the face in the image or the derivation of facial features from the extracted partial image failed (e.g. because the image did not contain any face or the quality was not good enough for further processing due to lack of contrast), the image was rejected and a corresponding error code was returned.
- Identification. A search image was passed to the facial recognition algorithm and first of all was put through the same steps as during enrolment: the face was localised and extracted, and then the features were extracted. With the resulting search template, a difference value was then calculated for each reference template contained in the database[4]. The reference template which differed the least from the search template was initially an intermediate result. However, the absolute value of this difference value was then compared with a predefined threshold value that constituted the maximum acceptable deviation. If the calculated difference between the best reference template and the search template lay below this threshold value, then the identity linked with that reference template was deemed to have been identified (see enrolment).
  If an image was not accepted by the feature extraction process (face-finding, derivation of

[4] To make the procedure easier to follow, it is presented simplified. A facial recognition system could also save the reference templates hierarchically in a tree, for example, and thus filter out in advance any part-trees that were not being considered so as to reduce the search set.

features) or the difference value of the best reference template lay above the threshold value, then no identification was returned, but a corresponding error value was reported back.
For BioFace, the stage of threshold value comparison was disabled, so as to permit access to the intermediate results. The effect of different threshold values could then be easily identified later on, e.g. using the diagrams. Moreover, in the project a group of the best hits was considered, e.g. three or ten hits, referred to in the descriptions as "results data record", rather than just the individual best hit.

- Verification. In this operating mode, a comparison image and an identity (e.g. a user name) were passed to the system. First of all once again the processing steps of face-finding and feature derivation were carried out for the comparison image so as to calculate the comparison template. The reference template(s) which went with the identity passed was then searched for and the corresponding difference values calculated. If at least one of the corresponding reference templates lay below a predefined maximum difference threshold, then the relevant identity was deemed to have been verified.
If a comparison image was not accepted by the feature extraction process (face-finding, derivation of features) or the difference value of the reference template that went with the identity lay above the threshold value, then no verification was returned, but a corresponding error value was reported back.
In the BioFace tests, the threshold value comparison was disabled and the difference values ("matching scores") were accessed directly so as to use the facial recognition algorithm as a metric for the similarity between the person in the reference image and the person in the comparison image.

5.7.1.2 Algorithm 2

Under Algorithm 2, it was observed that the number of hits specified in the interface description was not always returned independently of any other limits (see Table 12: enrolments and comparisons for Test Group 1 on page 26 and Table 16: number of identification results on page 41). As a result, the datasets for evaluation were in some cases a lot smaller than under the other algorithms.

The rising number of identification events in Table 16: number of identification results on page 41 suggests that internally there was a matching score threshold and that hits below that threshold were not returned.

5.7.1.3 Algorithm 3

Due to a software error, it was not possible to enrol all the images in Tests 1-4 using algorithm 3. Because enrolment was incomplete and the results were therefore not comparable to the results of the other systems, no analysis was carried out. Moreover, it is possible that the successful enrolments were impaired by the subsequent error.

Following termination of the investigations, for test purposes some software was made available which no longer demonstrated this error. To maintain comparability[5], no updates were carried out during the study.

5.7.2 Notes on the interpretation of the diagrams.

The diagrams generally make use of either green symbols and lines or red symbols and lines. Green stands here for the comparison between two images which represented the same person, and red for a pair of images that represented two different persons.

Some diagrams present combinations of other diagrams (e.g. Figure 44 combines Figure 45 and Figure 46). This has the effect of improving comparability, whereas the individual diagrams are easier to take in as they contain fewer symbols.

[5] See section 5.7.1 Peculiarities of the algorithms on page 23

In many of the diagrams, matching score is a dimension. It should be noted here that matching score depends on the facial recognition algorithm used, i.e. the absolute values of different algorithms and systems are not immediately comparable. Again, within the algorithm itself, the matching scores are not necessarily linear. Thus, a comparison of image A B with a score of 0.6 is not twice as good as a comparison image A C that produces a score of 0.3. The only conclusion that can be drawn from the resulting 0.6 and 0.3 scores is that in the first case the match was better.

In principle there are three types of diagram:

- Direct presentation of results. One symbol is shown for each comparison. The y-axis shows the matching score achieved, and the x-axis the person ID discovered during identification. For example, Figure 44 on page 63.
- Totals diagram. For the purposes of abstraction or because the number of individual results is so big that direct presentation of the results is not useful, a totals diagram has been created. Here the occurrence of individual matching scores (e.g. 0.3644) has been counted. A frequency distribution for the entire score range from 0 to 1 is then presented, whereby the line represents the assignment of matching score to frequency of occurrence. For example, Figure 47 on page 64.
- Receiver operating characteristics. In this type of diagram, the relationship between false acceptance rate (FAR) and false rejection rate (FRR) is plotted. If one were to choose a very low threshold value for a system, then few authorised persons would be falsely rejected (FRR low) but on the other hand many unauthorised persons would be falsely accepted (FAR high). If the threshold is positioned higher, the effect reverses: the FRR rises because more authorised persons are rejected, but the FAR falls because more unauthorised persons are correctly rejected as well. Generally, in an application where biometric systems might be used, there is no requirement for any vendor-specific threshold values; instead, requirements are expressed in terms of either a false rejection rate or a false acceptance rate or both. The ROC diagram shows quickly and clearly what the FRR is for a given FAR and vice versa. Where both values are required, it is possible to see whether the system is capable of satisfying these requirements or not. If the system is good, then the distance between the curves and the axes will be low. An example of an ROC diagram is Figure 29 on page 54.

5.7.3 Test Group 1 noise-free comparison

5.7.3.1 Explanation of results

In this group of tests, predominantly images of different people were compared with each other. The aim was to find out whether the number of similar faces would influence the average recognition performance of the algorithms as the number of enrolled images was increased. Hence, only the results of the comparisons between different persons were used for the analysis. The series began with tests 1-1, which contained 5,000 images in the reference set. The number of images in the reference set was successively increased to 10,000, 20,000 and 50,000.

Since an image compared with itself (not similar picture but bit-identical) naturally normally produced a high score, these self-agreements were excluded from the test, as they would have had the effect of disproportionately boosting the values.

If the number of enrolled images was influencing the results, the resulting distribution of matching scores would be expected to shift downwards, if the effect was to worsen scores, or upwards, if the effect was to improve the scores. However, under all three systems it was observed that both the maximum and also the shape of the curve self remained unaffected. Only the scatter of the results was smoothed out by a larger number (see Figure 15: change in scatter between 1-1 Algorithm 2 and 1-4- Algorithm 2).

5.7.3.2 Enrolments and comparisons

Test	Number of enrolments to be carried out	Number of successful enrolments	Number of relevant, successful comparisons[6]	Maximum frequency of a matching score
1-1 Algorithm 1	5,000	4,622	21,358,258	30,941
1-1 Algorithm 2	5,000	4,649	13,220,108[7]	7,013
1-1 Algorithm 3	5,000	5,000	24,994,996	38,680
1-2 Algorithm 1	10,000	9,262	85,775,374	123,244
1-2 Algorithm 2	10,000	9,264	53,074,099[7]	27,147
1-2 Algorithm 3	10,000	10,000	99,989,990	153,572
1-3 Algorithm 1	20,000	18,487	341,750,622	488,249
1-3 Algorithm 2	20,000	18,561	211,420,040[7]	107,629
1-3 Algorithm 3	20,000	20,000	399,979,920	615,288
1-4 Algorithm 1	50,000	46,217	2,135,964,534	3,054,121
1-4 Algorithm 2	50,000	46,328	1,314,049,251[7]	666,550
1-4 Algorithm 3	50,000	34,277[8]	-	-

Table 12: enrolments and comparisons for Test Group 1

5.7.3.3 Differences between 1-1 and 1-4

To investigate the change in matching scores due to database size, the values obtained in Test 1-1 were compared with those of Test 1-4. For this purpose, 1,000 comparisons were randomly selected out of the 4,999 comparison values calculated per image (self-comparisons were excluded). For these comparisons, the values in tests 1-1 and 1-4 were ascertained and compared. There were no differences between the algorithms of Algorithm 2 and Algorithm 1.

Since, due to an algorithm error, the enrolment of images was prematurely terminated during Test 1-4 Algorithm 3 (see section 5.7.1.3 Algorithm 3 on page 24), the differences could not be calculated here.

5.7.3.4 Differences in symmetries

To ascertain whether feature extraction during enrolment is different from during verification, the matching scores for symmetrical comparisons were also captured (enrolment image x compared with image y and enrolment for image y compared with image x). The following results are based on Test 1-4. Only comparisons that were successfully completed (enrol-

[6] In this connection, relevant means the images of different persons. Due to the random selection of images in the filler set, for a small number of persons (approx. 0.01%) there was more than one image in the set.

[7] The low number of analysable comparisons is due to the fact that not all the results were reported. See section 5.7.1.2 Algorithm 2 on page 24.

[8] See section 5.7.1.3 Algorithm 3 on page 24.

ment and comparison without error) and whose symmetric comparison was also successful have been used.

The deviations under Algorithm 2 extend over a wide area. This compares with only low deviations under Algorithm 3 and no deviations under Algorithm 1 during the symmetric comparisons.

	Algorithm 1	Algorithm 2	Algorithm 3
Number of comparisons that were no different from the symmetric comparisons	18,071,001	7,928	23,899,010

Table 13: number of comparisons without difference

5.7.3.5 Notes on the diagrams[9]

To create the diagrams, all the relevant matching scores and the frequency of their occurrence were counted (rounded to four decimal places). The results produced 10,000 frequency values, one for each matching score (0, 0.0001; 0.0002, 0.0003; …; 0.9998; 0.9999; 1). The corresponding frequencies were individually normalised for each test and vendor by dividing them by the total number of matching scores considered in the relevant test for the relevant vendor. For each diagram, this was arrived at by means of the integral encircled by the curve, i.e. an area corresponding exactly to size 1. In this way comparability of the test results was created which would not have been possible without normalisation, since, as the number of comparisons rose, so too did the frequency of occurrence of particular matching scores.

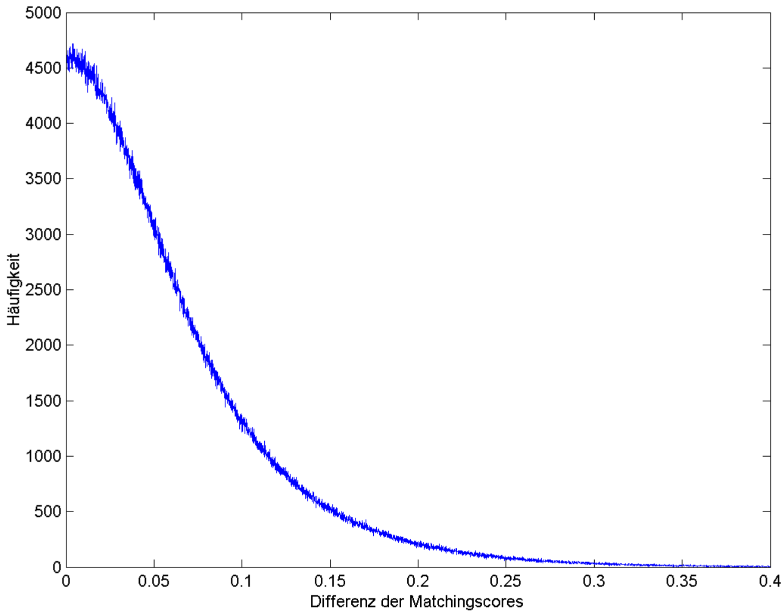


Figure 2: deviations in symmetric comparisons under 1-1 Algorithm 2

Häufigkeit	Frequency
Differenz der Matchingscores	Difference in matching scores

[9] See also 5.7.2 Notes on the interpretation of the diagrams.

In all the diagrams for this test group, the x-axis shows these normalised frequencies for the matching scores shown on the y-axis.

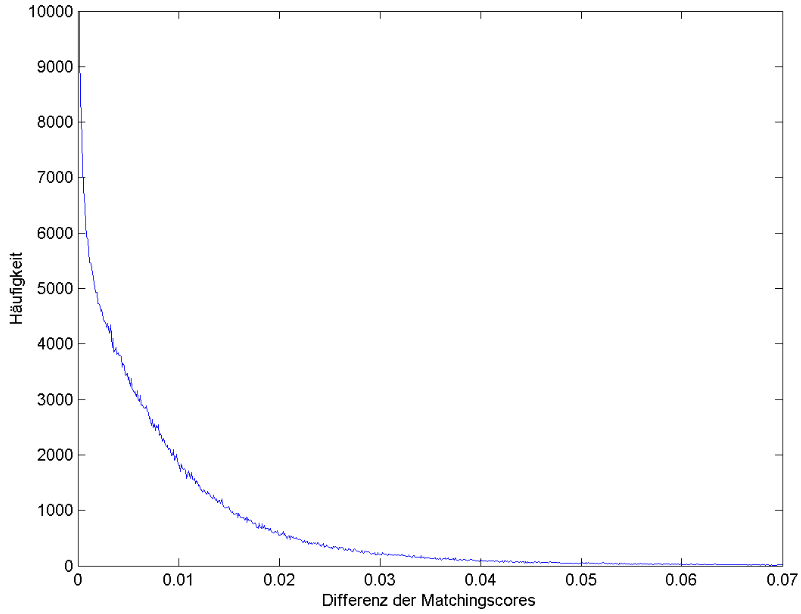


Figure 3: deviations in symmetric comparisons under 1-1 Algorithm 3

To improve comparability, the same scale for the x- and y-axes was always used on all diagrams for a given vendor. (The scaling for different vendors could differ). The scale of the x-axis was oriented towards the highest normalised frequency value that had occurred for the vendor in this test group. For the y-axis, the same matching score range from 0 to 1 was used on all the diagrams (both across tests and across vendors), even when only a part of that range was relevant. The absolute matching scores naturally depend on the vendor and algorithm and therefore are not directly comparable.

Häufigkeit	Frequency
Differenz der Matchingscores	Difference in matching scores

5.7.3.6 1-1 Algorithm 1

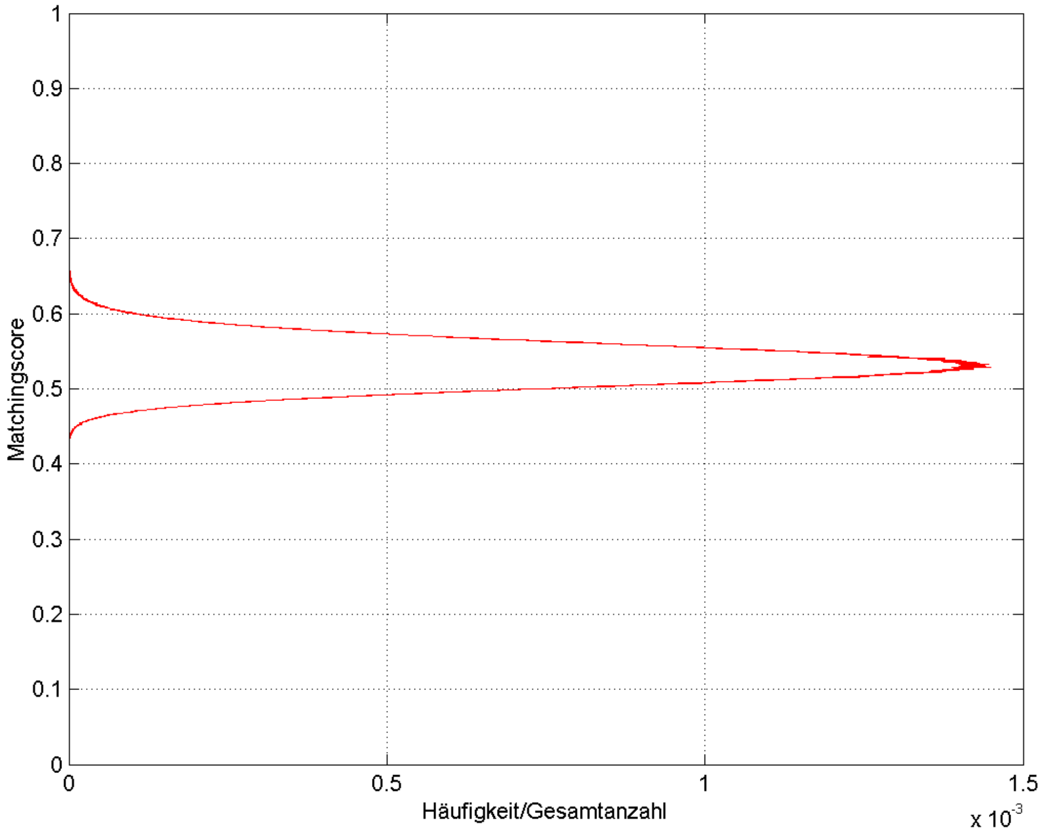


Figure 4: 1-1 Algorithm 1

Matchingscore	Matching score
Häufigkeit/Gesamtanzahl	Frequency/total number

5.7.3.7 1-1 Algorithm 2

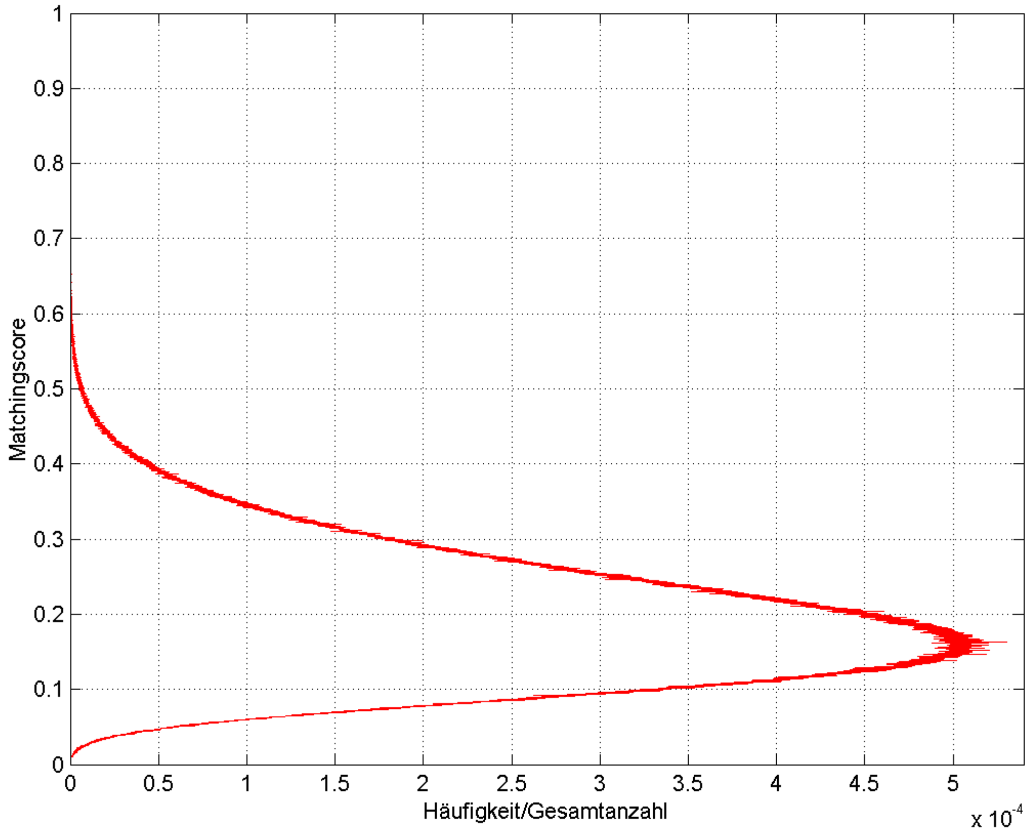


Figure 5: 1-1 Algorithm 2

Matchingscore	Matching score
Häufigkeit/Gesamtanzahl	Frequency/total number

5.7.3.8 Algorithm 3

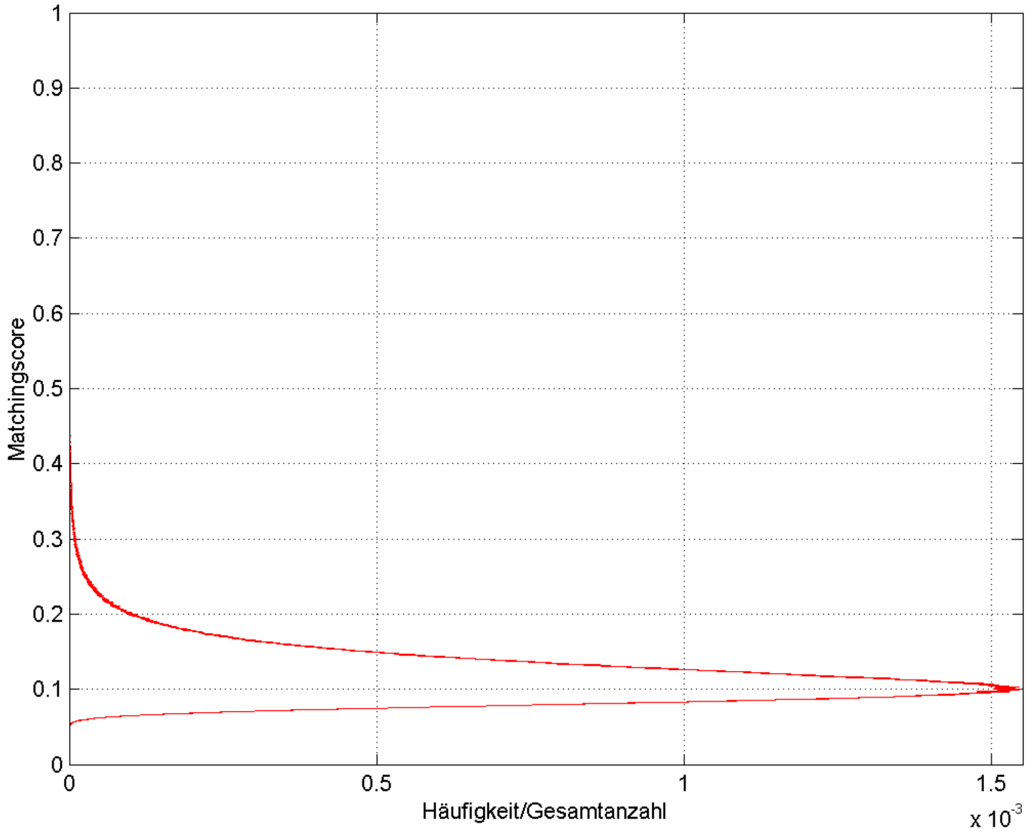


Figure 6: 1-1 Algorithm 3

Matchingscore	Matching score
Häufigkeit/Gesamtanzahl	Frequency/total number

5.7.3.9 1-2 Algorithm 1

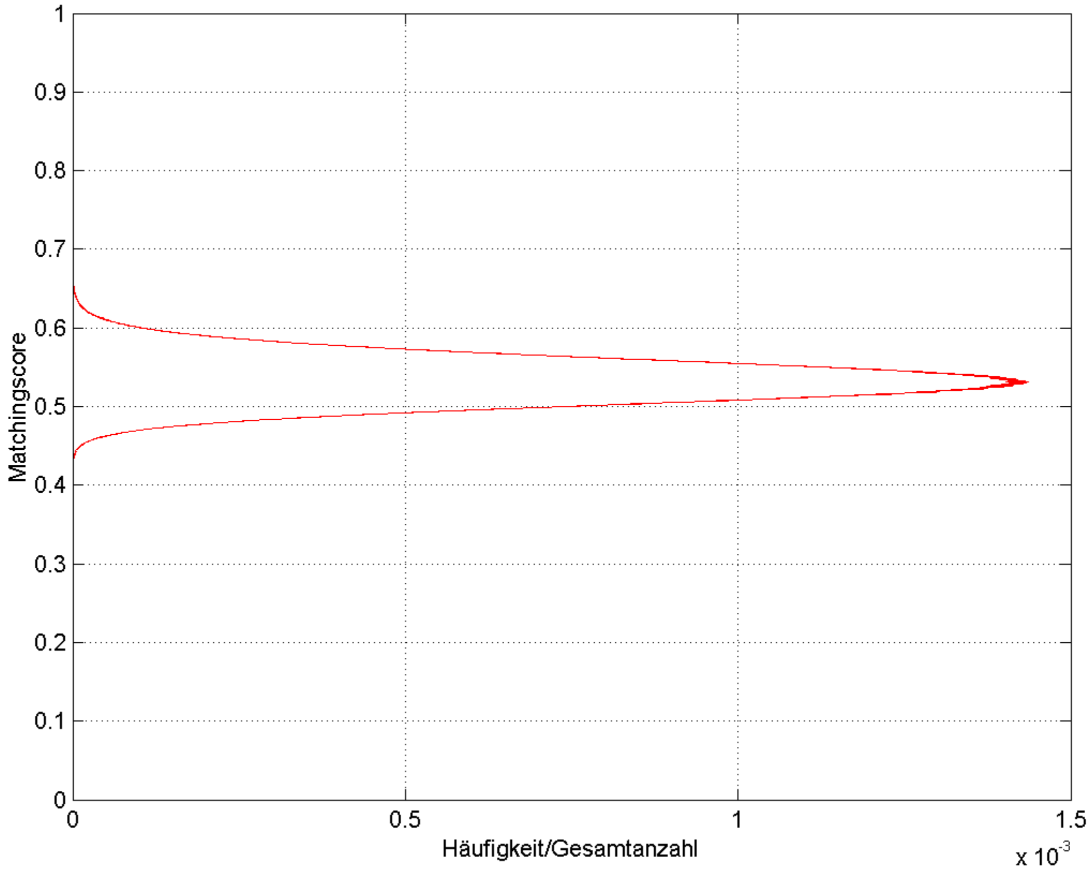


Figure 7: 1-2 Algorithm 1

Matchingscore	Matching score
Häufigkeit/Gesamtanzahl	Frequency/total number

5.7.3.10 1-2 Algorithm 2

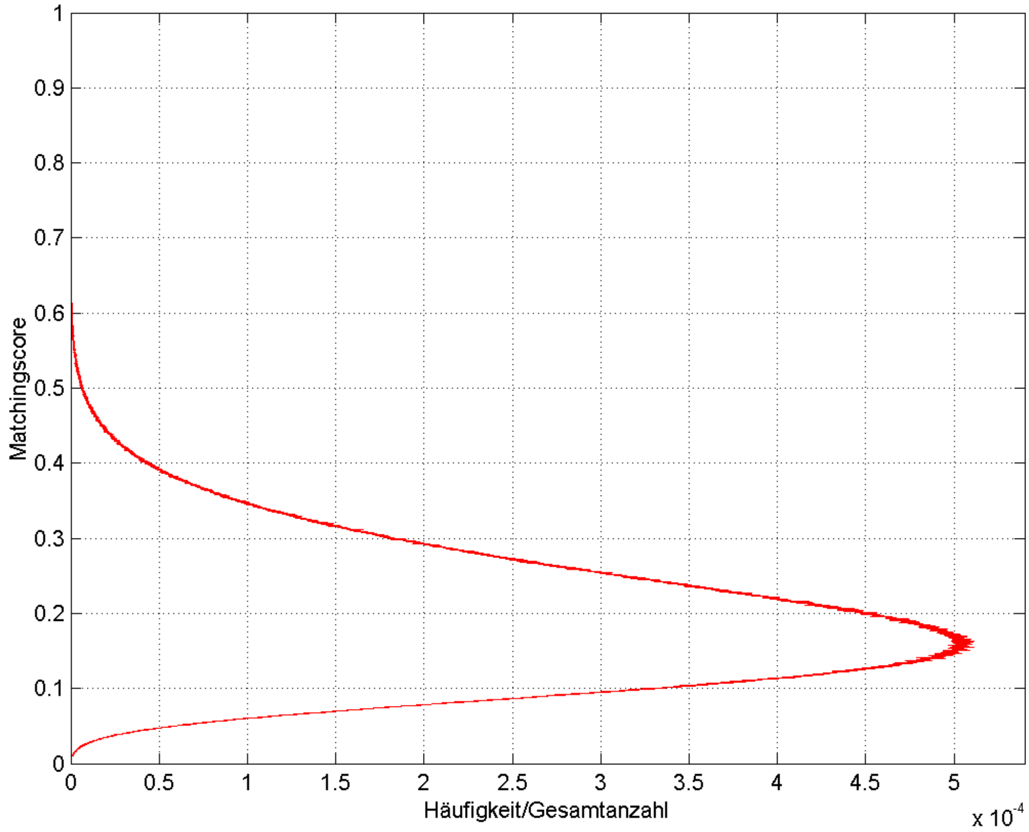


Figure 8: 1-2 Algorithm 2

Matchingscore	Matching score
Häufigkeit/Gesamtanzahl	Frequency/total number

5.7.3.11 1-2 Algorithm 3

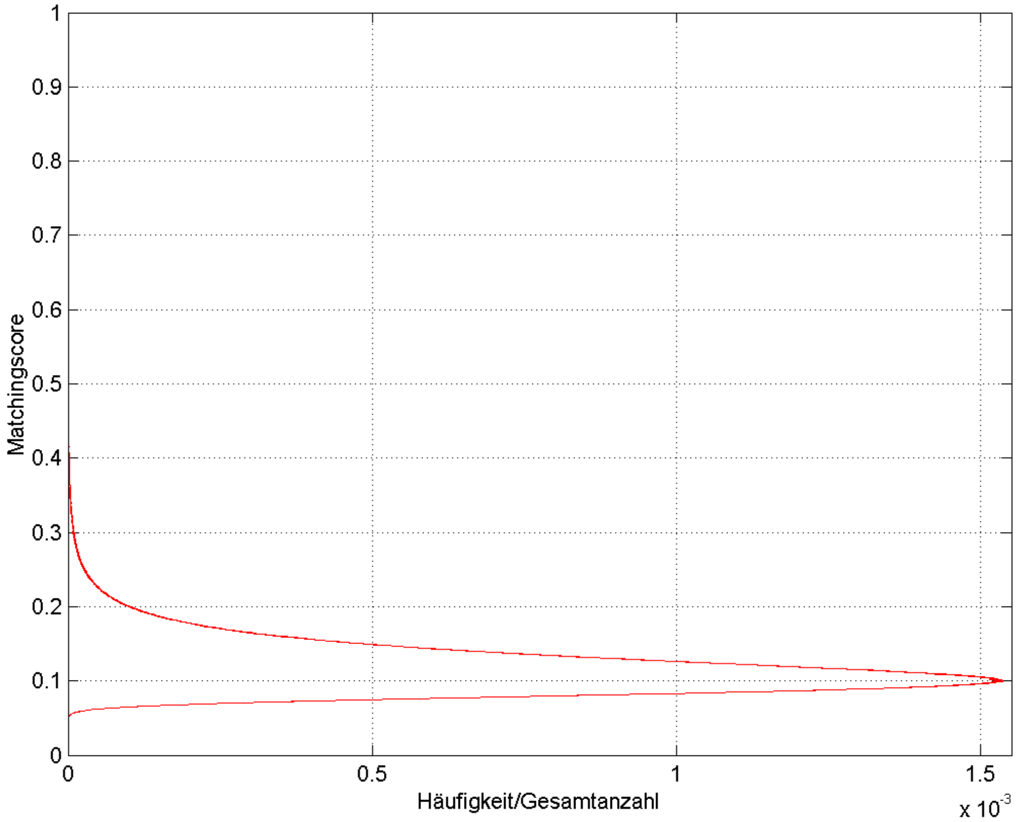


Figure 9: 1-2 Algorithm 3

Matchingscore	Matching score
Häufigkeit/Gesamtanzahl	Frequency/total number

5.7.3.12 1-3 Algorithm 1

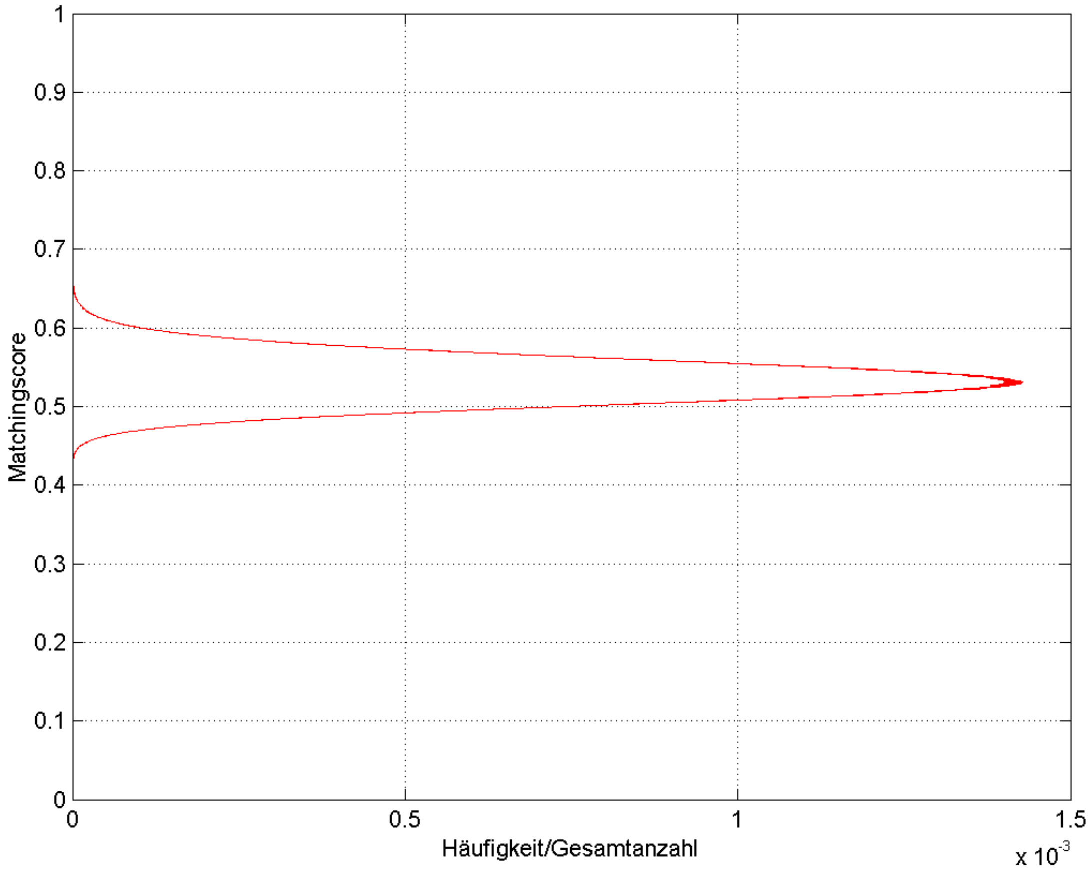


Figure 10: 1-3 Algorithm 1

Matchingscore	Matching score
Häufigkeit/Gesamtanzahl	Frequency/total number

5.7.3.13 1-3 Algorithm 2

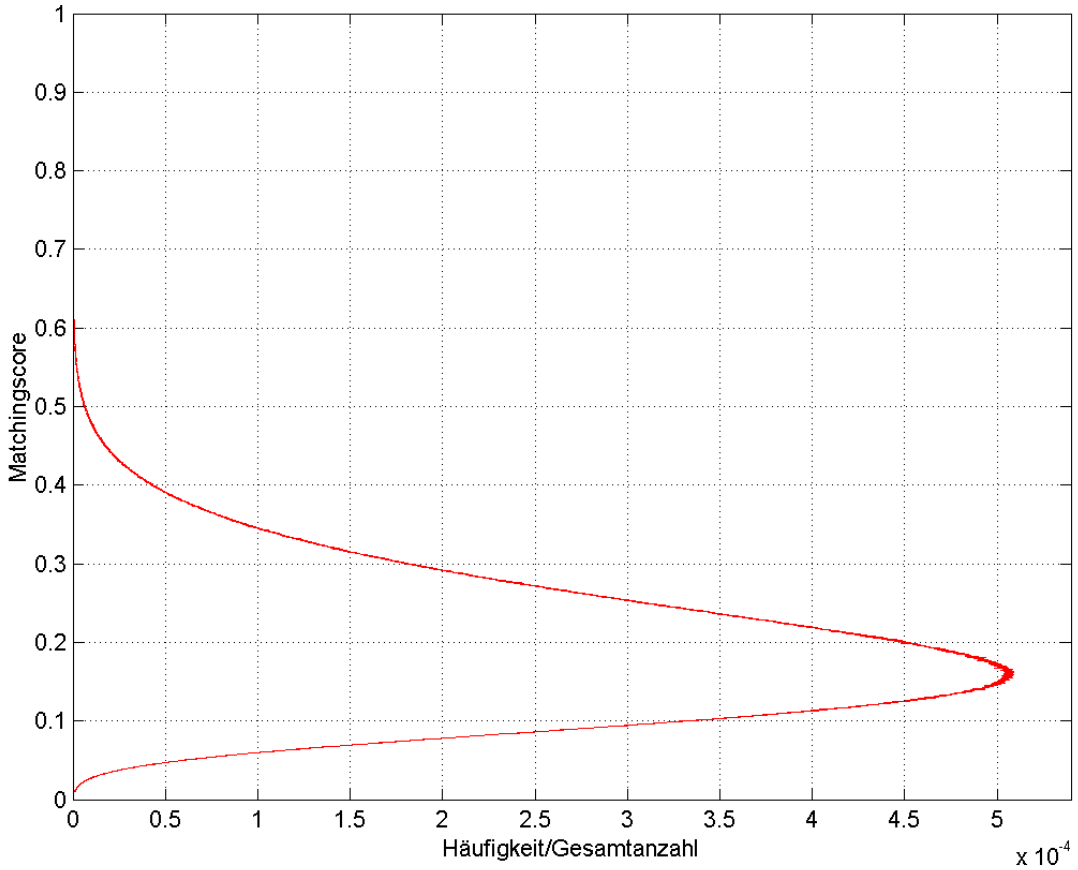


Figure 11: 1-3 Algorithm 2

Matchingscore	Matching score
Häufigkeit/Gesamtanzahl	Frequency/total number

5.7.3.14 1-3 Algorithm 3

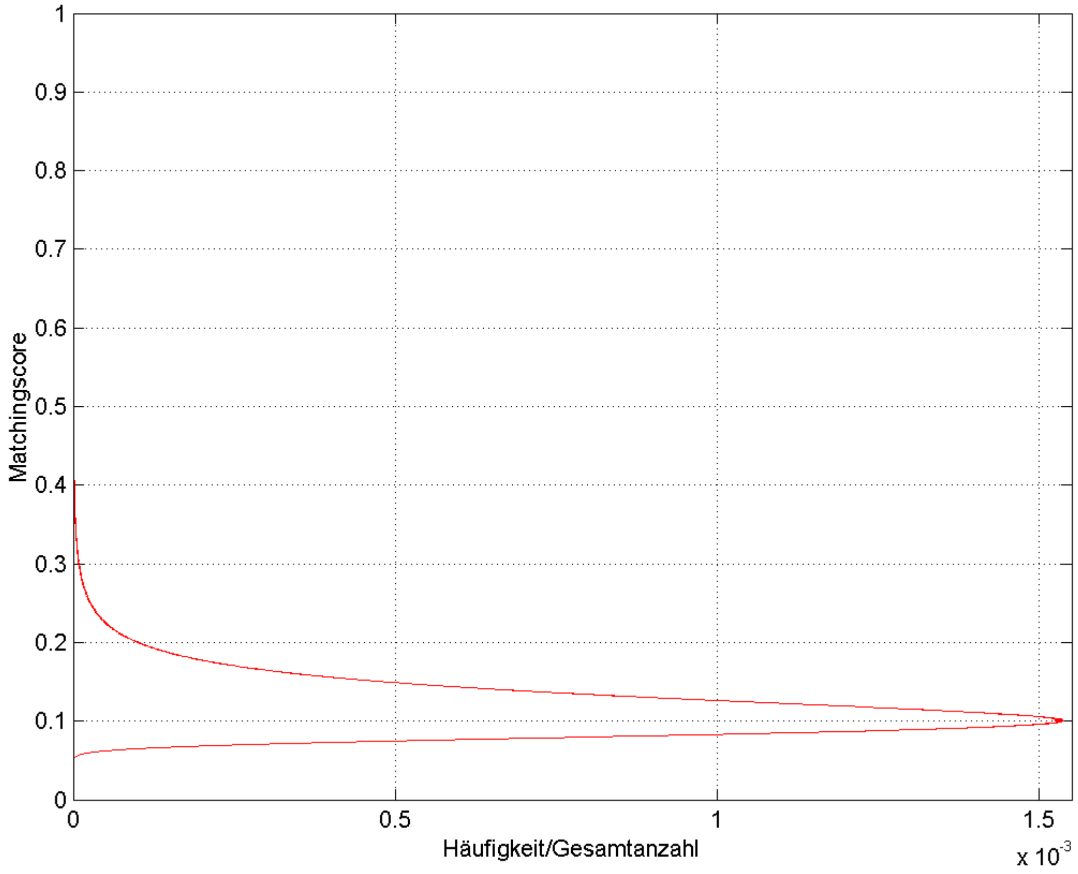


Figure 12: 1-3 Algorithm 3

Matchingscore	Matching score
Häufigkeit/Gesamtanzahl	Frequency/total number

5.7.3.15 1-4 Algorithm 1

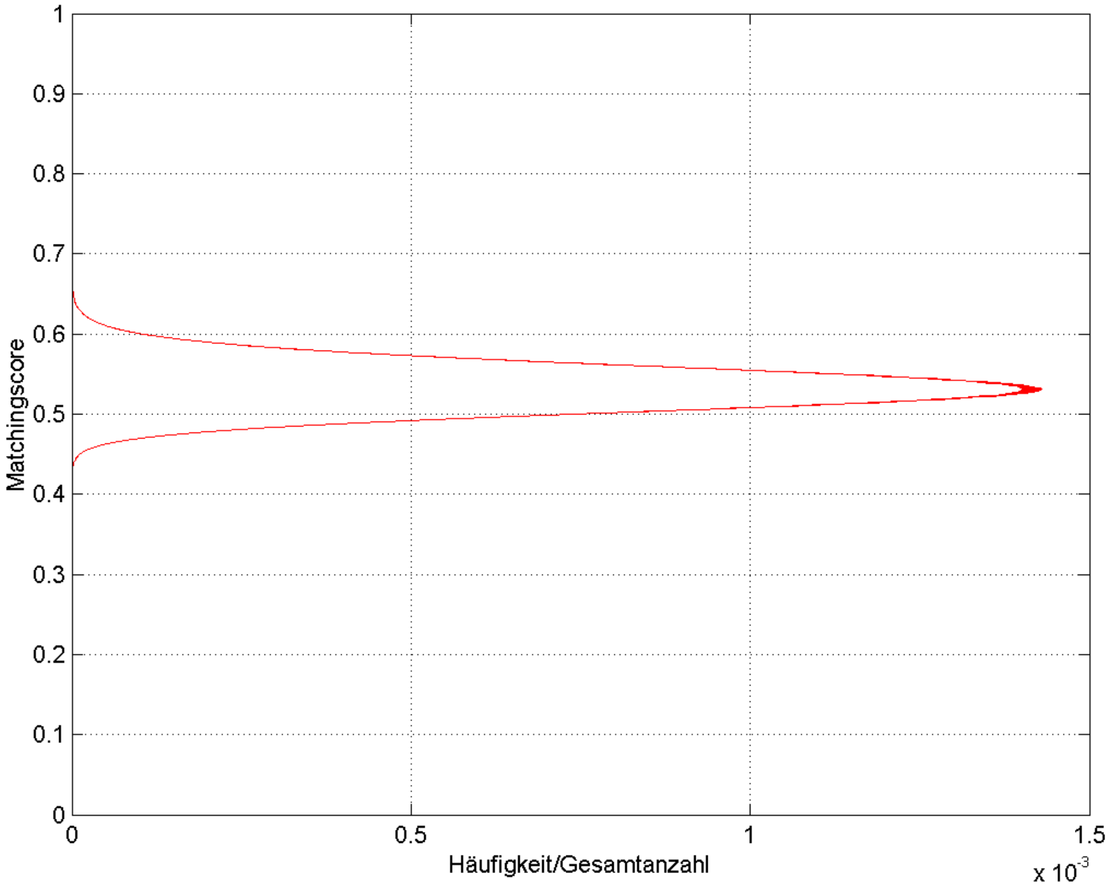


Figure 13: 1-4 Algorithm 1

Matchingscore	Matching score
Häufigkeit/Gesamtanzahl	Frequency/total number

5.7.3.16 1-4 Algorithm 2

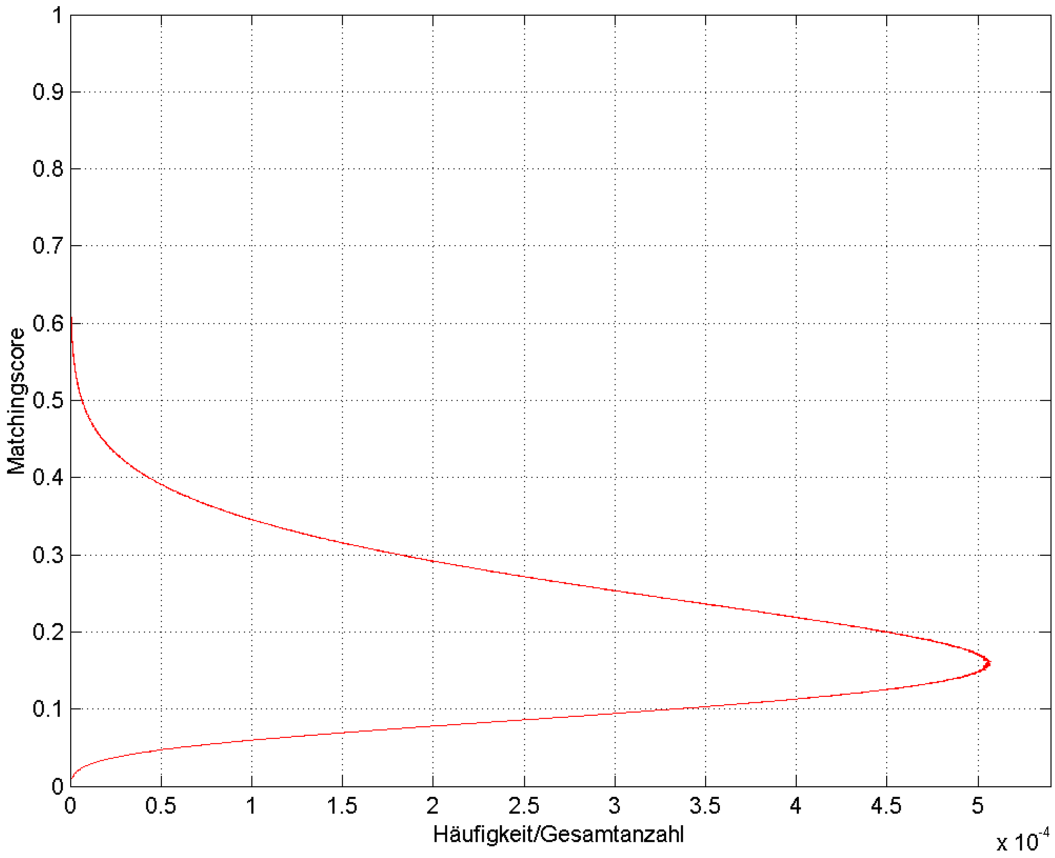


Figure 14: 1-4 Algorithm 2

Matchingscore	Matching score
Häufigkeit/Gesamtanzahl	Frequency/total number

5.7.3.17 Interpretation of results

In this group of tests, only comparisons between images of different persons were considered and investigated. The average matching scores thus represent the distribution of values for “non-hits” or “non-matches”. Increasing the size of the database could have three possible results:

- on average, a decline in values
- no change in values
- on average, an increase in values

A fall in values would be an improvement, as a non-match should achieve the smallest possible matching score, whereas a match should of course achieve the highest possible comparison value. The clearer this difference, the better a system’s recognition performance.

If on the other hand the values rise, it becomes more difficult to distinguish matches from non-matches. Genuine, correct hits would then be “submerged in the noise”.

In the case of the diagram types used here, a fall in values would suggest a downward shift of the value curve on the x-axis, whereas an upward shift would suggest a rise in values.

In this respect, direct comparison between the diagrams of different vendors is not sensible, as the absolute matching scores are neither comparable across algorithms nor normalised. Thus one can compare a score of 0.1 from vendor A, 0.34 from vendor B and 0.67 from vendor C, but as the values are algorithm-specific, it is not possible to arrive at any conclusions from this as to which system has the better recognition performance.

Even the recognition performance of one and the same system does not necessarily behave in a linear fashion, i.e. a comparison image that has achieved a comparison value of 0.6 compared with another comparison image with 0.3, is not necessarily twice as similar.

Against this background one can examine the diagrams for Tests 1-1 to 1-4 for a given vendor and observe whether any of the above changes has occurred. When one examines the data more closely, it is striking that for all manufacturers the curves are virtually identical even when superimposed on each other. Only the scatter of the frequency values declines as the number of test images rises. This is not surprising, as more comparisons were carried out, thus reducing the “noise”. This is evident in the curve extracts contained in Figure 15: change in scatter between 1-1 Algorithm 2 and 1-4- Algorithm 2.

This observation is interesting in that one can deduce from it that the number of non-matches with good and also with poor matching scores rises linearly in accordance with its distribution over the value spectrum. In other words, the frequency of occurrence simultaneously also represents the growth rate as the numbers of images and comparisons are increased.

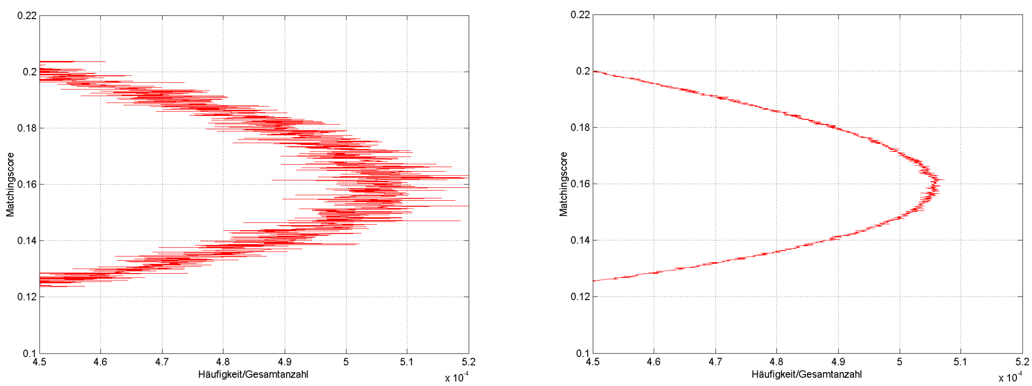


Figure 15: change in scatter between 1-1 Algorithm 2 and 1-4- Algorithm 2

Matchingscore	Matching score
Häufigkeit/Gesamtanzahl	Frequency/total number

5.7.4 Test Group 4 hunt

5.7.4.1 Comparisons and identifications

Test	Number of reference images	Algorithm 1	Algorithm 2	Algorithm 3
4-1	305	256 (83.9%)	251 (82.3%)	305 (100%)
4-2	1,305	1,192 (91.3%)	1,166 (89.3%)	1,305 (100%)
4-3	5,305	4,876 (91.9%)	4,799 (90.5%)	5,305 (100%)
4-4	50,305	46,396 (92.2%)	45,941 (91.3%)	50,305 (100%)

Table 14: enrolments

Test	Maximum number of search images	Algorithm 1	Algorithm 2	Algorithm 3
4-1	116	103 (88.7%)	108 (93.1%)	116 (100%)
4-2	116	103 (88.7%)	108 (93.1%)	116 (100%)
4-3	116	103 (88.7%)	108 (93.1%)	116 (100%)
4-4	116	103 (88.7%)	108 (93.1%)	116 (100%)

Table 15: accepted identifications Test Group 4

Test	Maximum number of identification results	Algorithm 1	Algorithm 2	Algorithm 3
4-1	1160	1030 (88.7%)	176[10] (15.2%)	1160 (100%)
4-2	1160	1030 (88.7%)	412[10] (36.3%)	1160 (100%)
4-3	1160	1030 (88.7%)	644[10] (55.5%)	1160 (100%)
4-4	1160	1030 (88.7%)	927[10] (79.9%)	1160 (100%)

Table 16: number of identification results

5.7.4.2 Scoring according to key figures

To gain a simple and rapid comparison of performance, two key figures were calculated from the results.

Key Figure 1: $$\frac{\text{Number of images found}}{\text{Number of images contained in the reference set}}$$

Key Figure 1 takes into account how many of the available images were found. The position within the results data record is irrelevant here. If the reference set contained more than ten images of the same person, then the denominator was limited to 10, since, because each

[10] The number of results returned deviates under Algorithm 2 from the formula *Number of accepted images * 10 return values*, as the required number of results (10) was not always returned. See section 5.7.1.2 Algorithm 2 on page 24.

results data record contained ten results, this was the maximum number of images that could be found.

Test	Maximum	Algorithm 1	Algorithm 2	Algorithm 3
4-1	115[11]	63.8	15.6	63.3
4-2	115[11]	52.9	15.4	55.3
4-3	115[11]	42.6	14.6	38.3
4-4	115[11]	33.7	13.0	25.4

Table 17: results Key Figure 1

Key Figure 2: $$\frac{\text{Number of images found} \times w_i}{\text{Number of images contained in the reference set}}$$

whereby w_i = (10, 9, …, 1) signify the weights and i refers to the position within the results data record (1,2,. .., 10).

Key Figure 2 is based on Key Figure 1, however in this case additional weights have been multiplied for the items in the results data record. For position 1, weight 10 applies, for position 2 weight 9 applies etc.

Test	Maximum	Algorithm 1	Algorithm 2	Algorithm 3
4-1	2,454	1,047.0	195.9	1,016.7
4-2	2,454	817.7	186.0	777.9
4-3	2,454	649.0	177.5	525.2
4-4	2,454	493.9	138.1	331.8

Table 18: results Key Figure 2

[11] The maximum deviates here from the number of persons (116), as there were no reference images for one person.

Value range	Algorithm 1	Algorithm 2	Algorithm 3
0-0.1	38	88	33
0.1-0.2	0	3	1
0.2-0.3	1	1	3
0.3-0.4	4	7	6
0.4-0.5	0	1	1
0.5-0.6	13	4	12
0.6-0.7	7	2	10
0.7-0.8	1	0	0
0.8-0.9	3	0	7
0.9-1	48	9	42

Table 19: distribution of results Key Figure 1 for Test 4-1

Value range	Algorithm 1	Algorithm 2	Algorithm 3
0-0.1	47	88	40
0.1-0.2	1	3	1
0.2-0.3	3	1	3
0.3-0.4	4	7	5
0.4-0.5	1	2	1
0.5-0.6	11	4	18
0.6-0.7	5	1	8
0.7-0.8	5	0	1
0.8-0.9	0	0	2
0.9-1	38	9	36

Table 20: distribution of results Key Figure 1 for Test 4-2

Value range	Algorithm 1	Algorithm 2	Algorithm 3
0-0.1	56	90	58
0.1-0.2	2	3	2
0.2-0.3	4	2	2
0.3-0.4	5	5	9
0.4-0.5	1	1	3
0.5-0.6	9	4	12
0.6-0.7	5	1	2
0.7-0.8	5	0	3

0.8-0.9	0	0	1
0.9-1	28	9	23

Table 21: distribution of results Key Figure 1 for Test 4-3

Value range	Algorithm 1	Algorithm 2	Algorithm 3
0-0.1	67	93	77
0.1-0.2	1	3	4
0.2-0.3	3	3	0
0.3-0.4	5	3	4
0.4-0.5	4	1	4
0.5-0.6	8	3	8
0.6-0.7	2	0	1
0.7-0.8	3	0	0
0.8-0.9	0	0	0
0.9-1	22	9	17

Table 22: distribution of results Key Figure 1 for Test 4-4

Value range	Algorithm 1	Algorithm 2	Algorithm 3
0-5	49	97	50
5-10	15	6	12
10-15	29	11	29
15-20	8	0	10
20-25	2	0	3
25-30	4	1	7
30-35	2	0	2
35-40	2	0	1
40-45	3	0	1
45-50	1	0	0
50-55	0	0	0

Table 23: distribution of results Key Figure 2 for Test 4-1

Value range	Algorithm 1	Algorithm 2	Algorithm 3
0-5	61	98	56
5-10	13	7	20
10-15	24	9	23
15-20	5	0	8
20-25	2	0	4
25-30	4	1	1
30-35	6	0	2
35-40	0	0	0
40-45	0	0	1
45-50	0	0	0
50-55	0	0	0

Table 24: distribution of results Key Figure 2 for Test 4-2

Value range	Algorithm 1	Algorithm 2	Algorithm 3
0-5	72	100	75
5-10	6	6	13
10-15	25	8	16
15-20	4	0	8
20-25	0	0	2
25-30	5	1	1
30-35	3	0	0
35-40	0	0	0
40-45	0	0	0
45-50	0	0	0
50-55	0	0	0

Table 25: distribution of results Key Figure 2 for Test 4-3

Value range	Algorithm 1	Algorithm 2	Algorithm 3
0-5	75	103	83
5-10	14	4	14
10-15	18	7	14
15-20	4	1	4
20-25	1	0	0
25-30	3	0	0
30-35	0	0	0
35-40	0	0	0
40-45	0	0	0
45-50	0	0	0
50-55	0	0	0

Table 26: distribution of results Key Figure 2 for Test 4-4

Detailed tables by Person ID are contained in the annex (see Annex A Additional results for Test Group 4 on page 126).

5.7.4.3 Notes on the diagrams[12]

Seven diagrams have been prepared for each test. The first three in each case present the hits returned for each of the 116 identifications carried out. Every correct hit (the right person identified) is represented by a green star, every false hit (wrong person identified) by a red cross.

The first diagram presents these values all in a single diagram. Ideally, it would be possible to see 1,160 symbols, i.e. 160 identifications each with 10 hits returned[13]. However, certain images were not accepted as search image, in which case the relevant column is then empty. The next two diagrams then show the correct and false hits once again, but this time separately.

The next group of graphs (4, 5 and 6) total up the matching scores that have occurred. Each matching score was rounded to two decimal places and the frequencies of occurrence were counted separately for correct and false hits. The values were normalised across the resulting total number of correct or false hits. Once again, the first diagram in the group of four presents the two curves together, the two next curves showed the individual curves on their own.

The fourth and last diagram shows the relationship between correct and false hits, known as “receiver operating characteristics”.

[12] See also 5.7.2 Notes on the interpretation of the diagrams.

[13] The number of results returned deviates under Algorithm 2 from the formula *Number of accepted images * 10 return values*, as the required number of results (10) was not always returned. See section 5.7.1.2 Algorithm 2 on page 24.

5.7.4.4 4-1 Algorithm 1

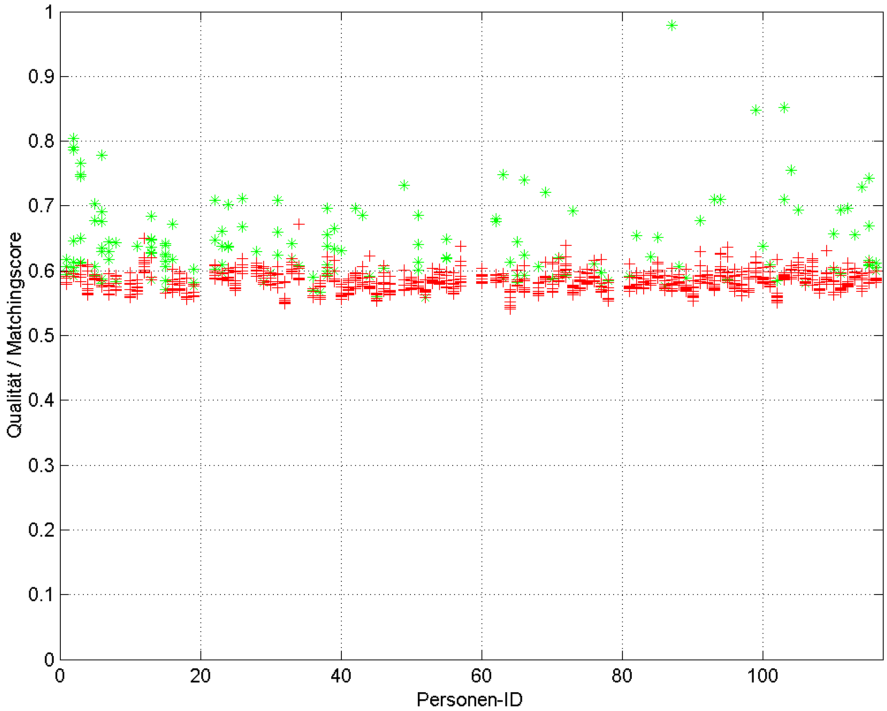


Figure 16: 4-1 Algorithm 1 correct and false hits

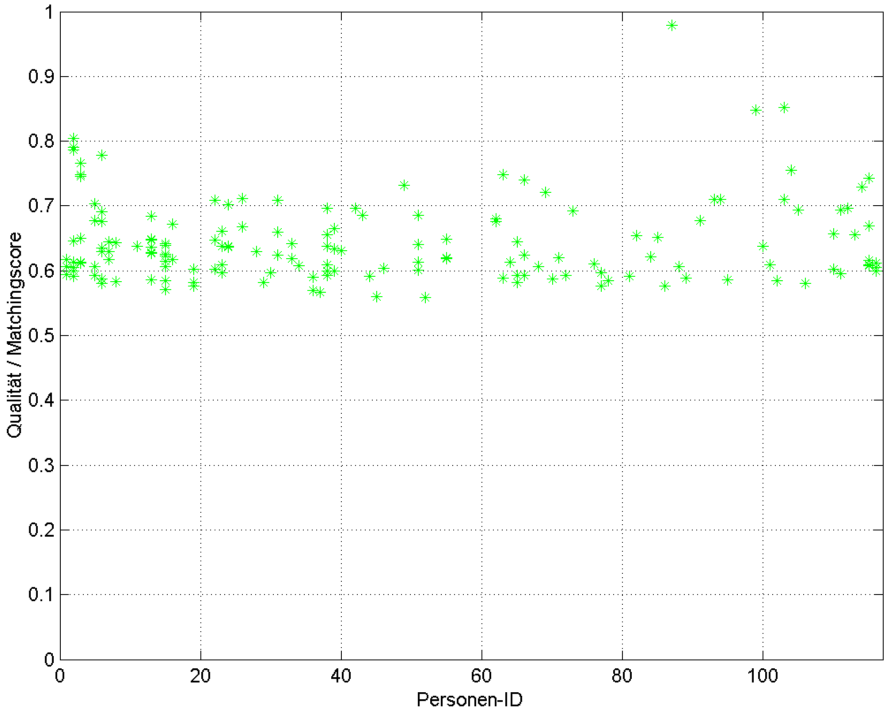


Figure 17: 4-1 Algorithm 1 correct hits

Qualität/Matchingscore	Quality/matching score
Personen-ID	Person ID

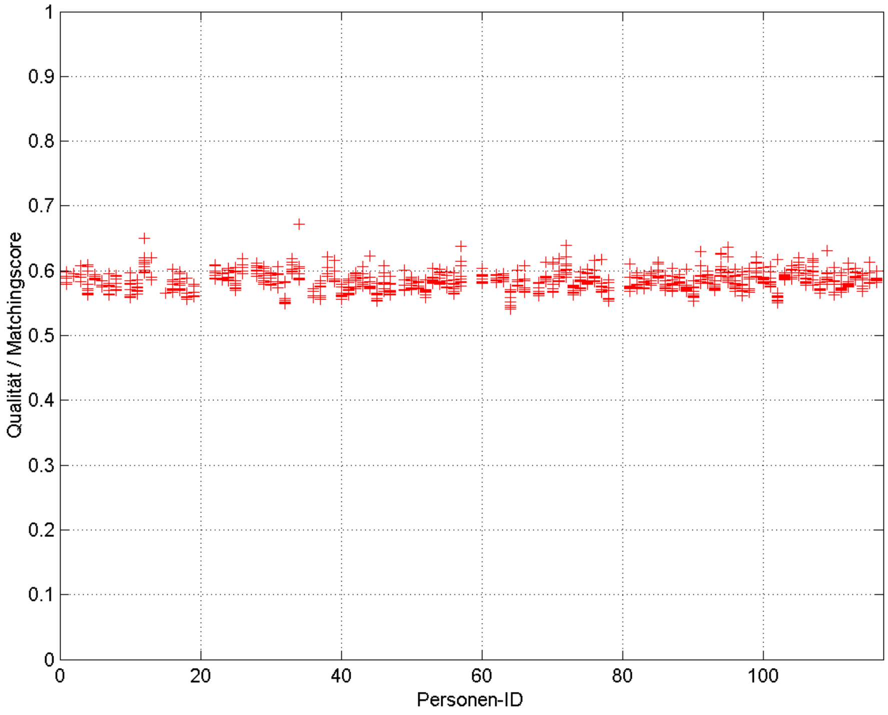


Figure 18: 4-1 Algorithm 1 false hits

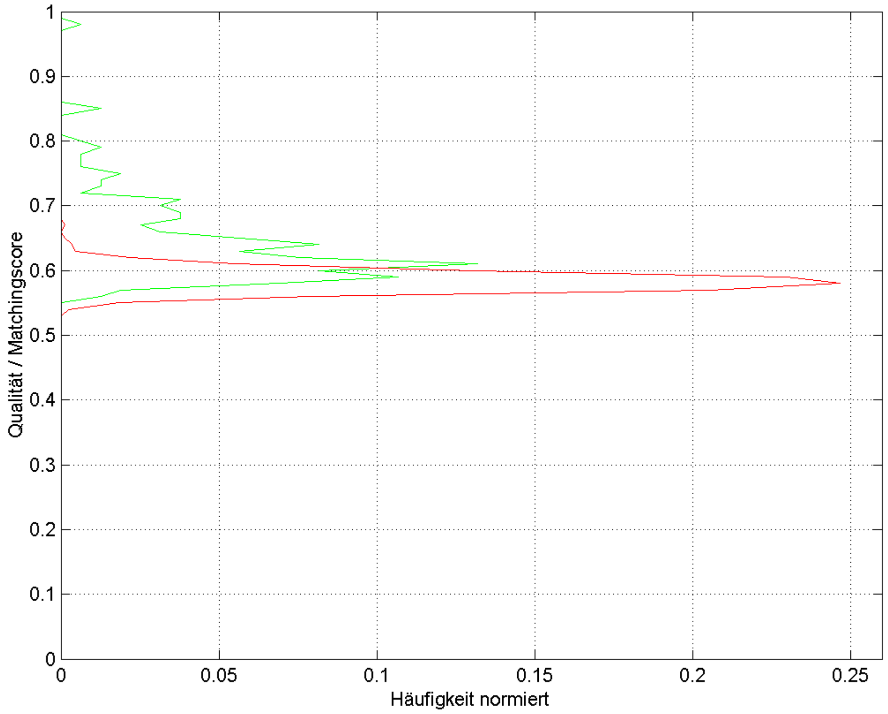


Figure 19: 4-1 Algorithm 1 total correct and false hits

Qualität/Matchingscore	Quality/matching score
Personen-ID	Person ID
Häufigkeit normiert	Normalised frequency

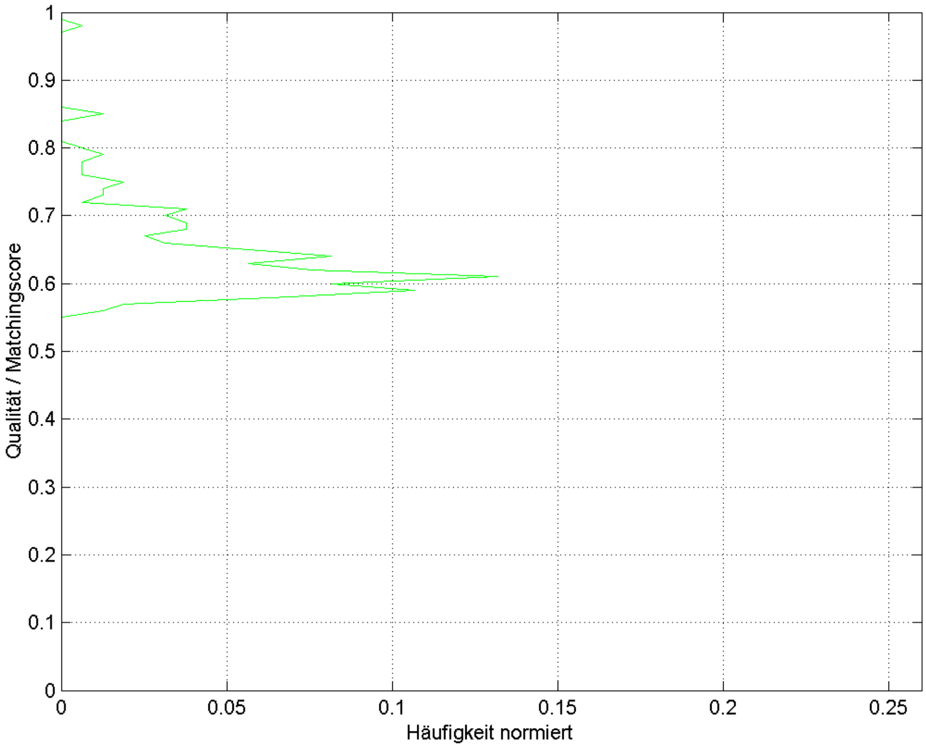


Figure 20: 4-1 Algorithm 1 total correct hits

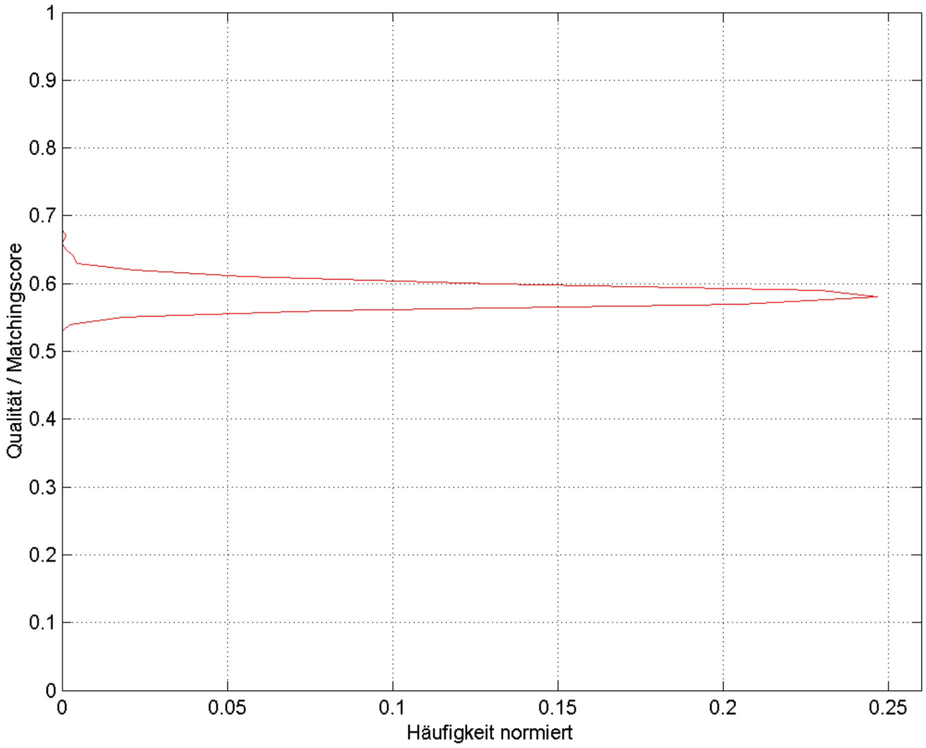


Figure 21: 4-1 Algorithm 1 total false hits

Qualität/Matchingscore	Quality/matching score
Häufigkeit normiert	Normalised frequency

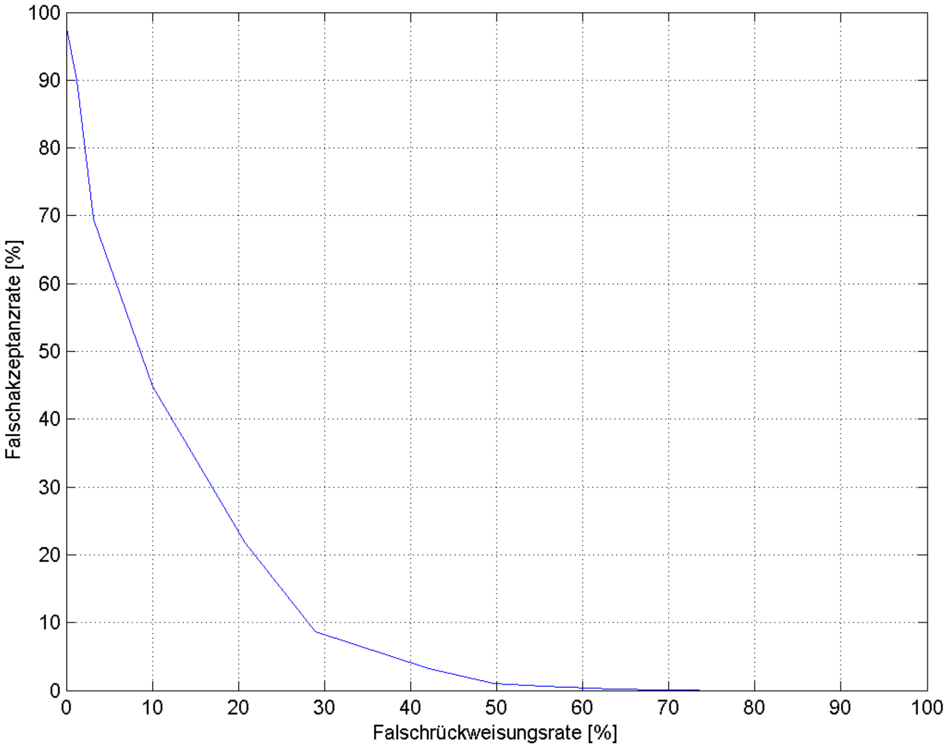


Figure 22: 4-1 Algorithm 1 receiver operating characteristics

Falschakzeptanzrate [%]	False acceptance rate [%]
Falschrückweisungsrate [%]	False rejection rate [%]

5.7.4.5 4-2-Algorithm 1

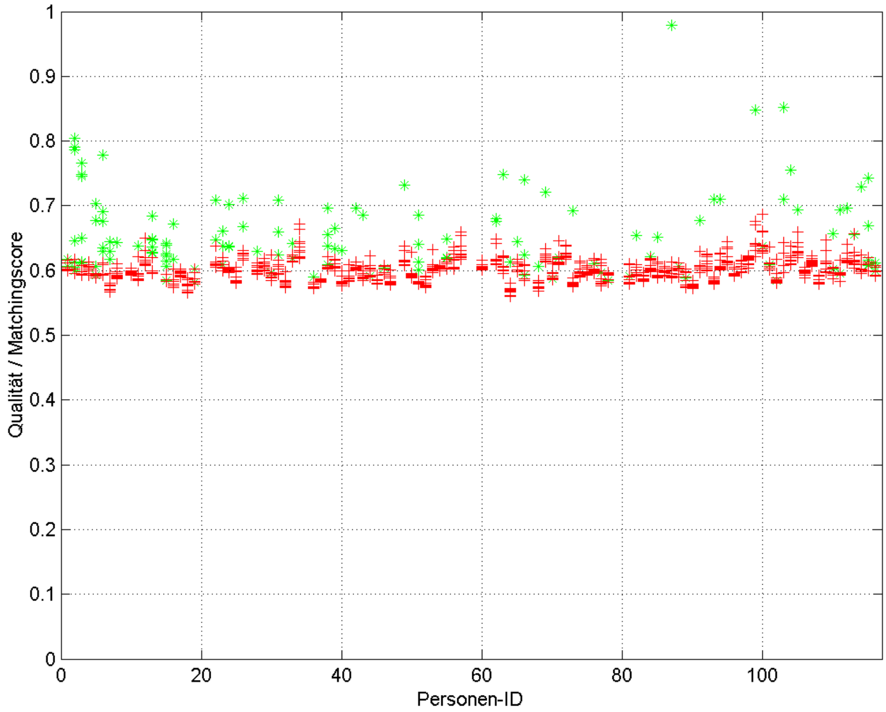


Figure 23: 4-2 Algorithm 1 correct and false hits

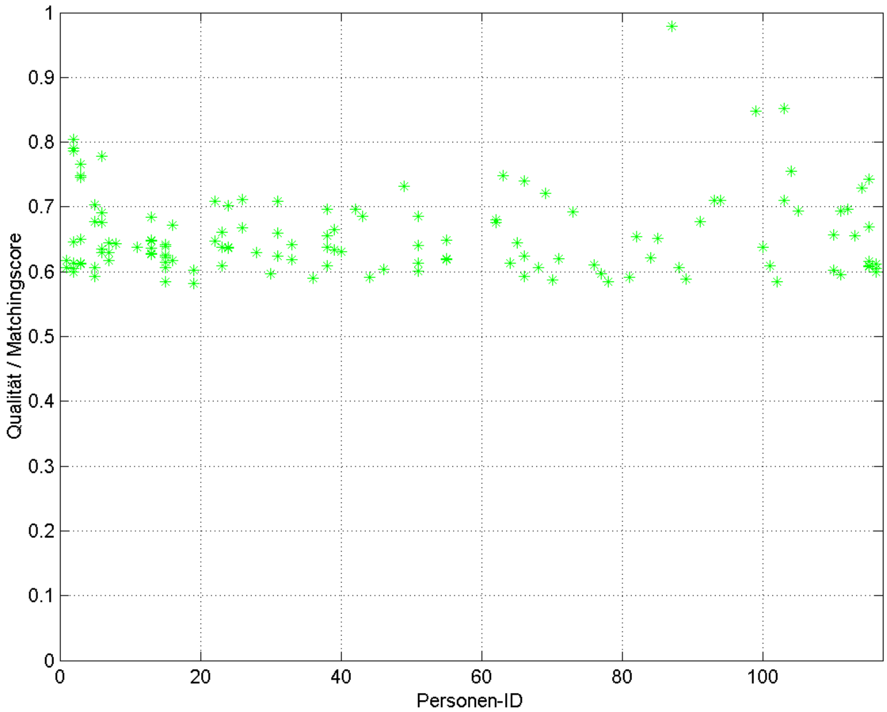


Figure 24: 4-2 Algorithm 1 correct hits

Qualität/Matchingscore	Quality/matching score
Personen-ID	Person ID

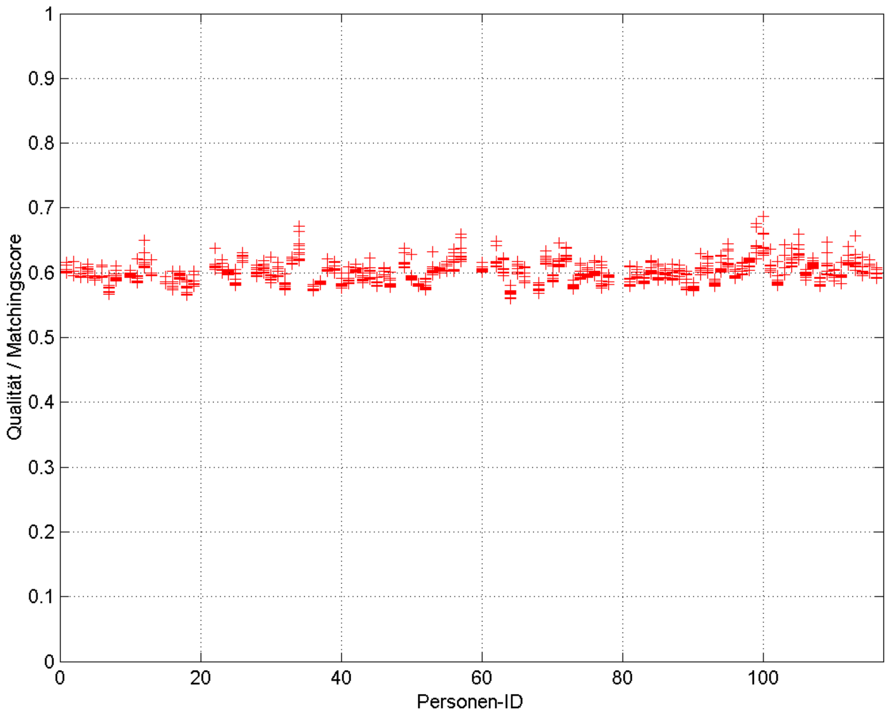


Figure 25: 4-2 Algorithm 1 false hits

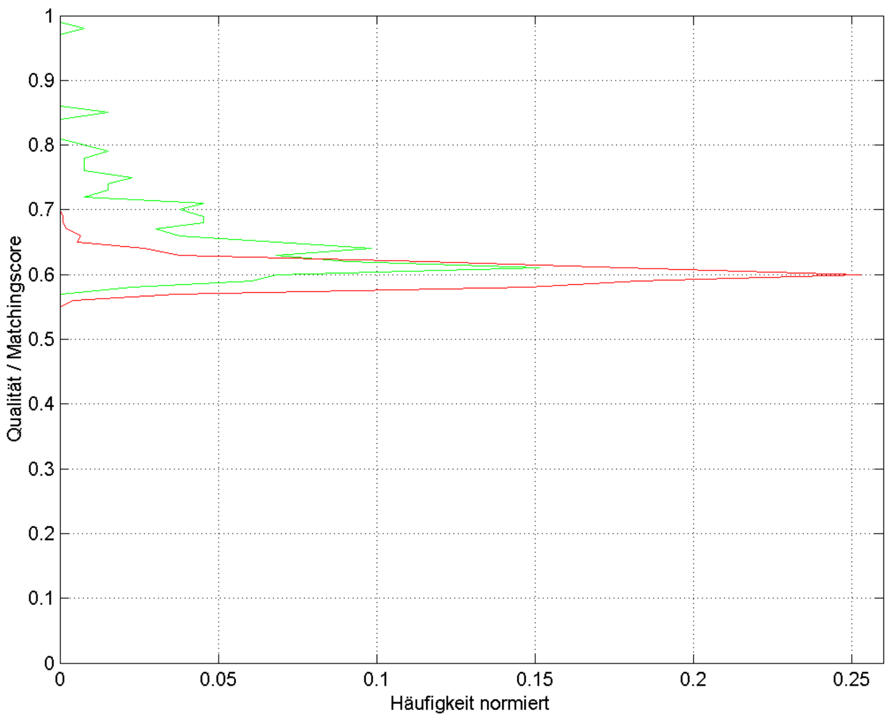


Figure 26: 4-2 Algorithm 1 total correct and false hits

Qualität/Matchingscore	Quality/matching score
Personen-ID	Person ID
Häufigkeit normiert	Normalised frequency

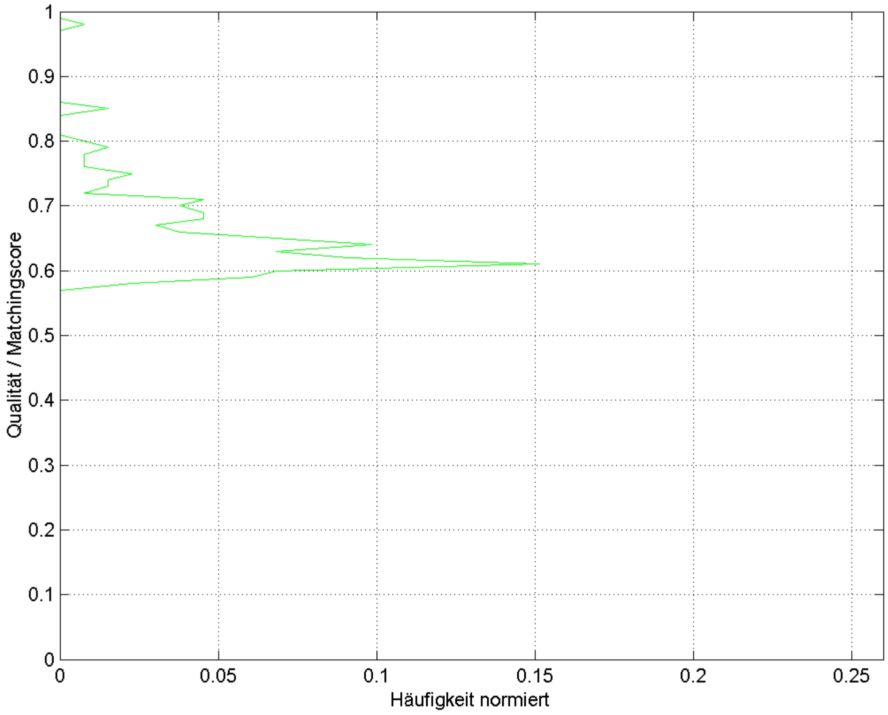


Figure 27: 4-2 Algorithm 1 total correct hits

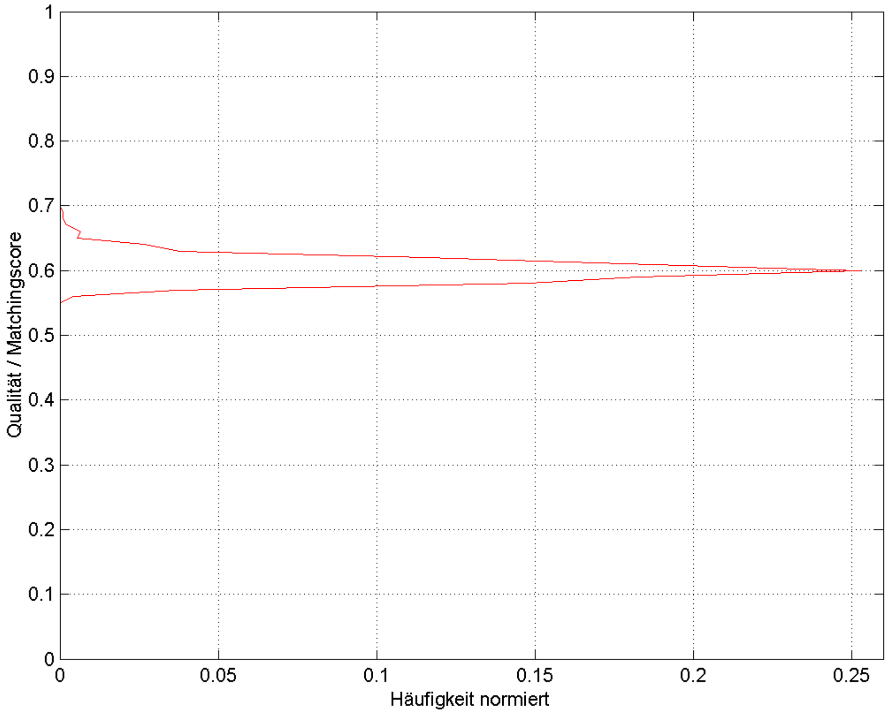


Figure 28: 4-2 Algorithm 1 total false hits

Qualität/Matchingscore	Quality/matching score
Häufigkeit normiert	Normalised frequency

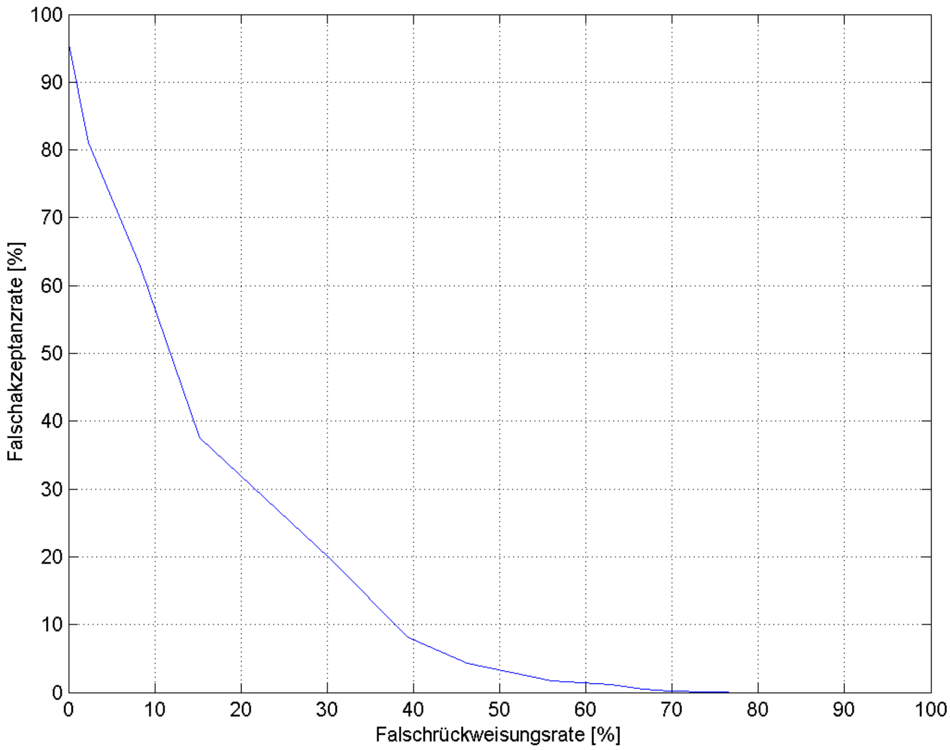


Figure 29: 4-2 Algorithm 1 receiver operating characteristics

Falschakzeptanzrate [%]	False acceptance rate [%]
Falschrückweisungsrate [%]	False rejection rate [%]

5.7.4.6 4-3-Algorithm 1

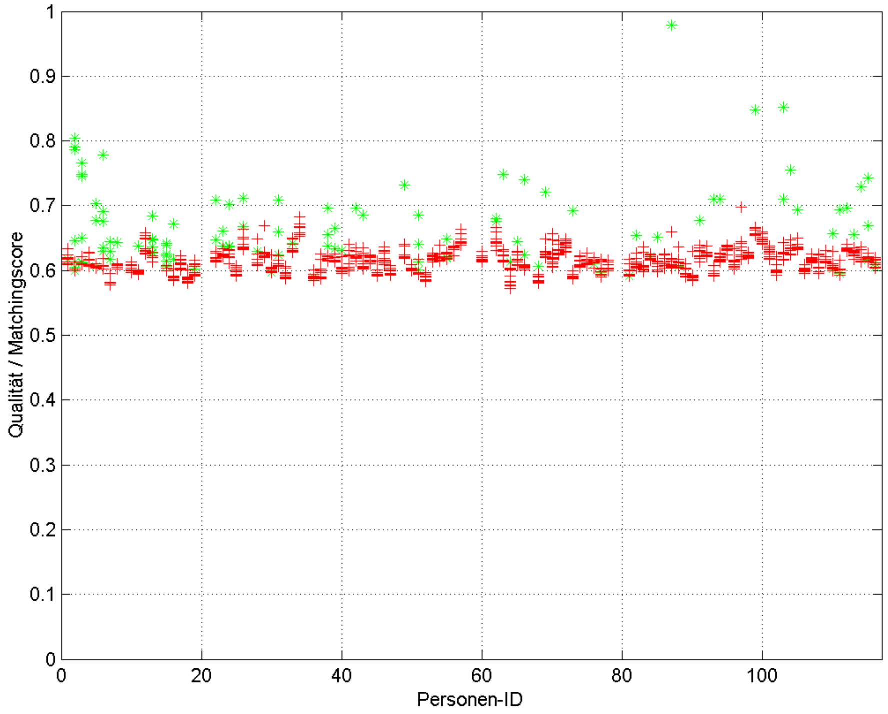


Figure 30: 4-3 Algorithm 1 correct and false hits

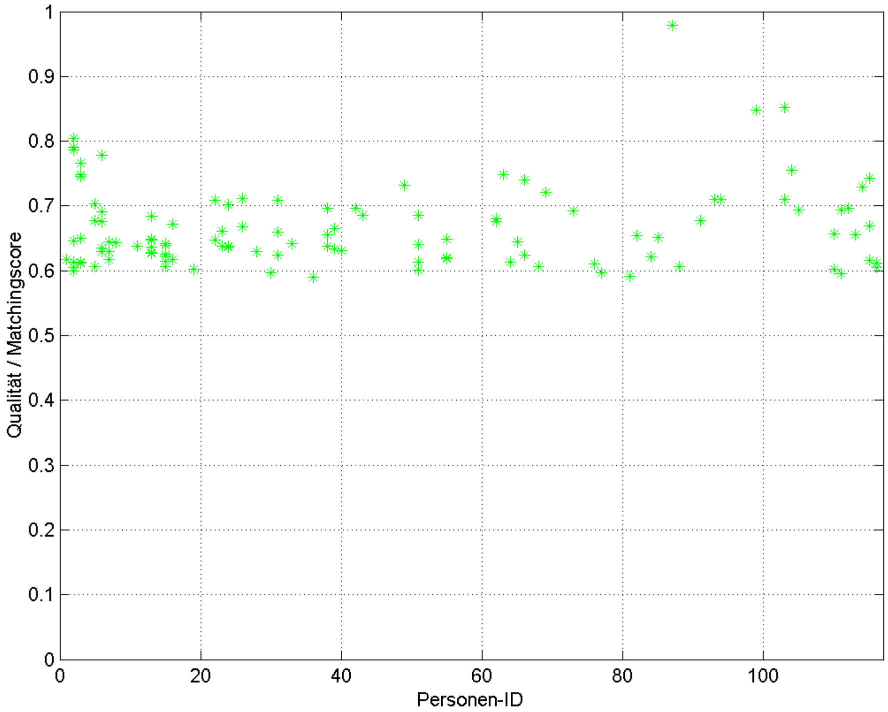


Figure 31: 4-3 Algorithm 1 correct hits

Qualität/Matchingscore	Quality/matching score
Personen-ID	Person ID

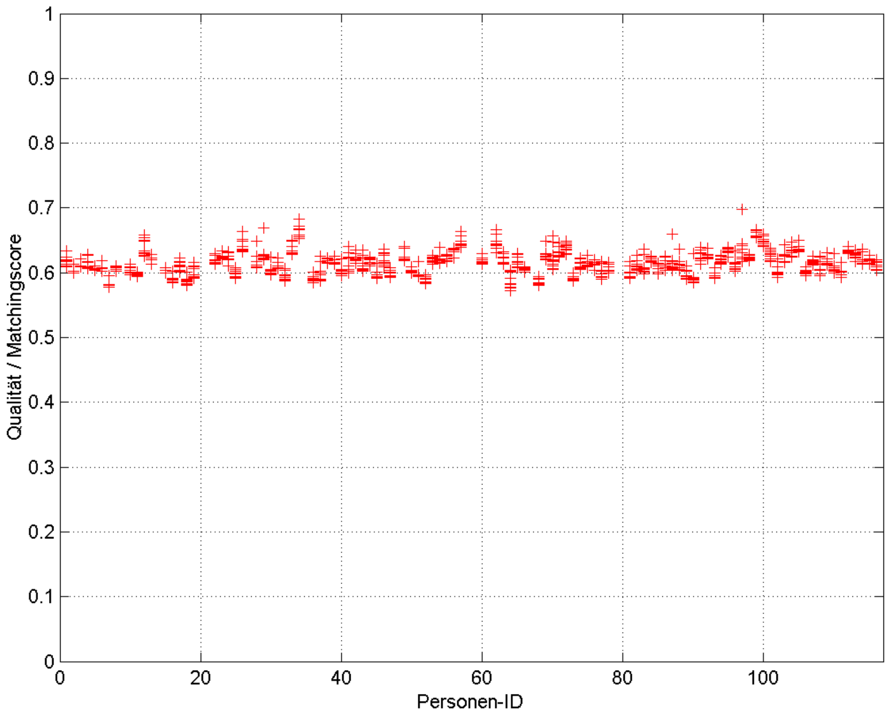


Figure 32: 4-3 Algorithm 1 false hits

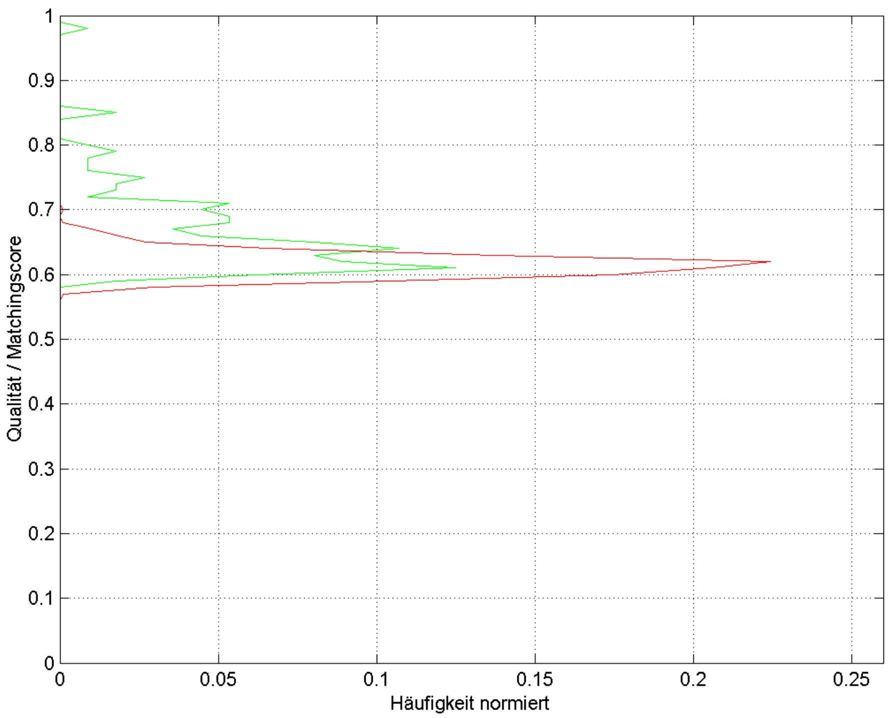


Figure 33: 4-3 Algorithm 1 total correct and false hits

Qualität/Matchingscore	Quality/matching score
Personen-ID	Person ID
Häufigkeit normiert	Normalised frequency

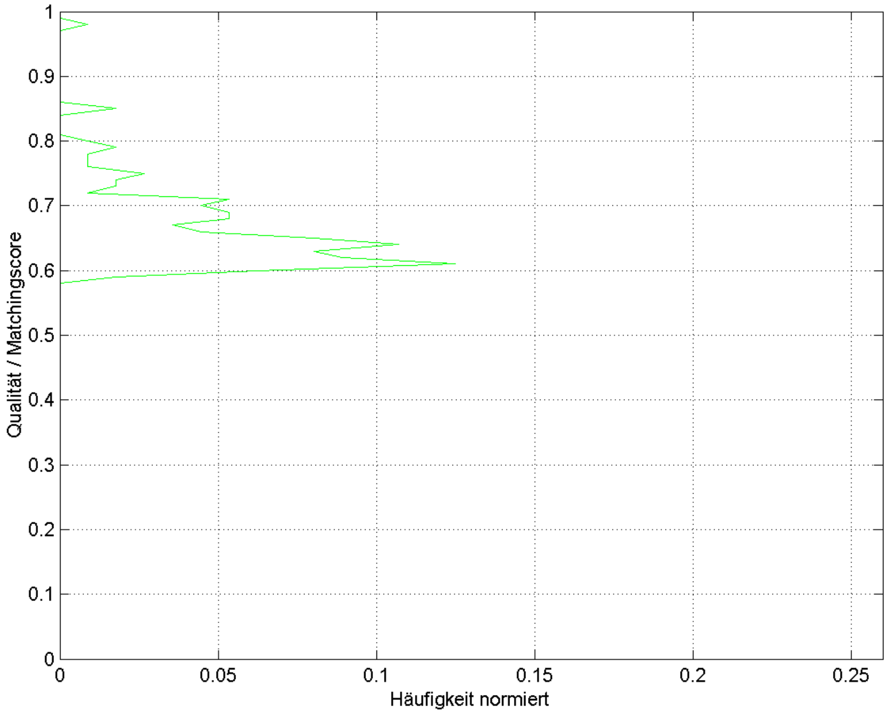


Figure 34: 4-3 Algorithm 1 total correct hits

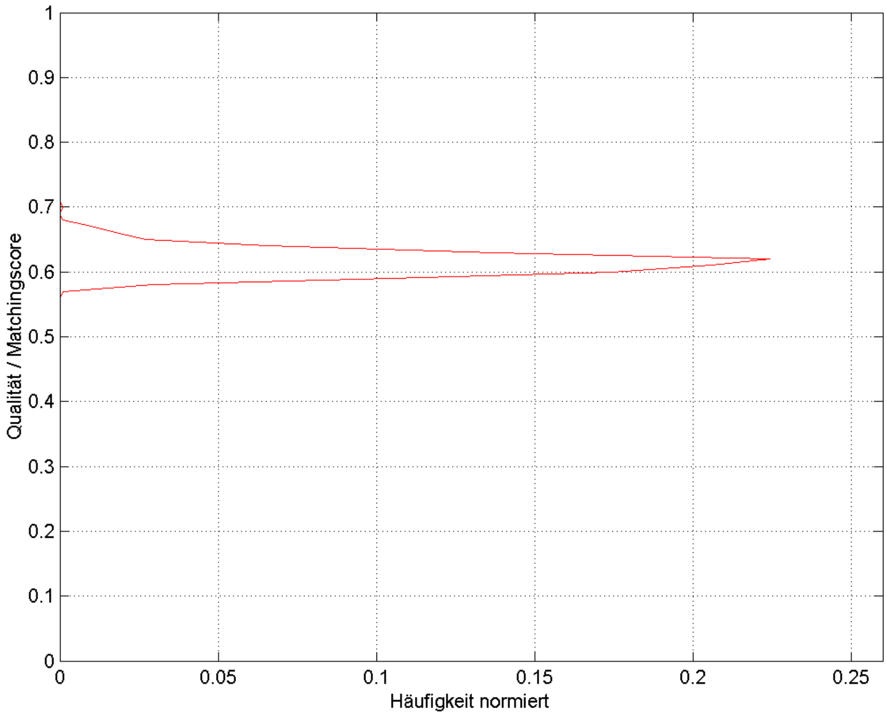


Figure 35: 4-3 Algorithm 1 total false hits

Qualitat/Matchingscore	Quality/matching score
Häufigkeit normiert	Normalised frequency

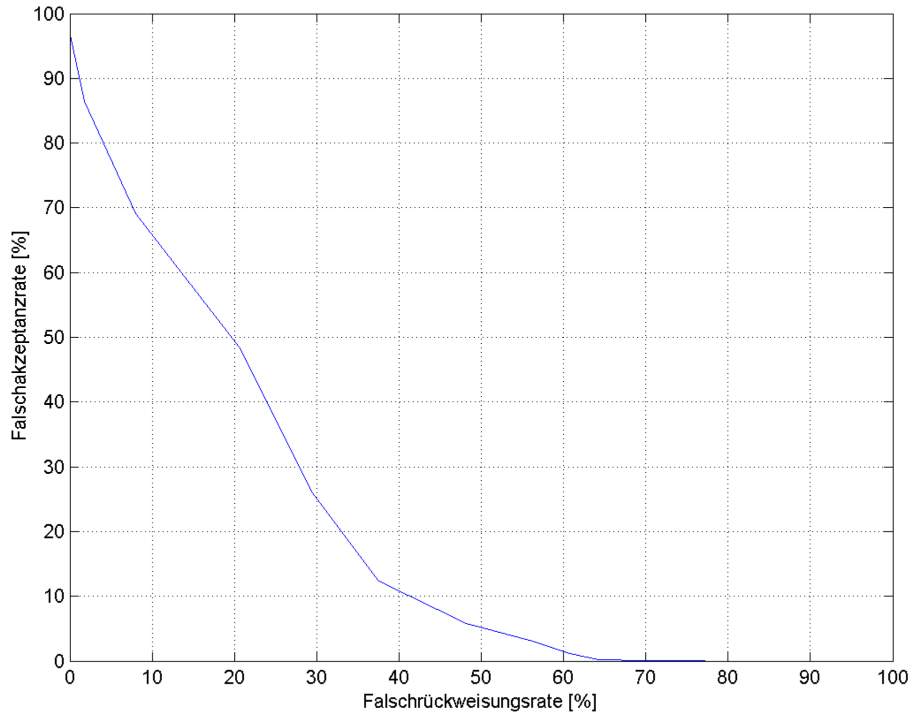


Figure 36: 4-3 Algorithm 1 receiver operating characteristics

Falschakzeptanzrate [%]	False acceptance rate [%]
Falschrückweisungsrate [%]	False rejection rate [%]

5.7.4.7 4-4-Algorithm 1

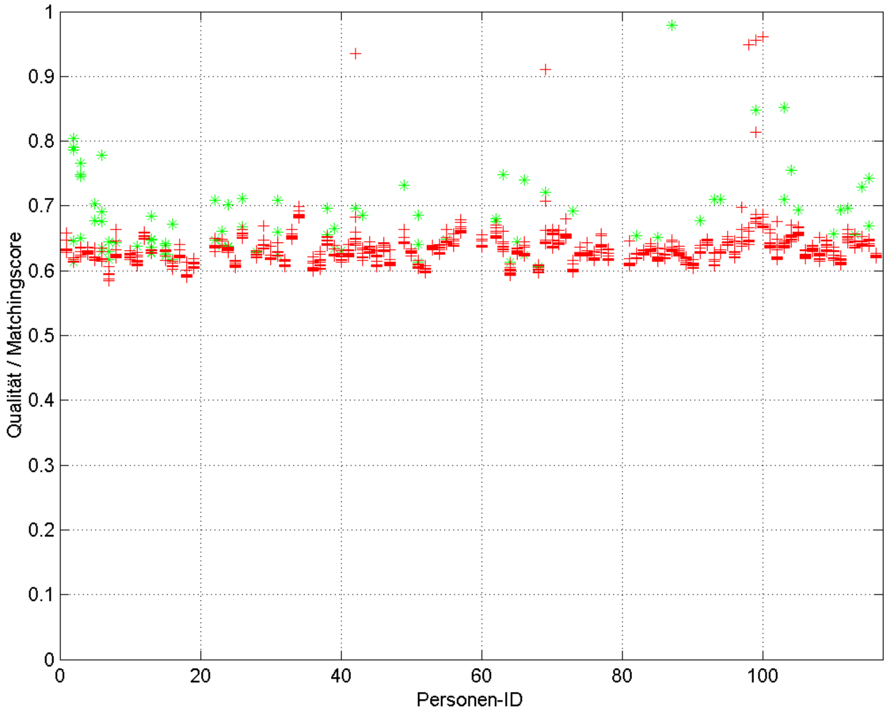


Figure 37: 4-4 Algorithm 1 correct and false hits

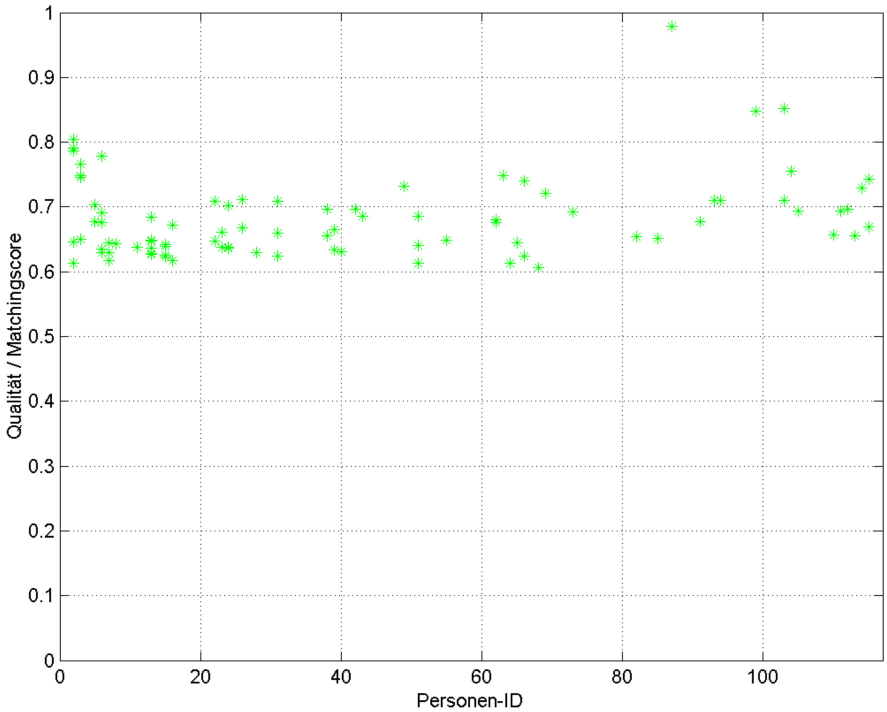


Figure 38: 4-4 Algorithm 1 correct hits

Qualität/Matchingscore	Quality/matching score
Personen-ID	Person ID

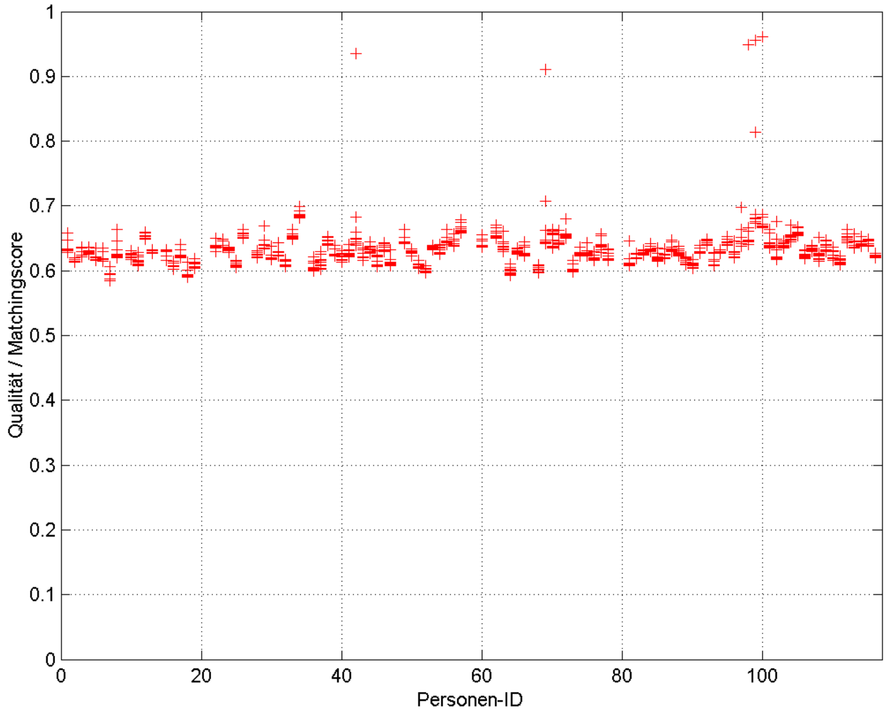


Figure 39: 4-4 Algorithm 1 false hits

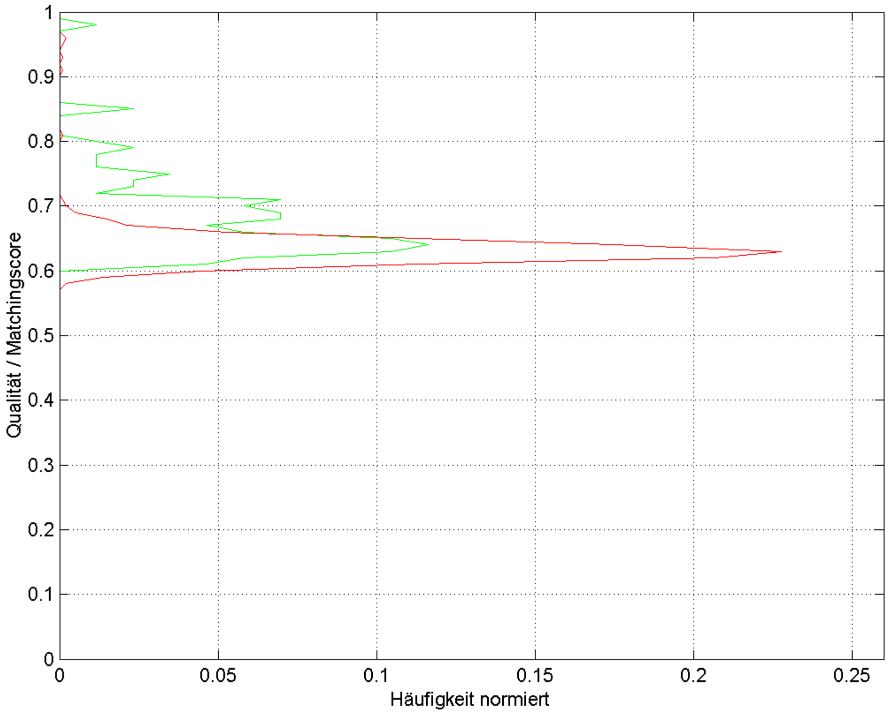


Figure 40: 4-4 Algorithm 1 total correct and false hits

Qualität/Matchingscore	Quality/matching score
Personen-ID	Person ID
Häufigkeit normiert	Normalised frequency

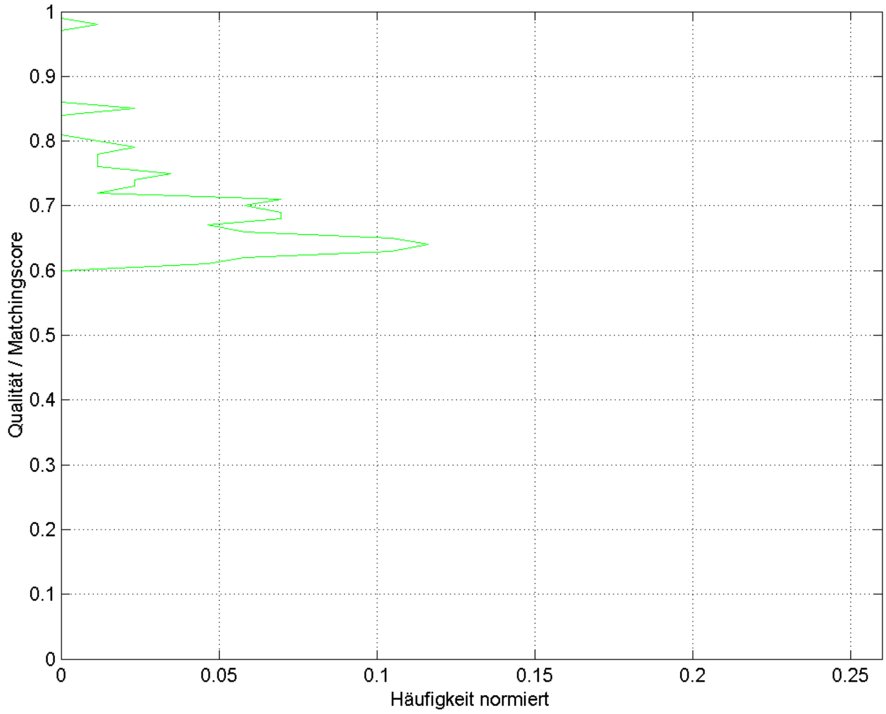


Figure 41: 4-4 Algorithm 1 total correct hits

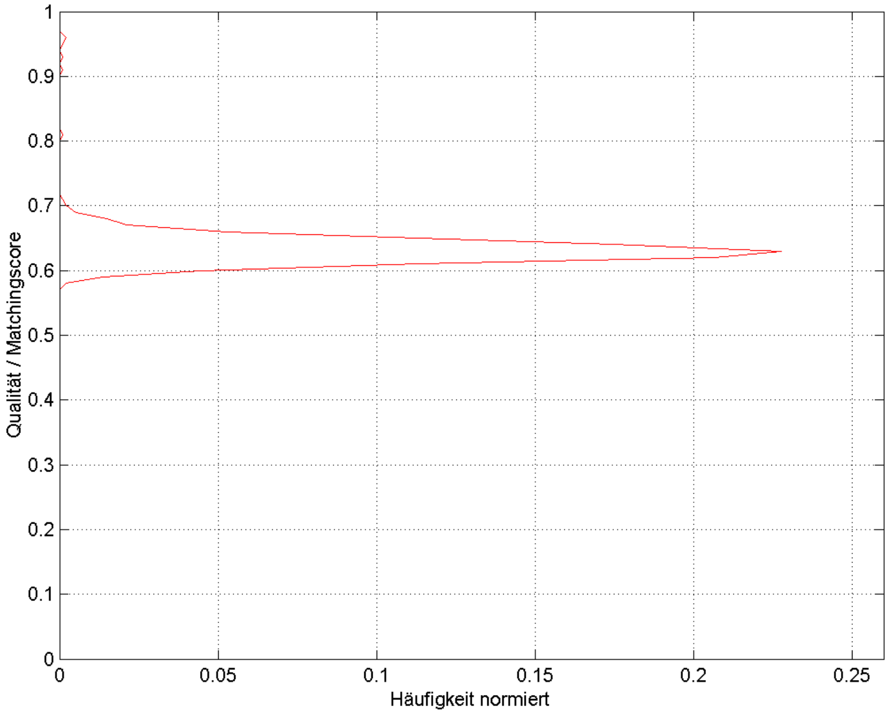


Figure 42: 4-4 Algorithm 1 total false hits

Qualität/Matchingscore	Quality/matching score
Häufigkeit normiert	Normalised frequency

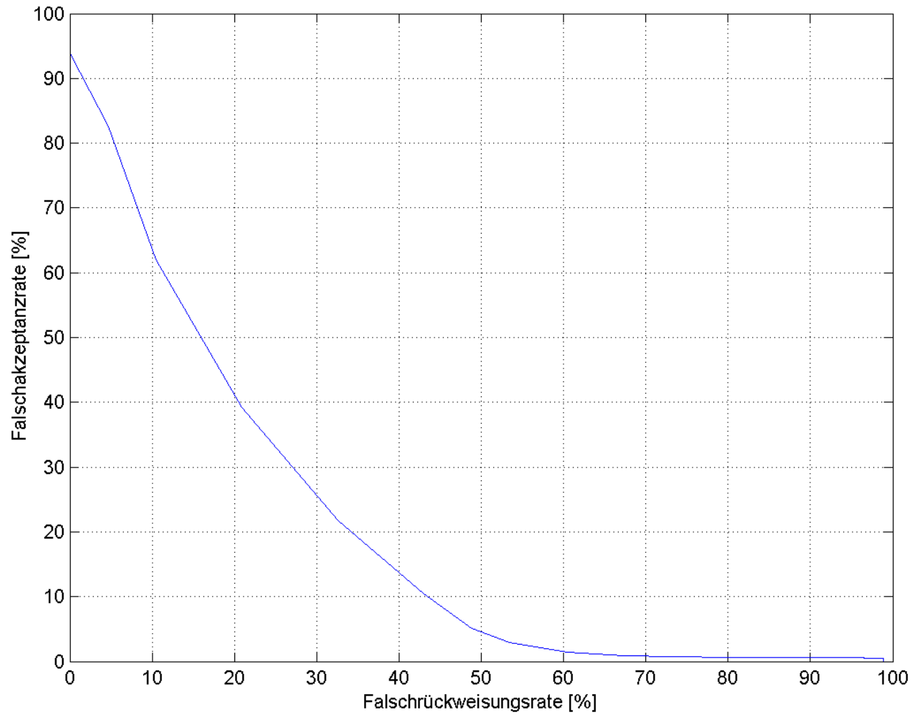


Figure 43: 4-4 Algorithm 1 receiver operating characteristics

Falschakzeptanzrate [%]	False acceptance rate [%]
Falschrückweisungsrate [%]	False rejection rate [%]

5.7.4.8 4-1-Algorithm 2

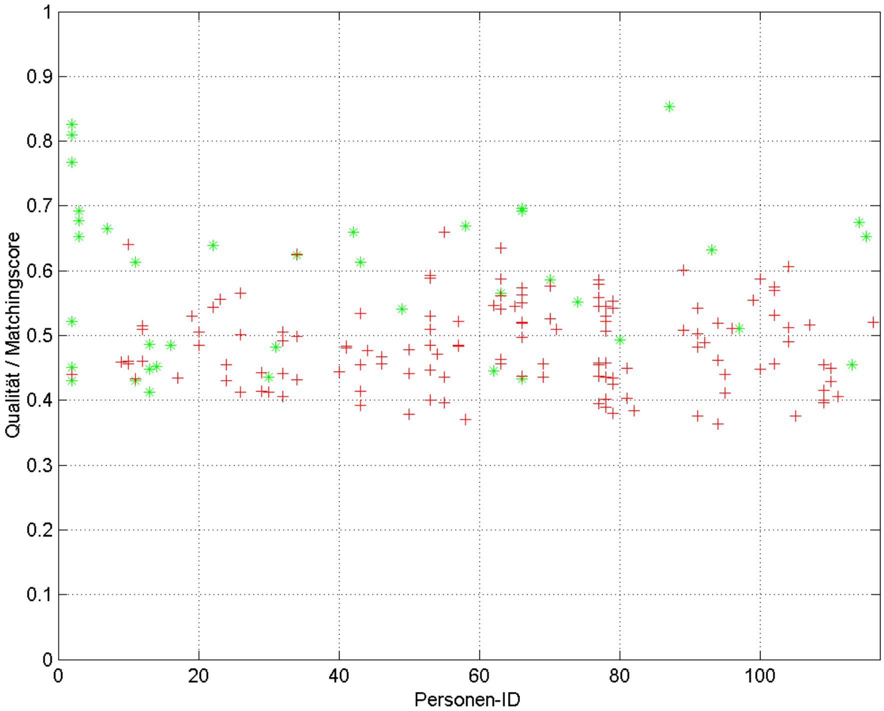


Figure 44: 4-1 Algorithm 2 correct and false hits

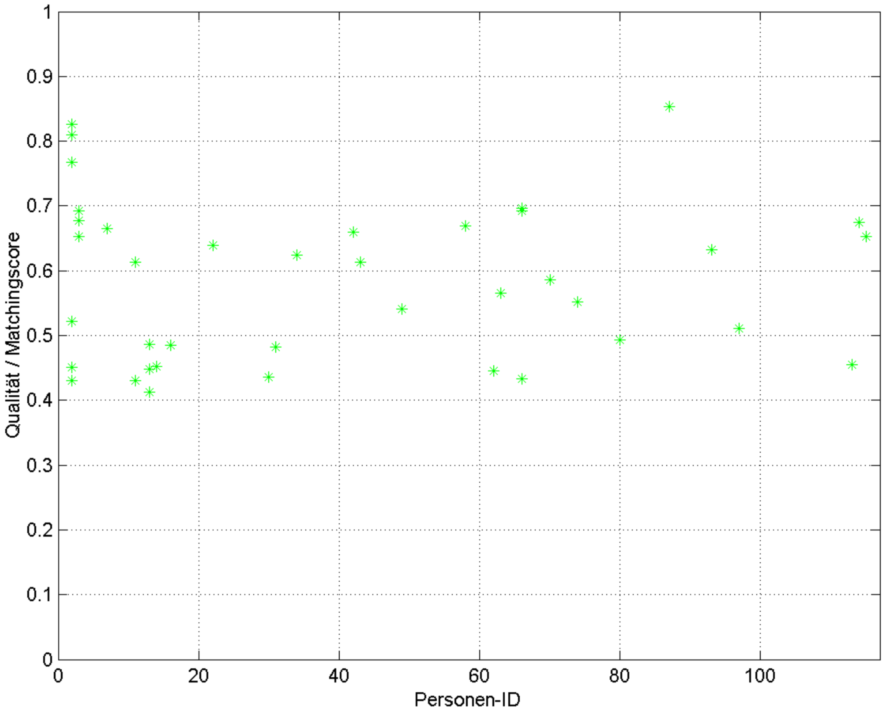


Figure 45: 4-1 Algorithm 2 correct hits

Qualität/Matchingscore	Quality/matching score
Personen-ID	Person ID

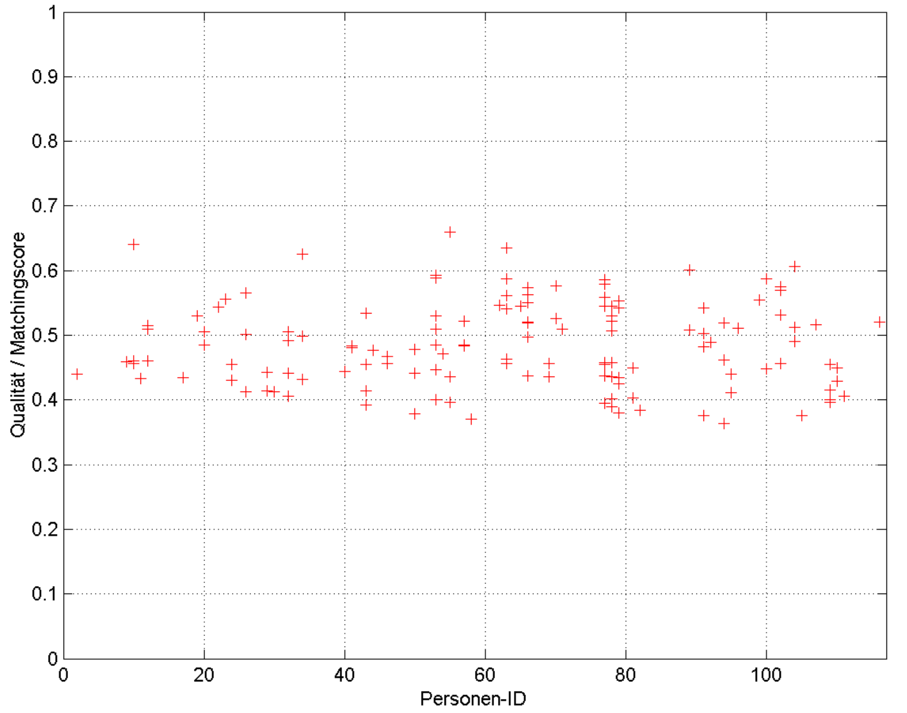


Figure 46: 4-1 Algorithm 2 false hits

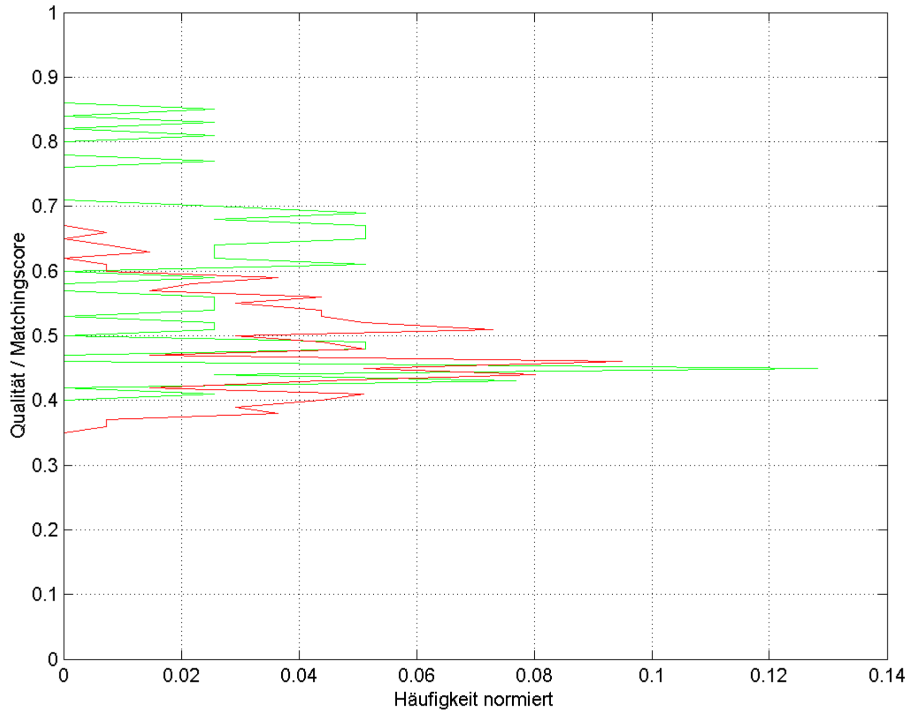


Figure 47: 4-1 Algorithm 2 total correct and false hits

Qualität/Matchingscore	Quality/matching score
Personen-ID	Person ID
Häufigkeit normiert	Normalised frequency

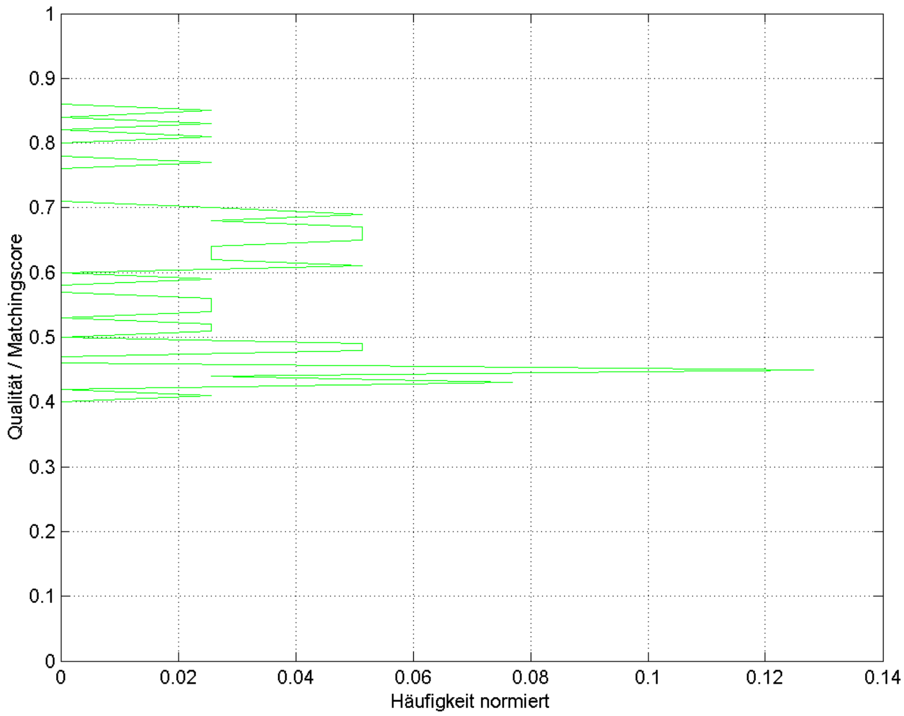


Figure 48: 4-1 Algorithm 2 total correct hits

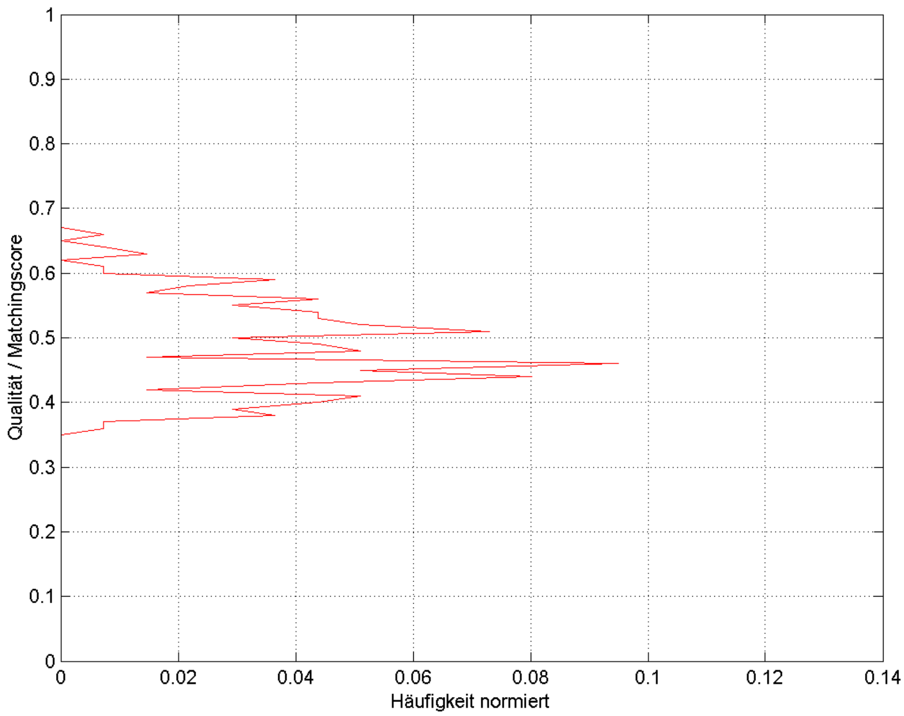


Figure 49: 4-1 Algorithm 2 total false hits

Qualität/Matchingscore	Quality/matching score
Häufigkeit normiert	Normalised frequency

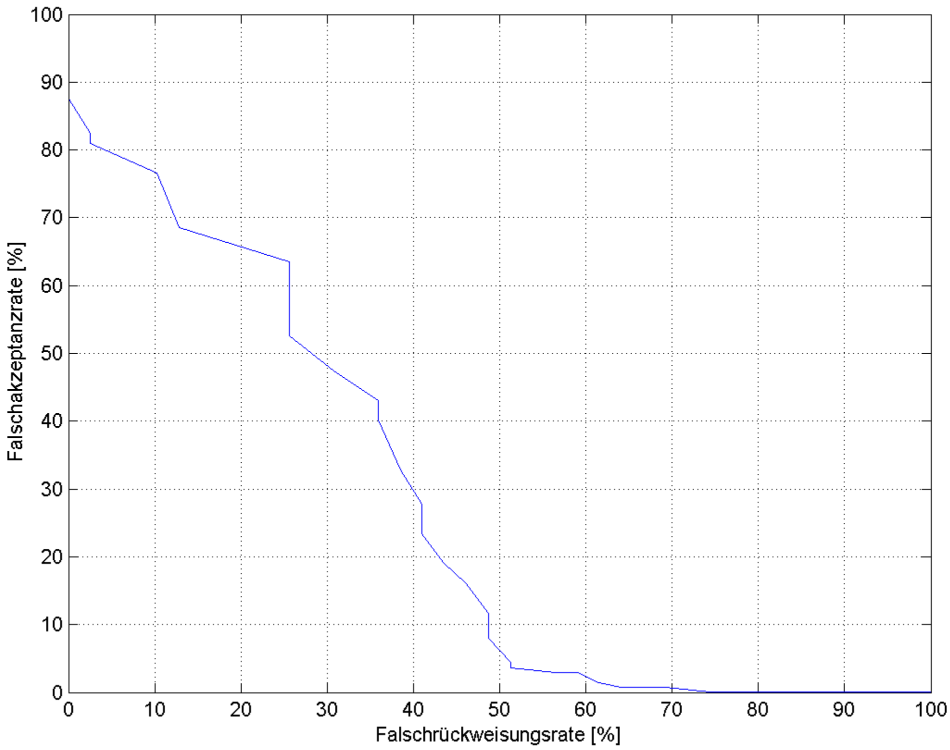


Figure 50: 4-1 Algorithm 2 receiver operating characteristics

Falschakzeptanzrate [%]	False acceptance rate [%]
Falschrückweisungsrate [%]	False rejection rate [%]

5.7.4.9 4-2-Algorithm 2

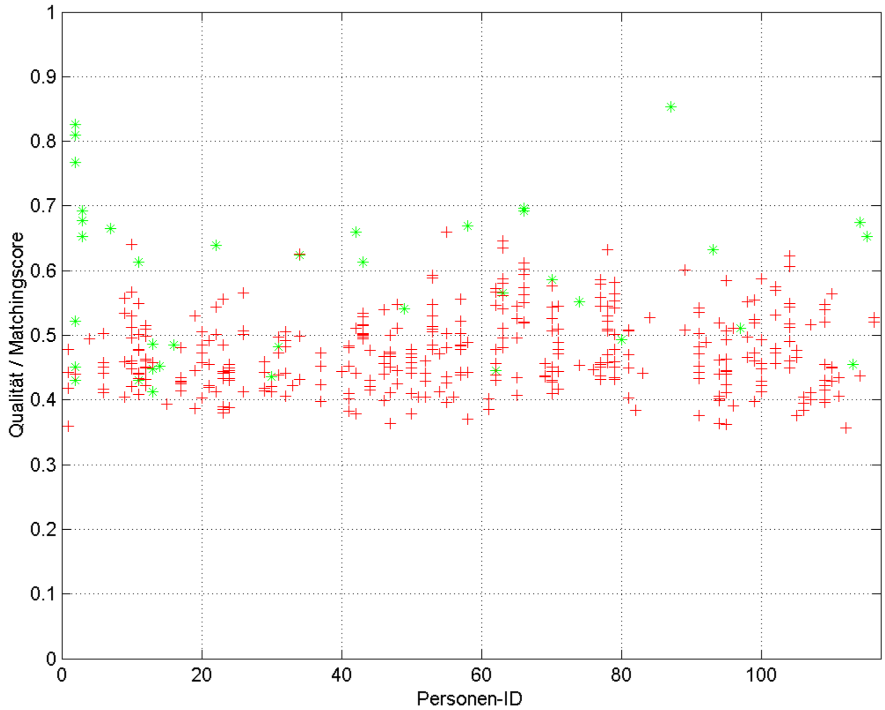


Figure 51: 4-2 Algorithm 2 correct and false hits

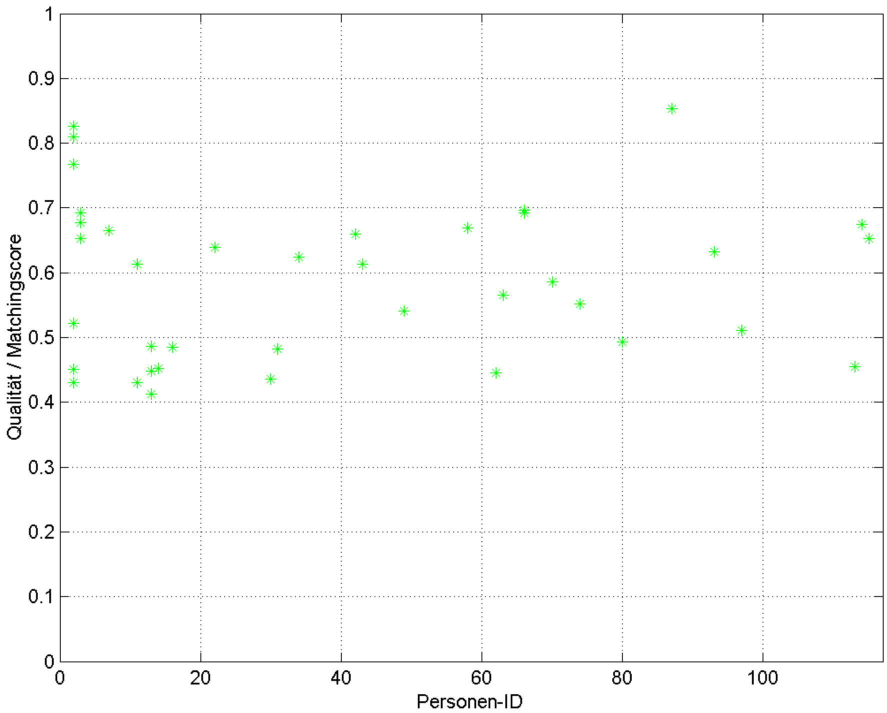


Figure 52: 4-2 Algorithm 2 correct hits

Qualität/Matchingscore	Quality/matching score
Personen-ID	Person ID

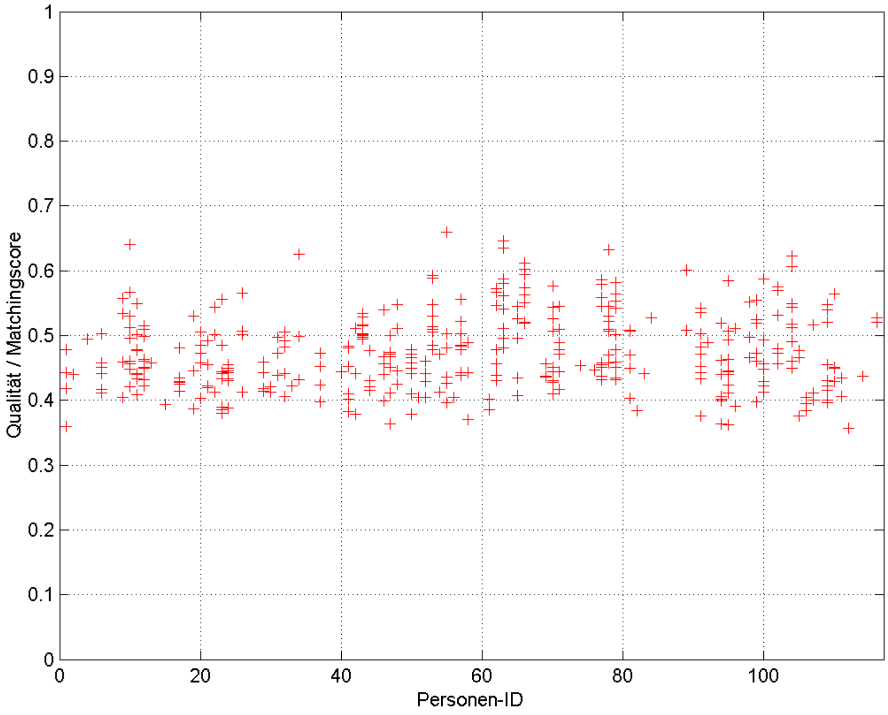


Figure 53: 4-2 Algorithm 2 false hits

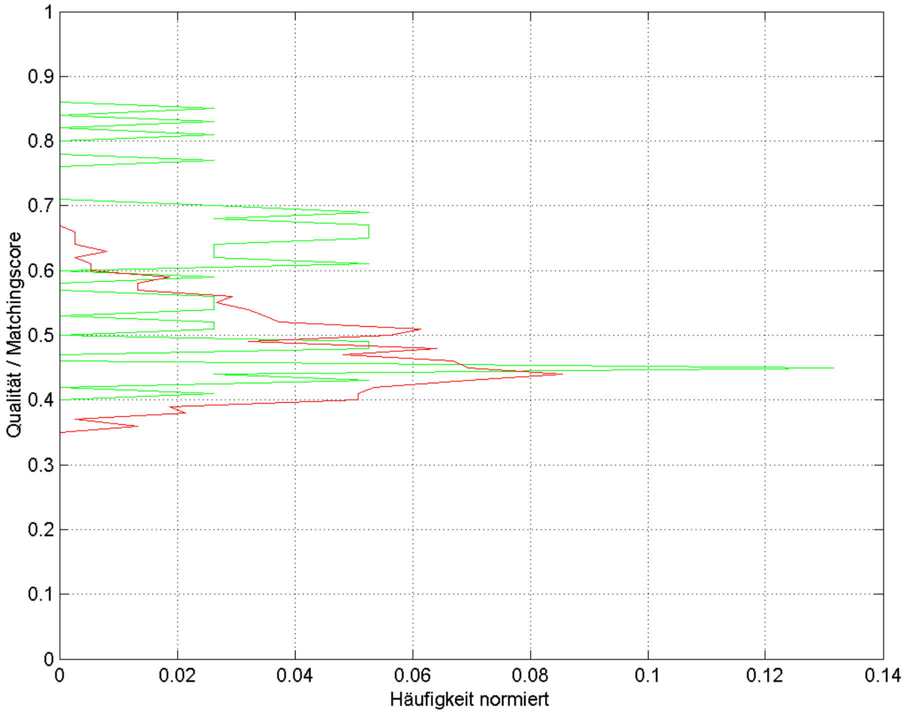


Figure 54: 4-2 Algorithm 2 total correct and false hits

Qualität/Matchingscore	Quality/matching score
Personen-ID	Person ID
Häufigkeit normiert	Normalised frequency

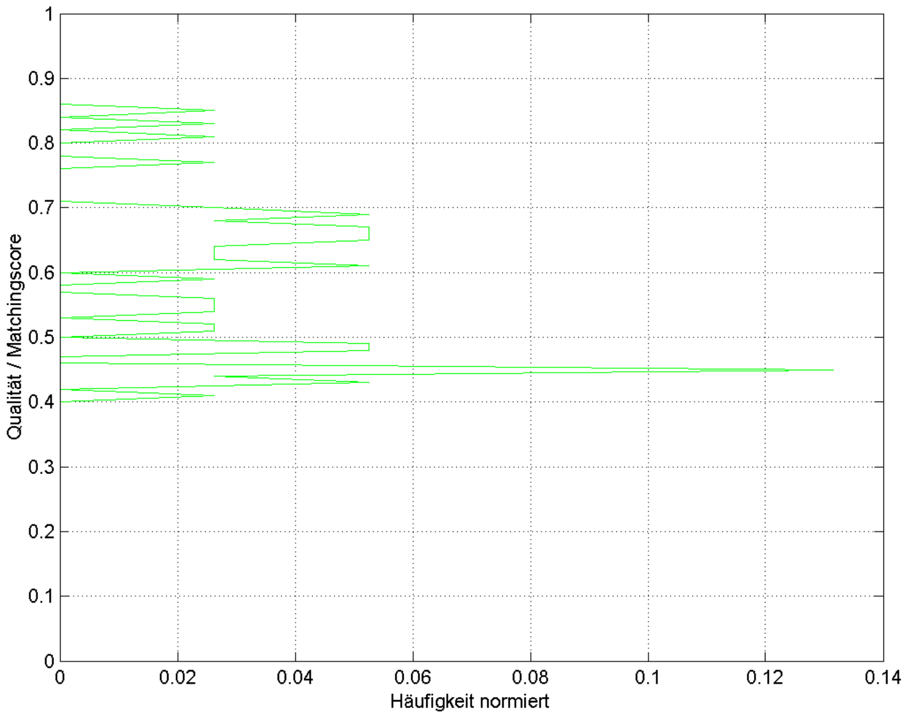


Figure 55: 4-2 Algorithm 2 total correct hits

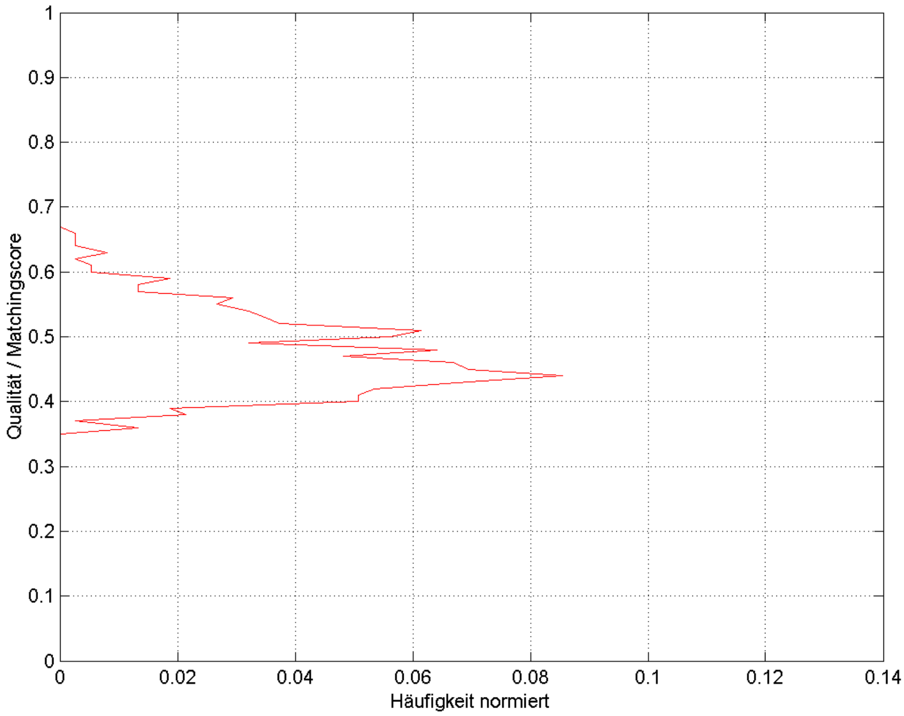


Figure 56: 4-2 Algorithm 2 total false hits

Qualität/Matchingscore	Quality/matching score
Häufigkeit normiert	Normalised frequency

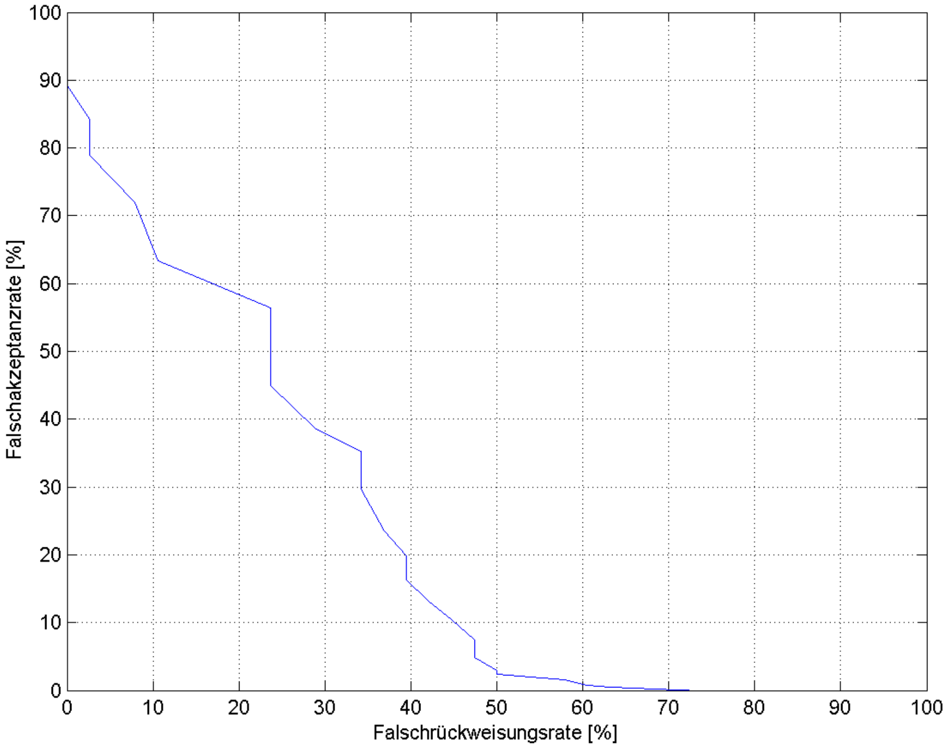


Figure 57: 4-2 Algorithm 2 receiver operating characteristics

Falschakzeptanzrate [%]	False acceptance rate [%]
Falschrückweisungsrate [%]	False rejection rate [%]

5.7.4.10 4-3-Algorithm 2

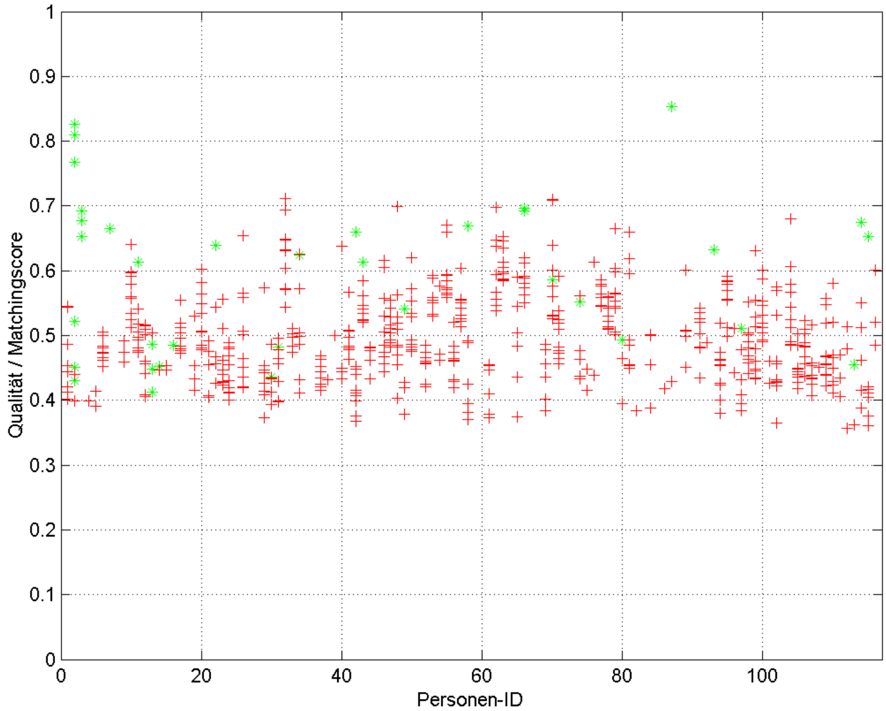


Figure 58: 4-3 Algorithm 2 correct and false hits

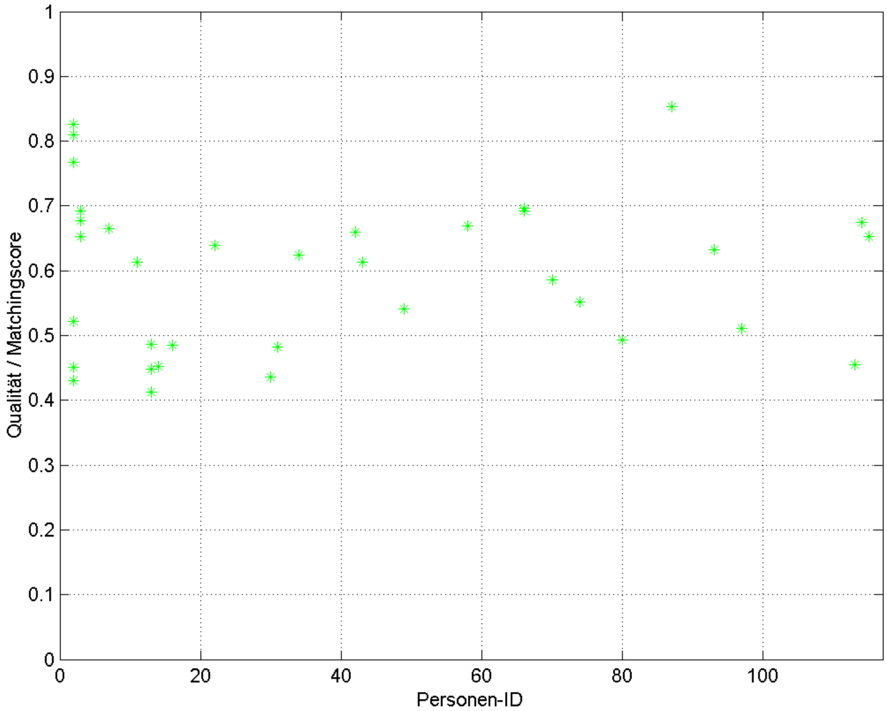


Figure 59: 4-3 Algorithm 2 correct hits

Qualität/Matchingscore	Quality/matching score
Personen-ID	Person ID

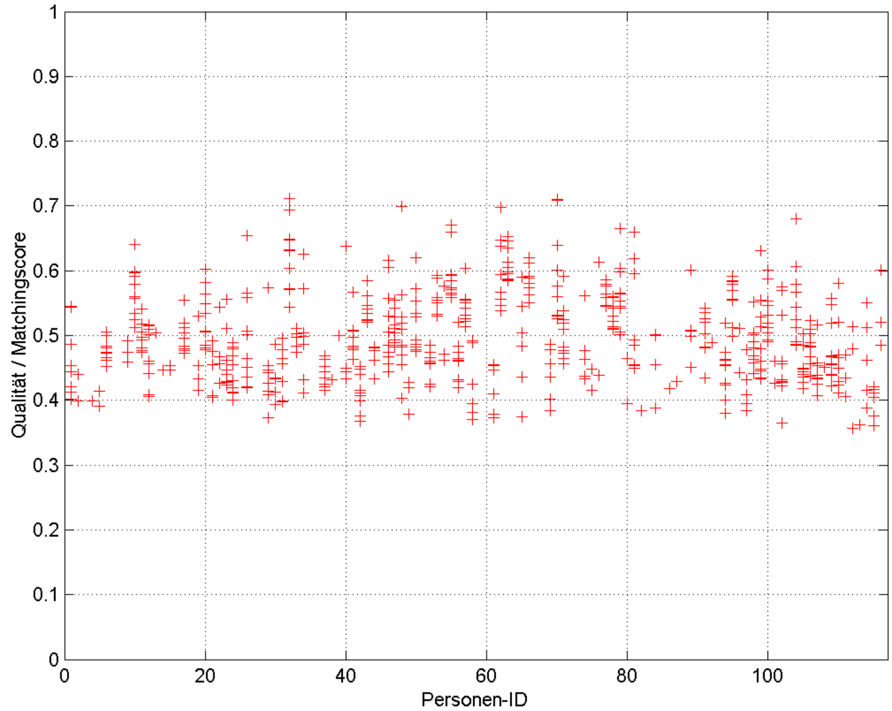


Figure 60: 4-3 Algorithm 2 false hits

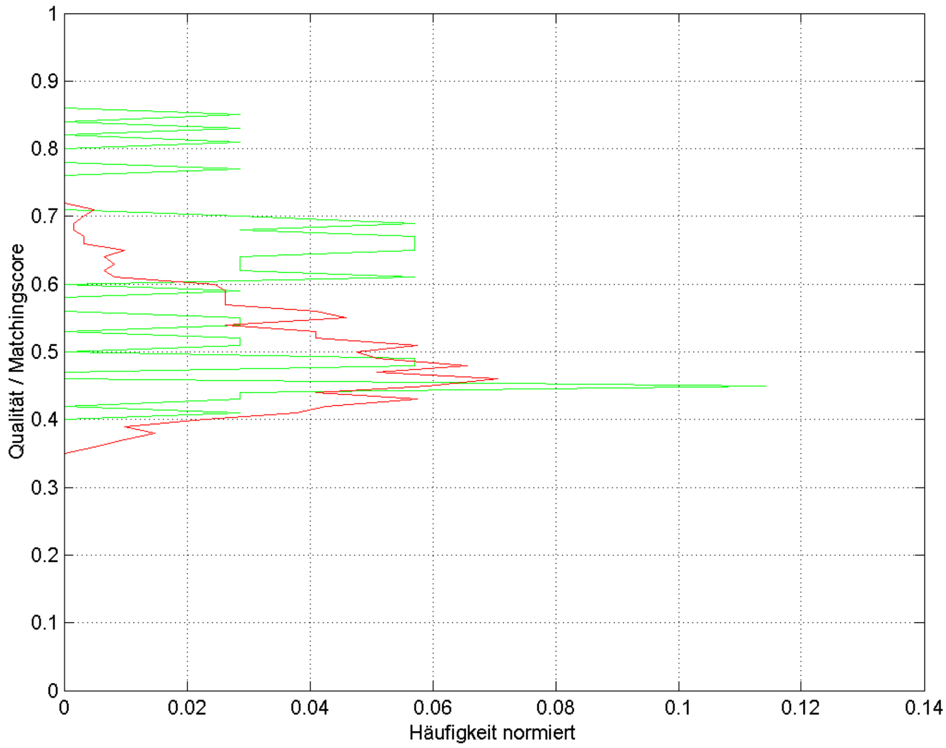


Figure 61: 4-3 Algorithm 2 total correct and false hits

Qualität/Matchingscore	Quality/matching score
Personen-ID	Person ID
Häufigkeit normiert	Normalised frequency

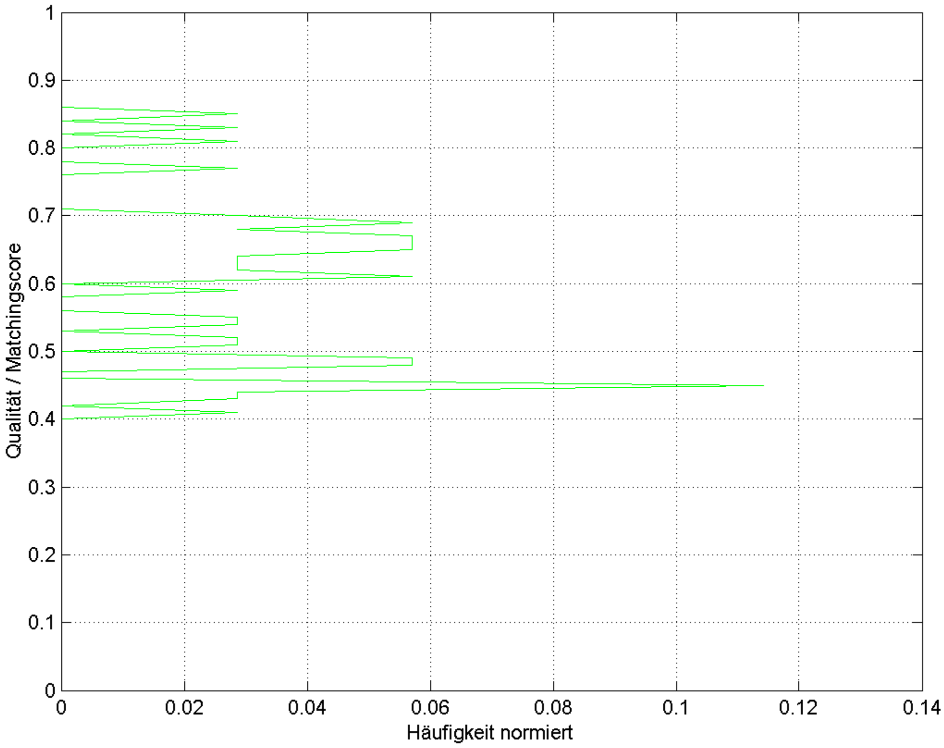


Figure 62: 4-3 Algorithm 2 total correct hits

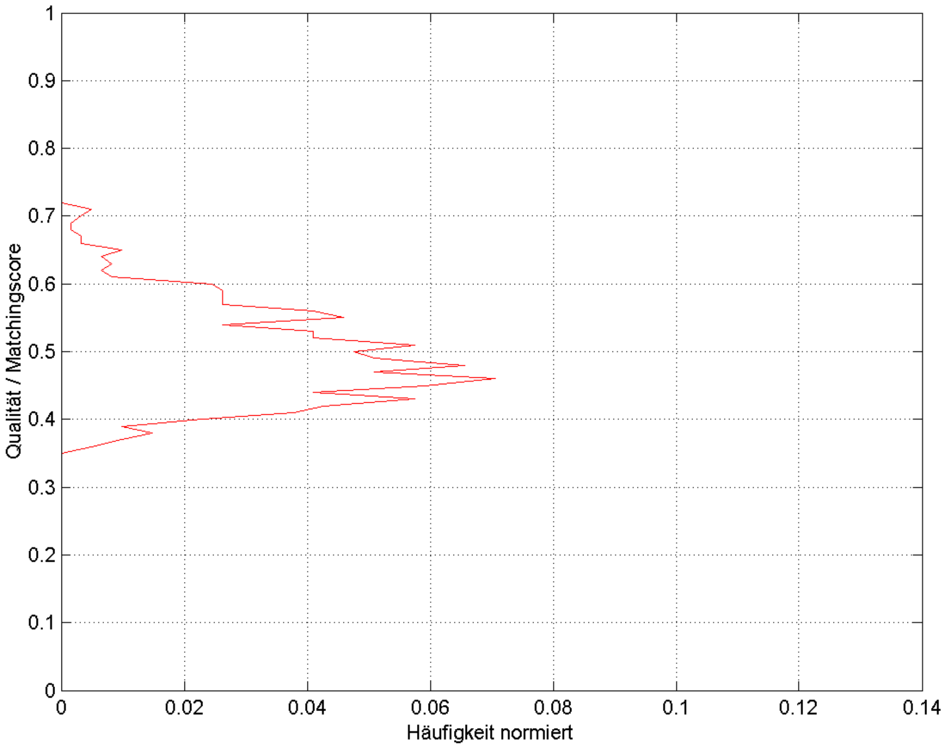


Figure 63: 4-3 Algorithm 2 total false hits

Qualität/Matchingscore	Quality/matching score
Häufigkeit normiert	Normalised frequency

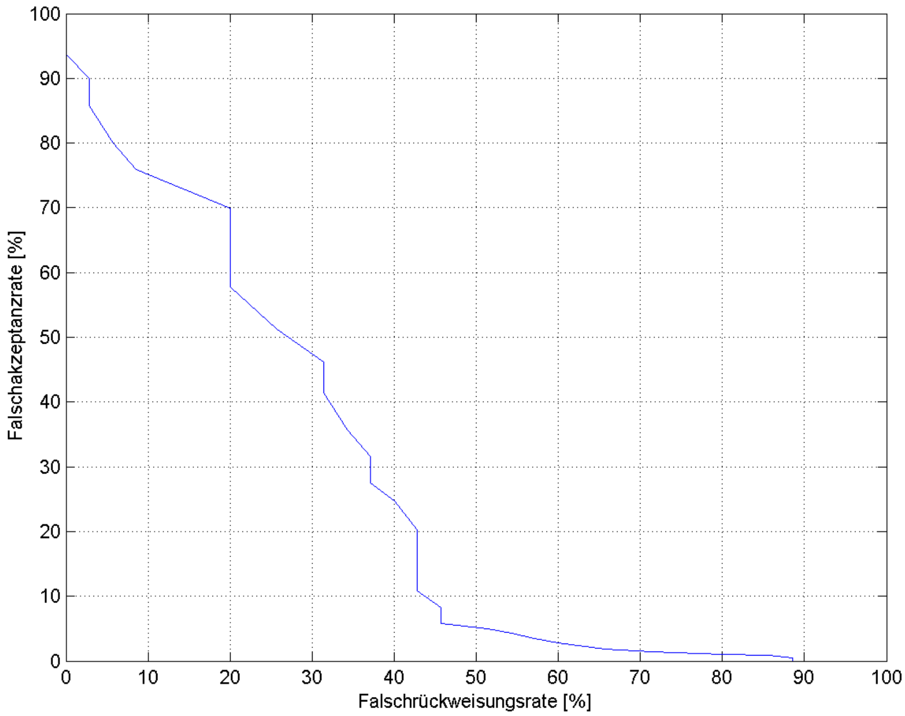


Figure 64: 4-3 Algorithm 2 receiver operating characteristics

Falschakzeptanzrate [%]	False acceptance rate [%]
Falschrückweisungsrate [%]	False rejection rate [%]

5.7.4.11 4-4 Algorithm 2

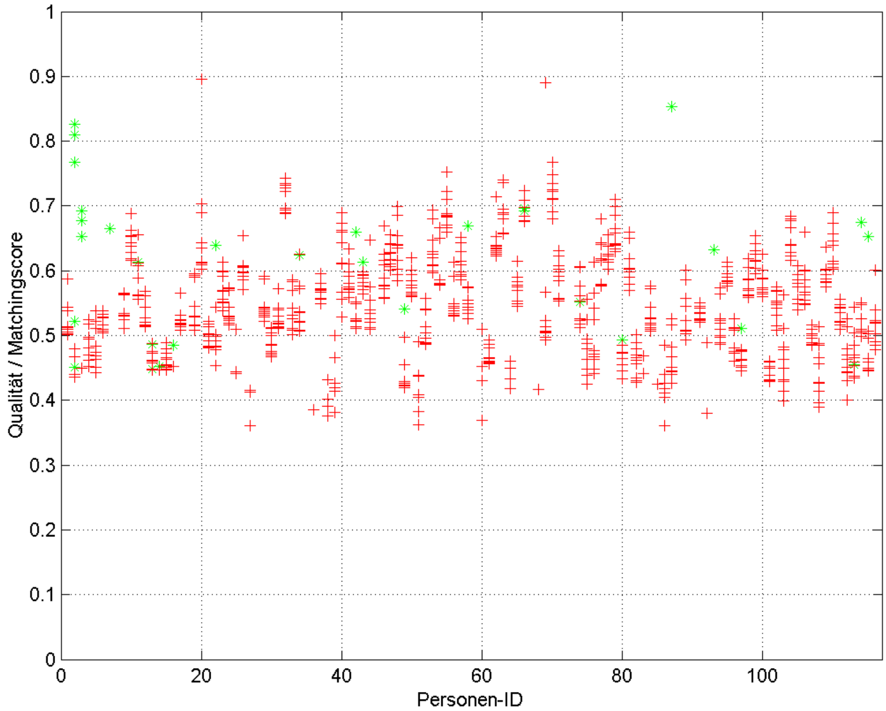


Figure 65: 4-4 Algorithm 2 correct and false hits

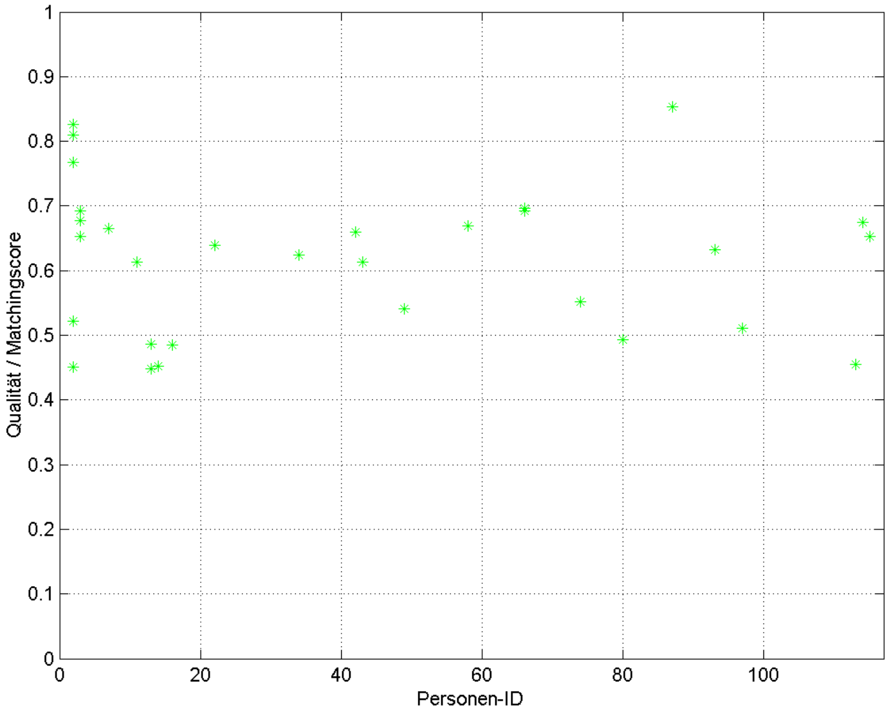


Figure 66: 4-4 Algorithm 2 correct hits

Qualität/Matchingscore	Quality/matching score
Personen-ID	Person ID

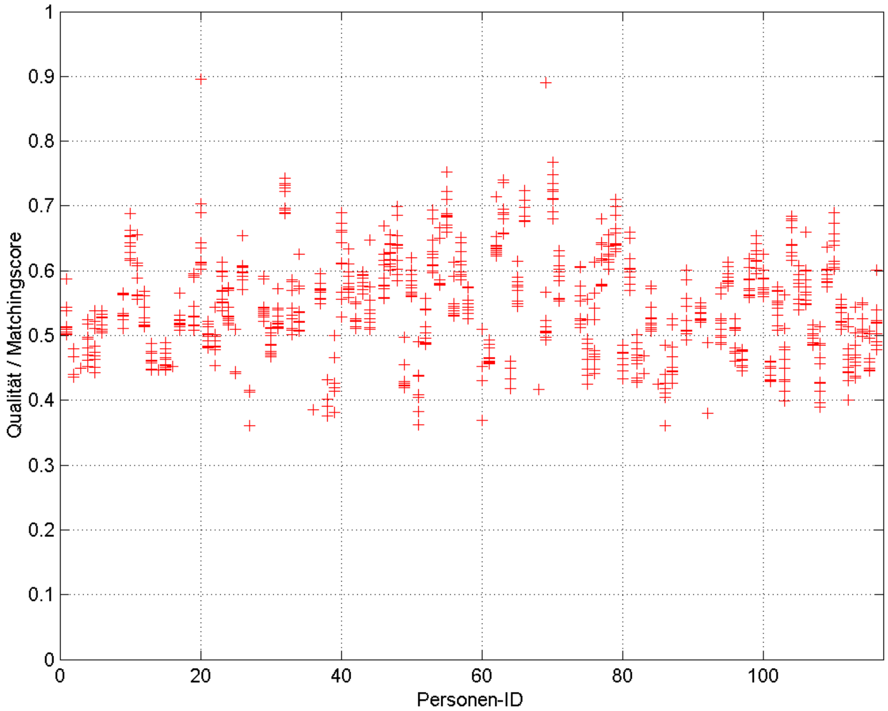


Figure 67: 4-4 Algorithm 2 false hits

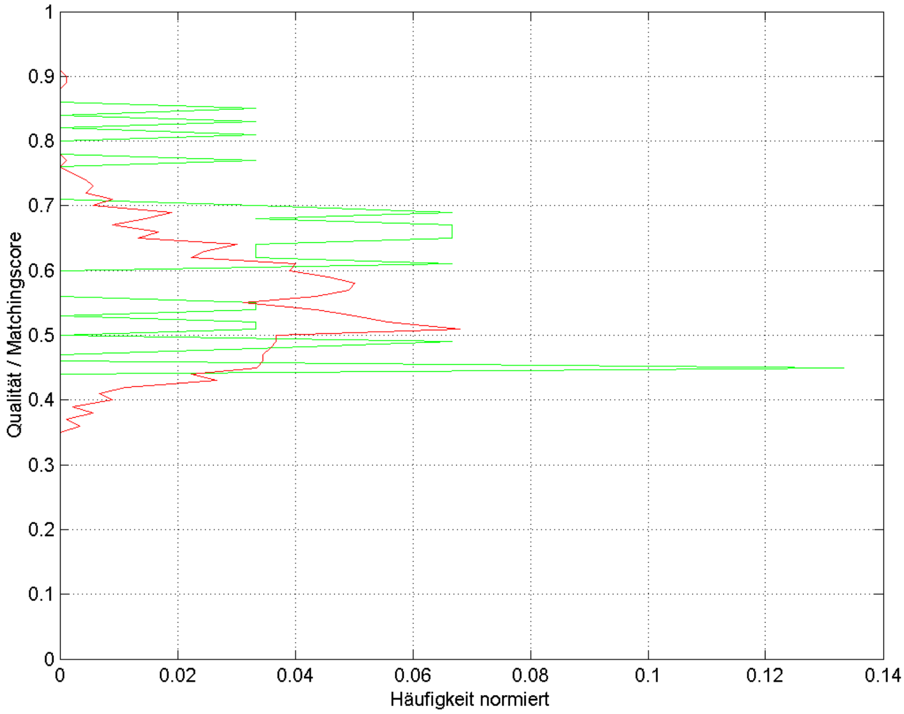


Figure 68: 4-4 Algorithm 2 total correct and false hits

Qualität/Matchingscore	Quality/matching score
Personen-ID	Person ID
Häufigkeit normiert	Normalised frequency

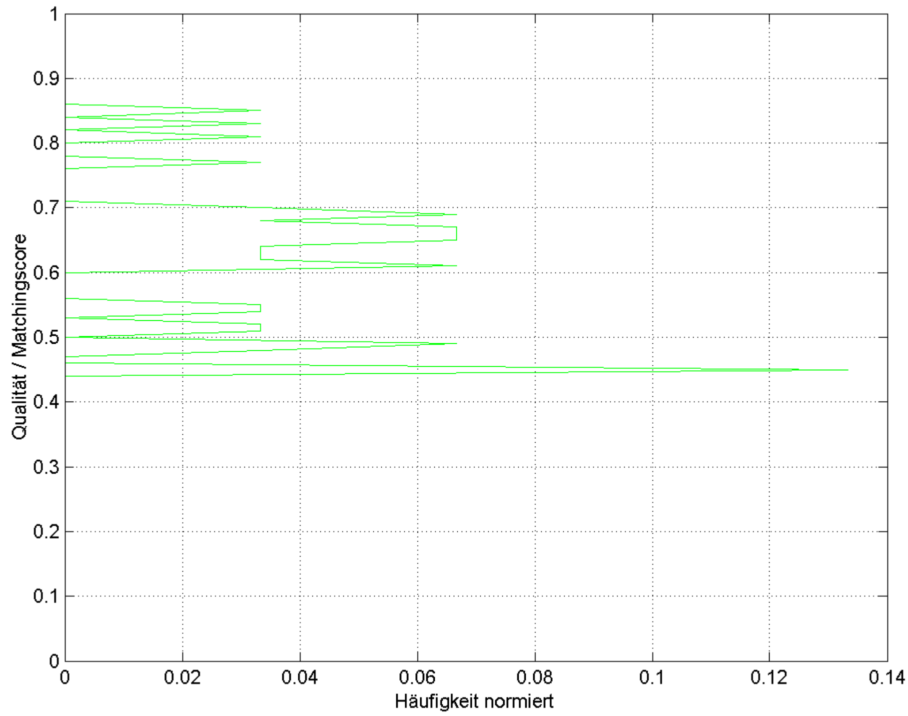


Figure 69: 4-4 Algorithm 2 total correct hits

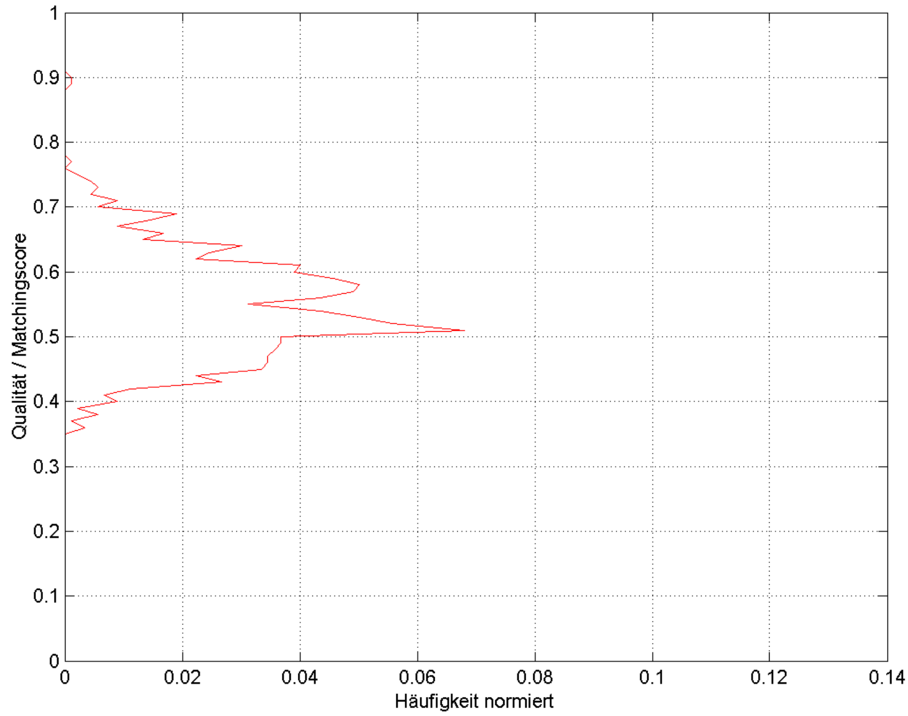


Figure 70: 4-4 Algorithm 2 total false hits

Qualität/Matchingscore	Quality/matching score
Häufigkeit normiert	Normalised frequency

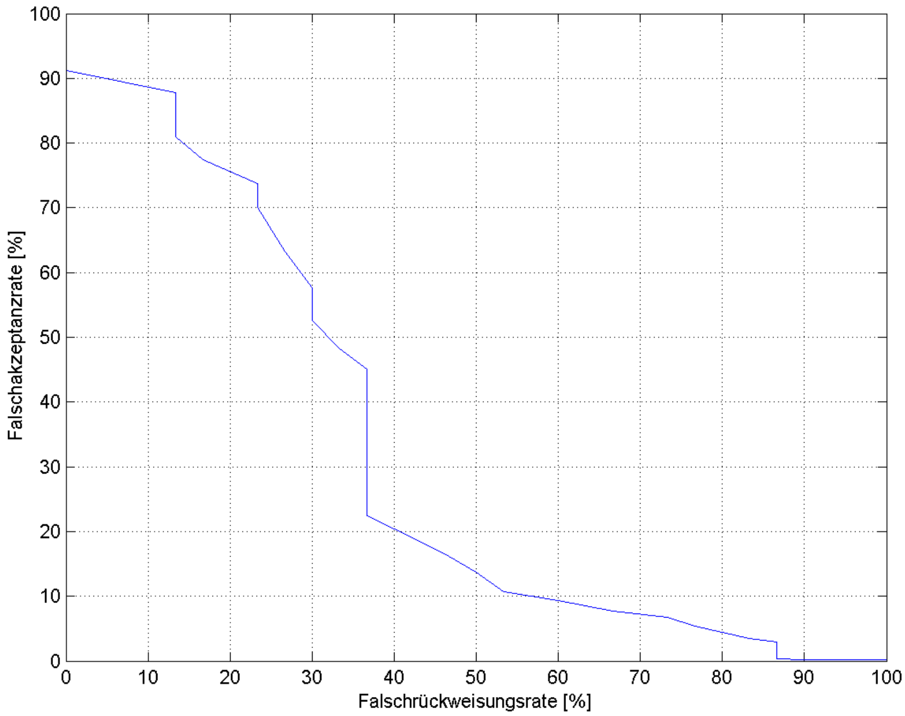


Figure 71: 4-4 Algorithm 2 receiver operating characteristics

Falschakzeptanzrate [%]	False acceptance rate [%]
Falschrückweisungsrate [%]	False rejection rate [%]

5.7.4.12 4-1-Algorithm 3

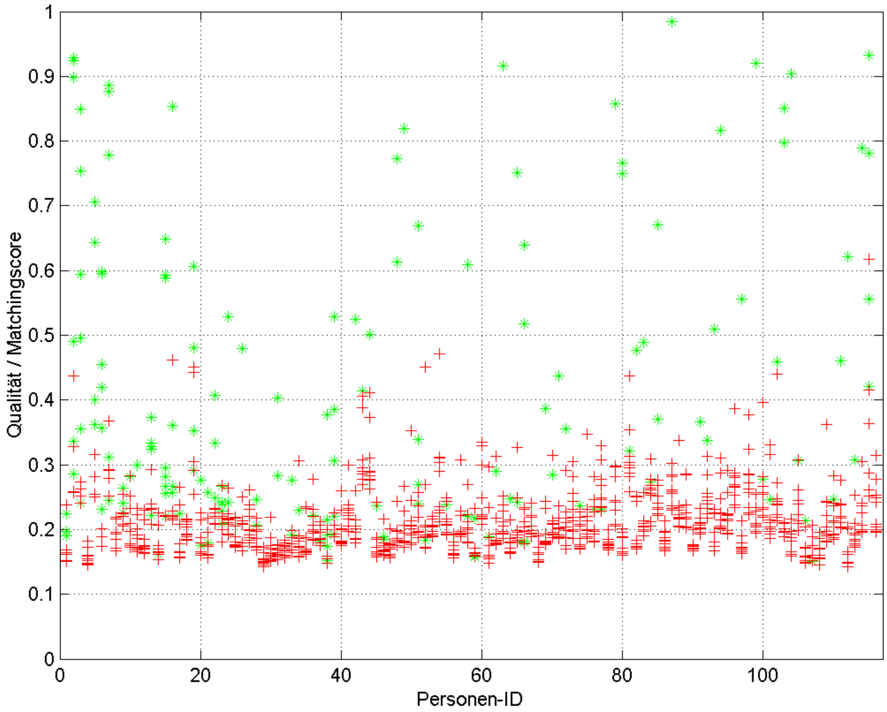


Figure 72: 4-1 Algorithm 3 correct and false hits

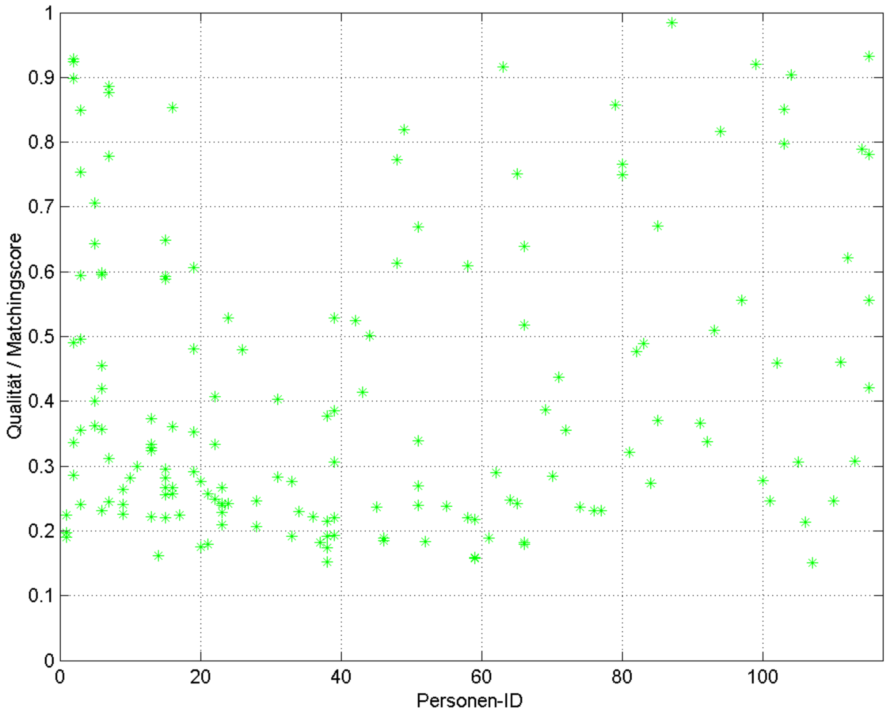


Figure 73: 4-1 Algorithm 3 correct hits

Qualität/Matchingscore	Quality/matching score
Personen-ID	Person ID

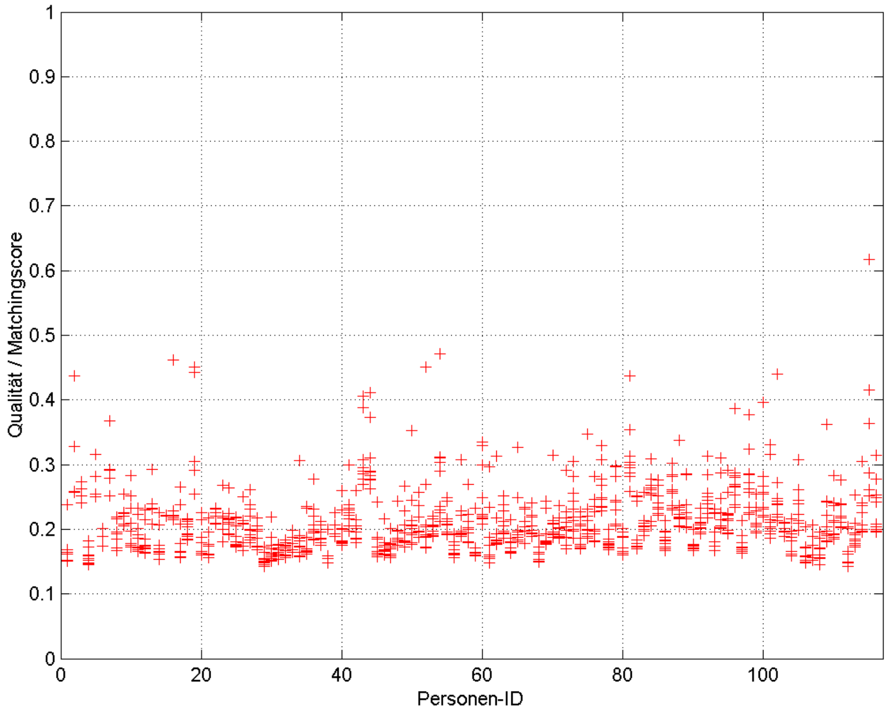


Figure 74: 4-1 Algorithm 3 false hits

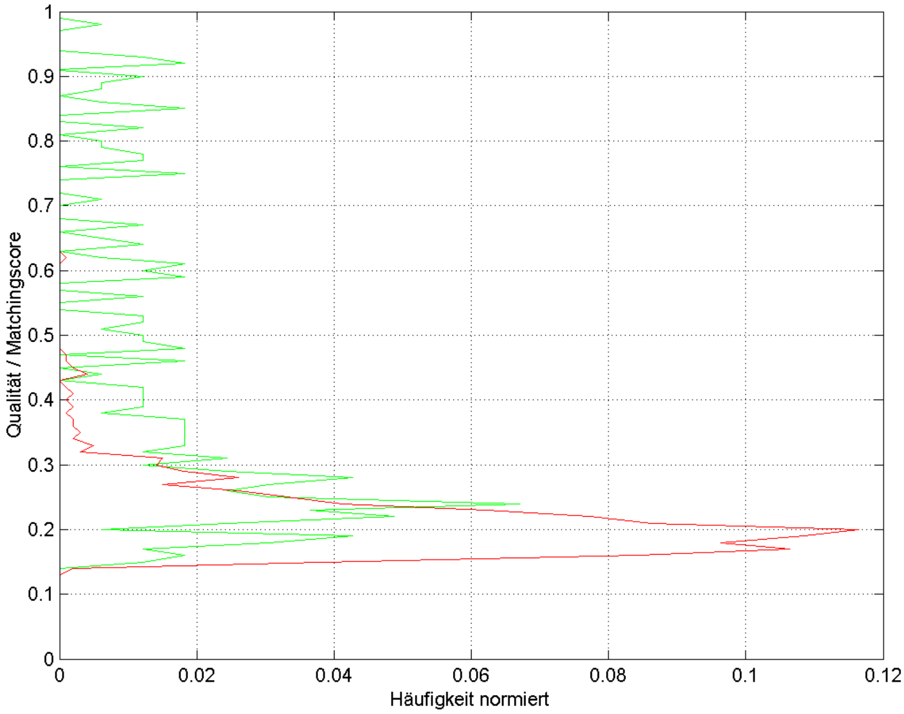


Figure 75: 4-1 Algorithm 3 total correct and false hits

Qualität/Matchingscore	Quality/matching score
Personen-ID	Person ID
Häufigkeit normiert	Normalised frequency

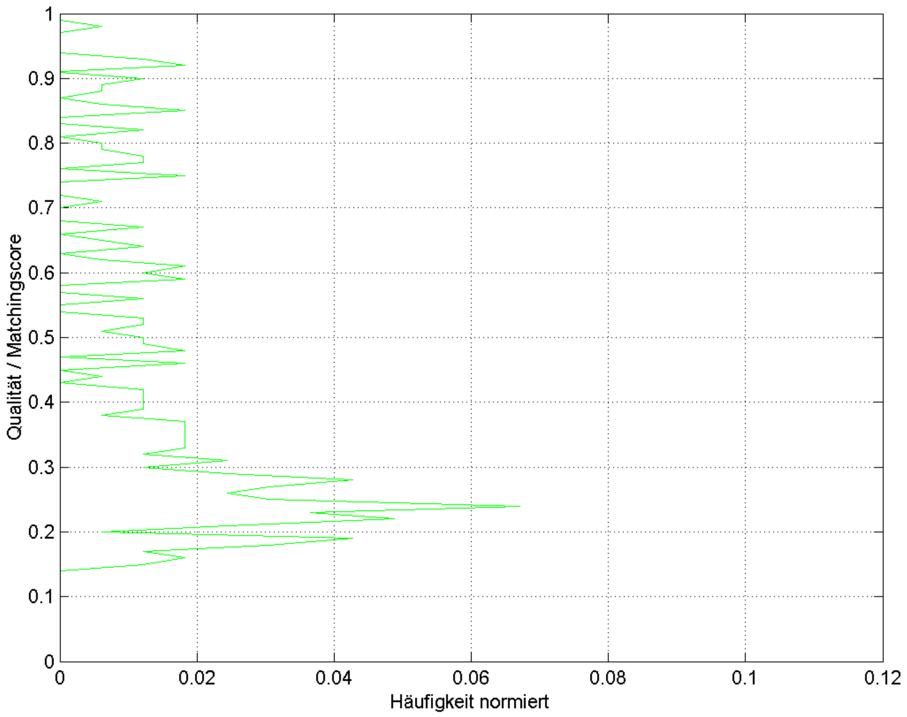


Figure 76: 4-1 Algorithm 3 total correct hits

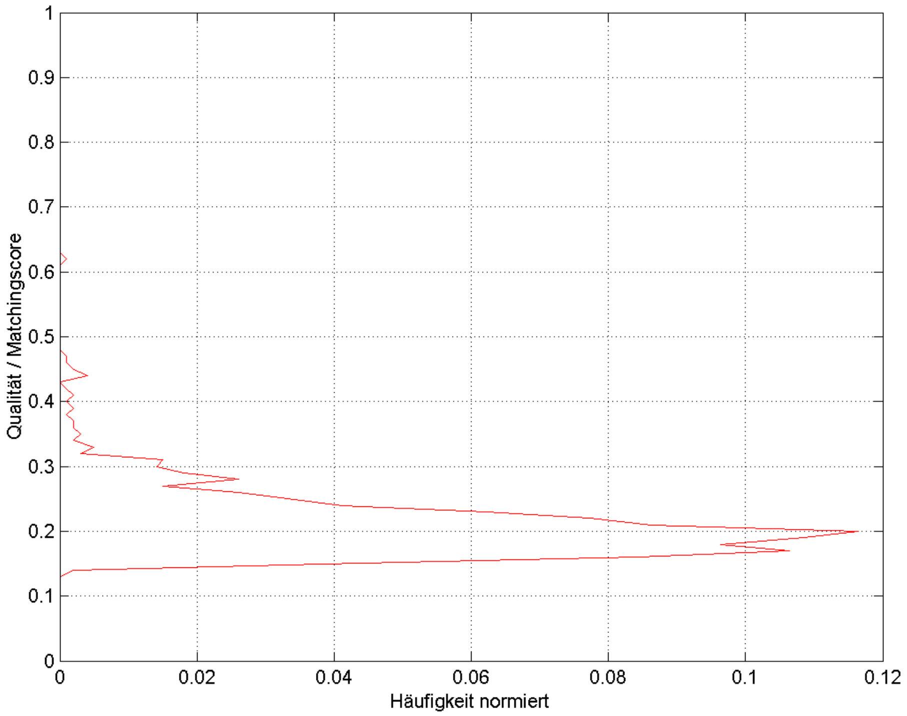


Figure 77: 4-1 Algorithm 3 total false hits

Qualität/Matchingscore	Quality/matching score
Häufigkeit normiert	Normalised frequency

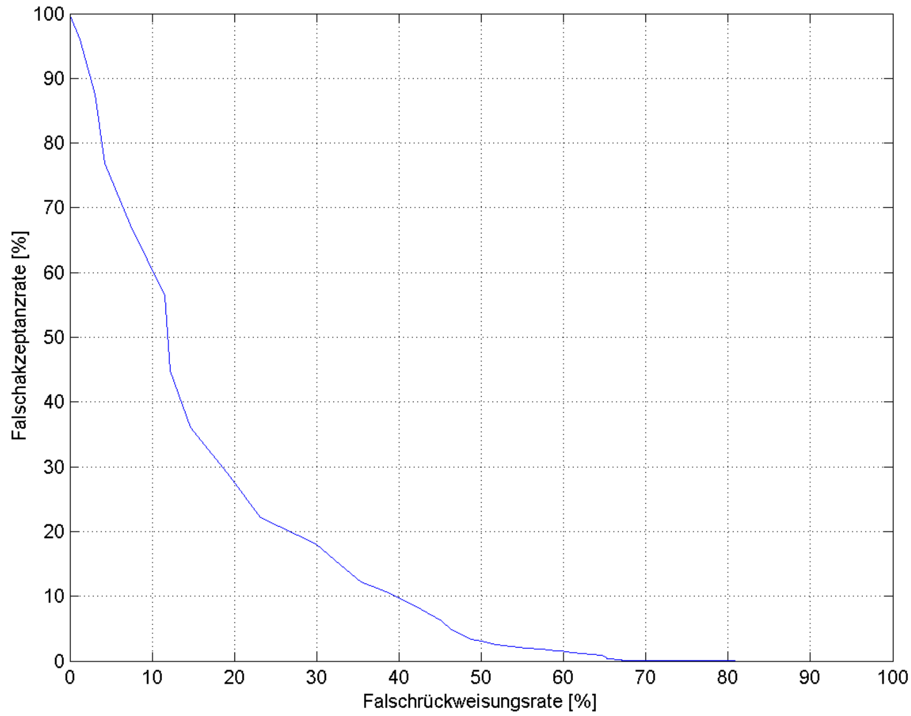


Figure 78: 4-1 Algorithm 3 receiver operating characteristics

Falschakzeptanzrate [%]	False acceptance rate [%]
Falschrückweisungsrate [%]	False rejection rate [%]

5.7.4.13 4-2-Algorithm 3

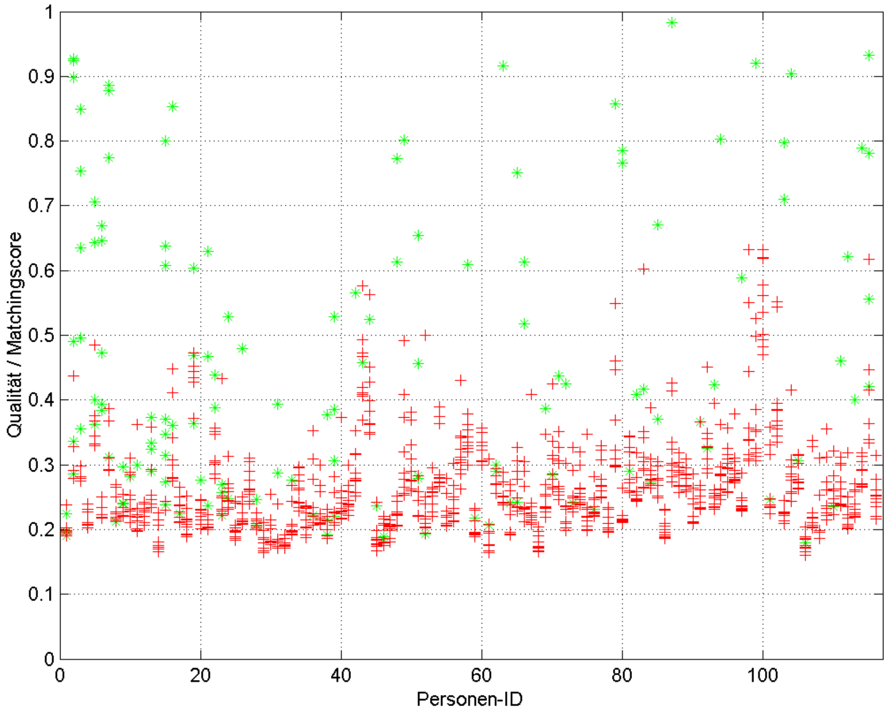


Figure 79: 4-2 Algorithm 3 correct and false hits

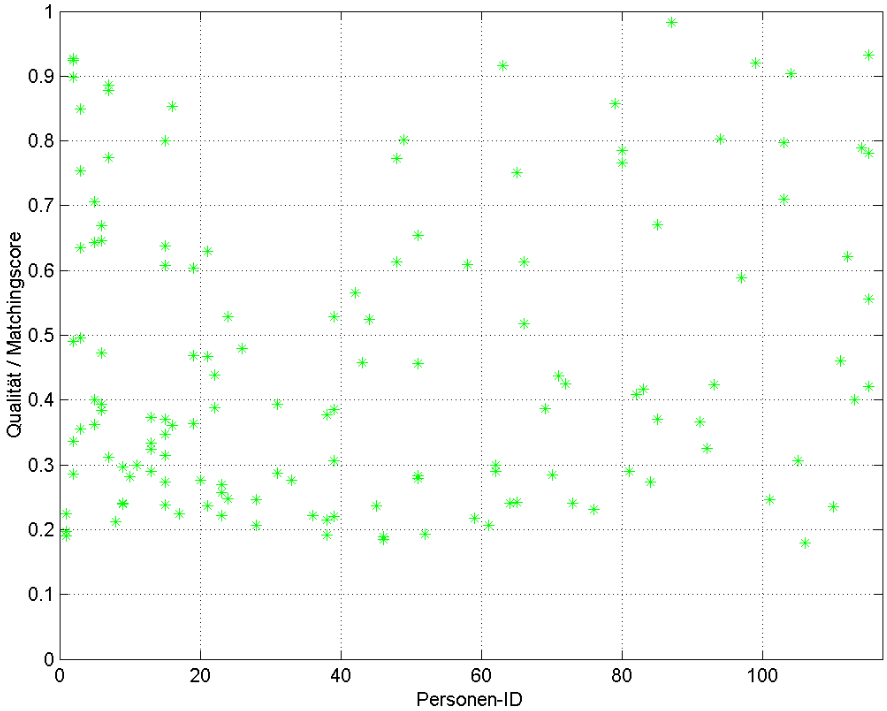


Figure 80: 4-2 Algorithm 3 correct hits

Qualität/Matchingscore	Quality/matching score
Personen-ID	Person ID

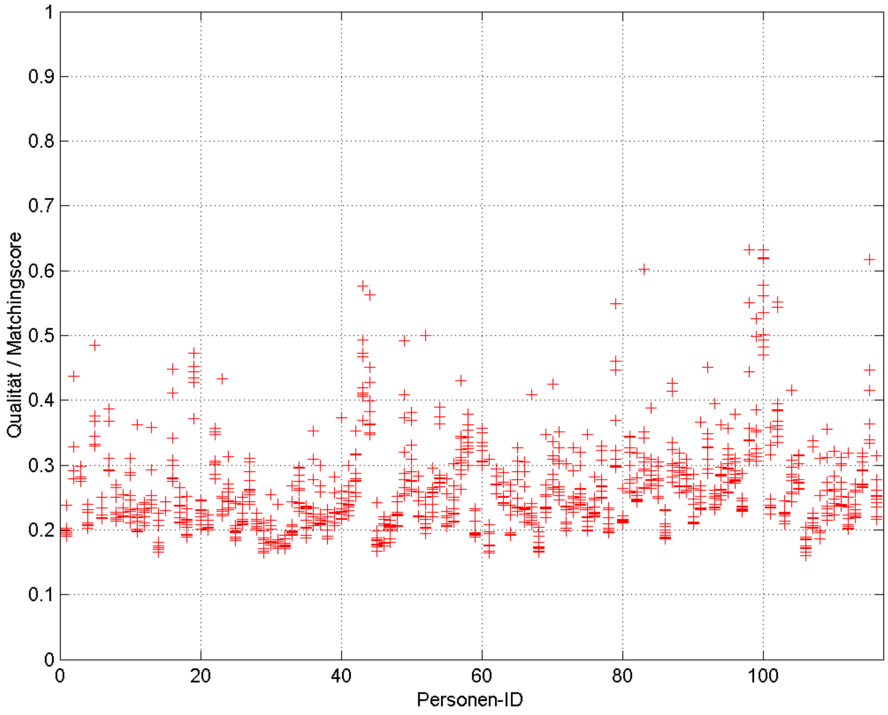


Figure 81: 4-2 Algorithm 3 false hits

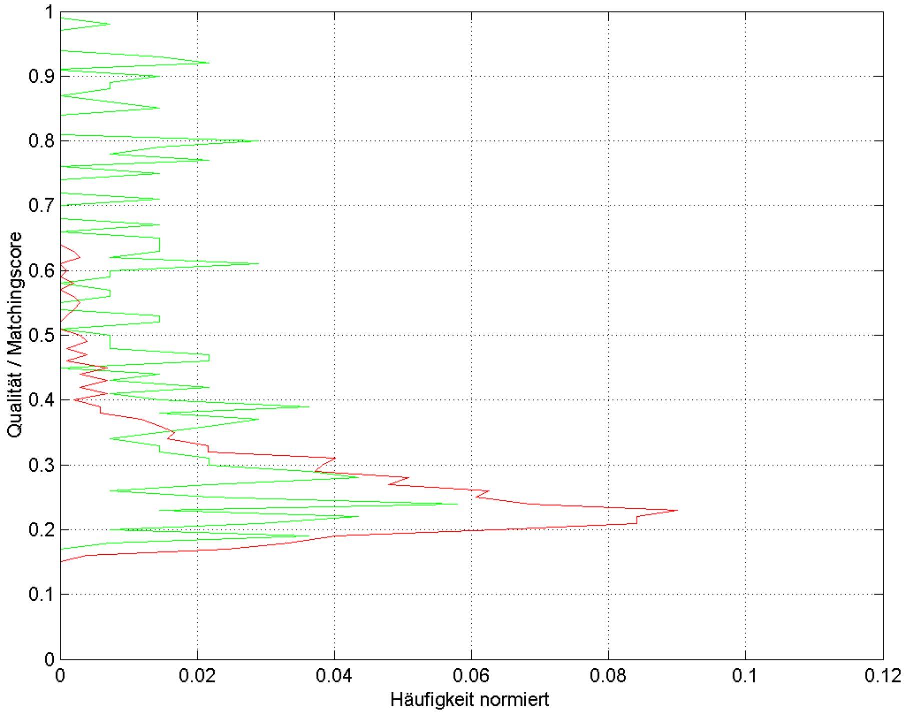


Figure 82: 4-2 Algorithm 3 total correct and false hits

Qualität/Matchingscore	Quality/matching score
Personen-ID	Person ID
Häufigkeit normiert	Normalised frequency

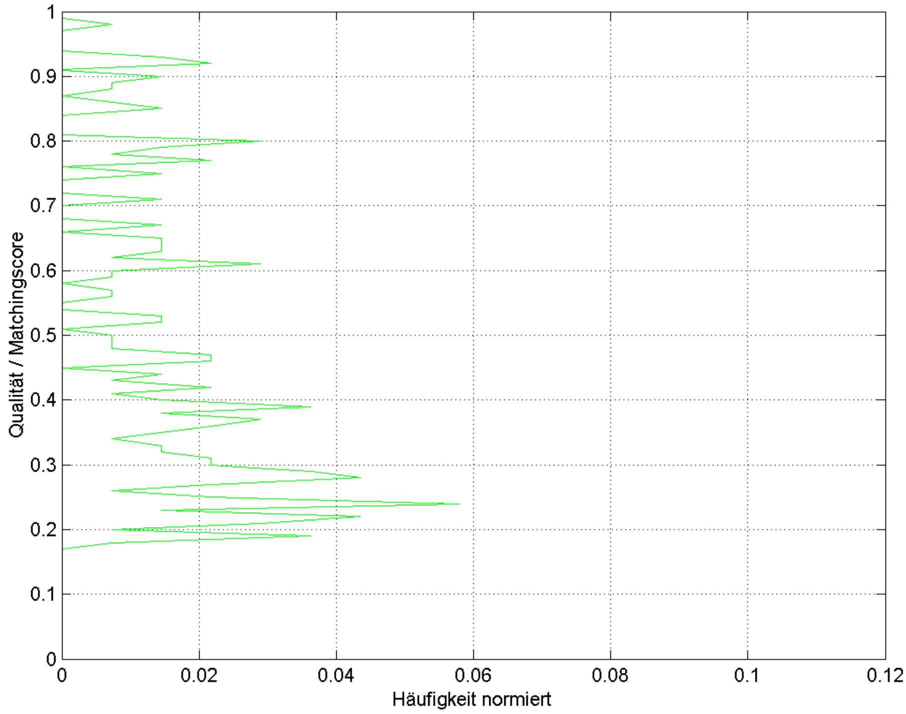


Figure 83: 4-2 Algorithm 3 total correct hits

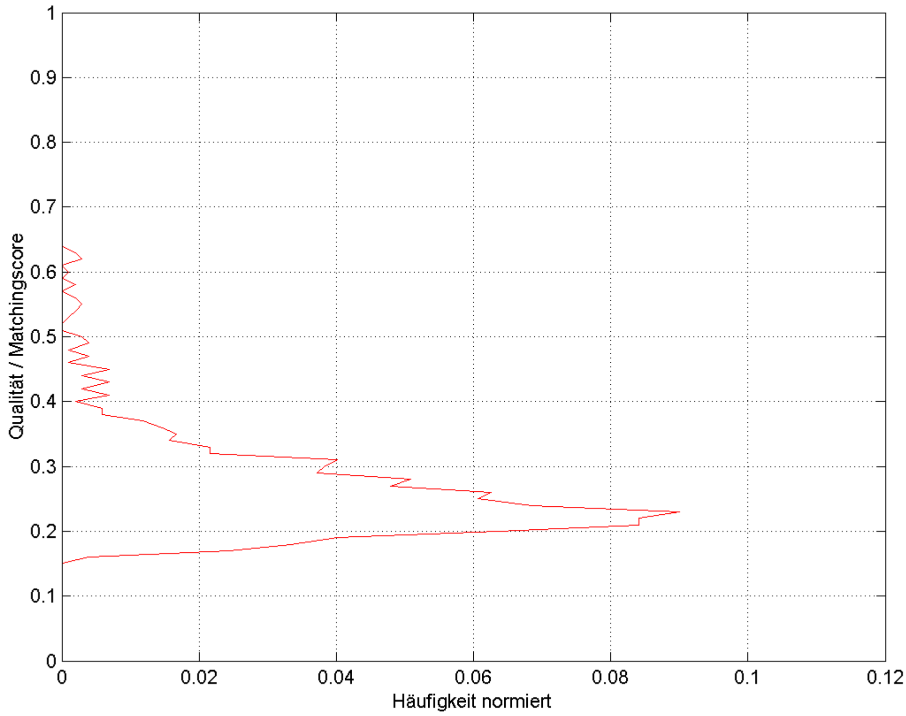


Figure 84: 4-2 Algorithm 3 total false hits

Qualität/Matchingscore	Quality/matching score
Häufigkeit normiert	Normalised frequency

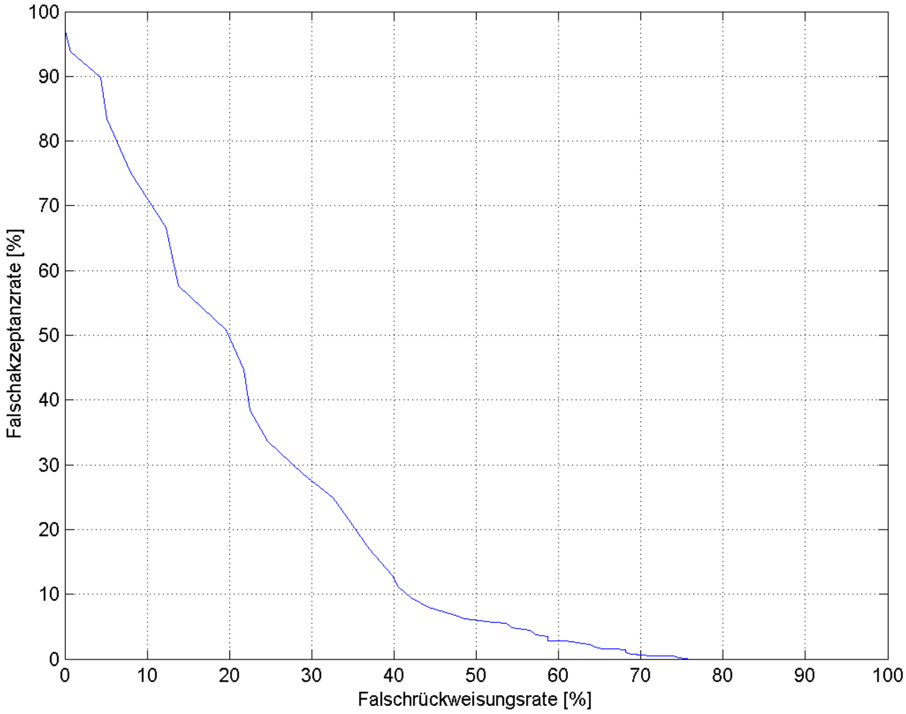


Figure 85: 4-2 Algorithm 3 receiver operating characteristics

Falschakzeptanzrate [%]	False acceptance rate [%]
Falschrückweisungsrate [%]	False rejection rate [%]

5.7.4.14 4-3-Algorithm 3

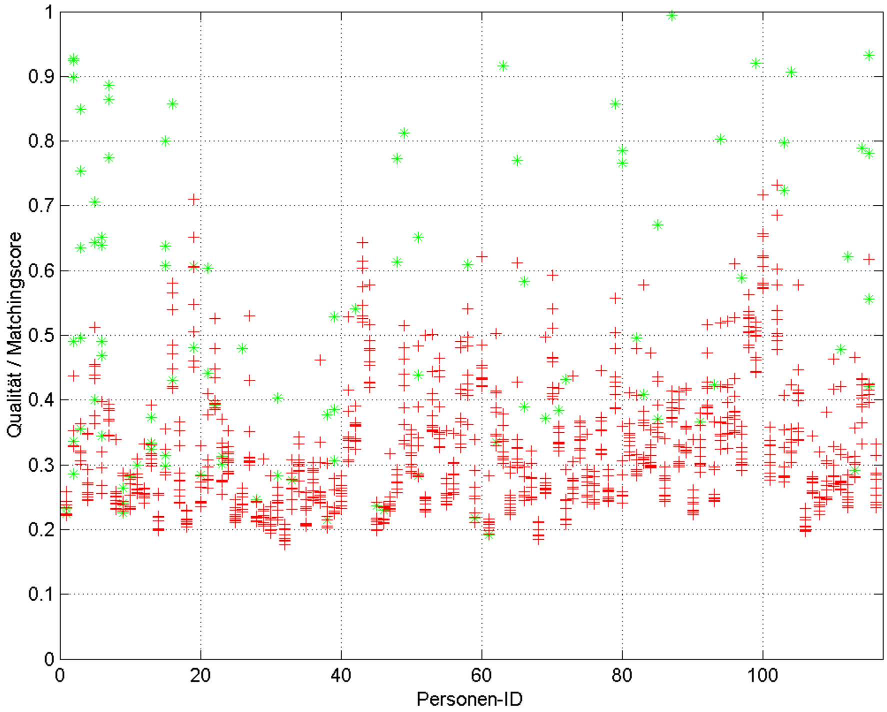


Figure 86: 4-3 Algorithm 3 correct and false hits

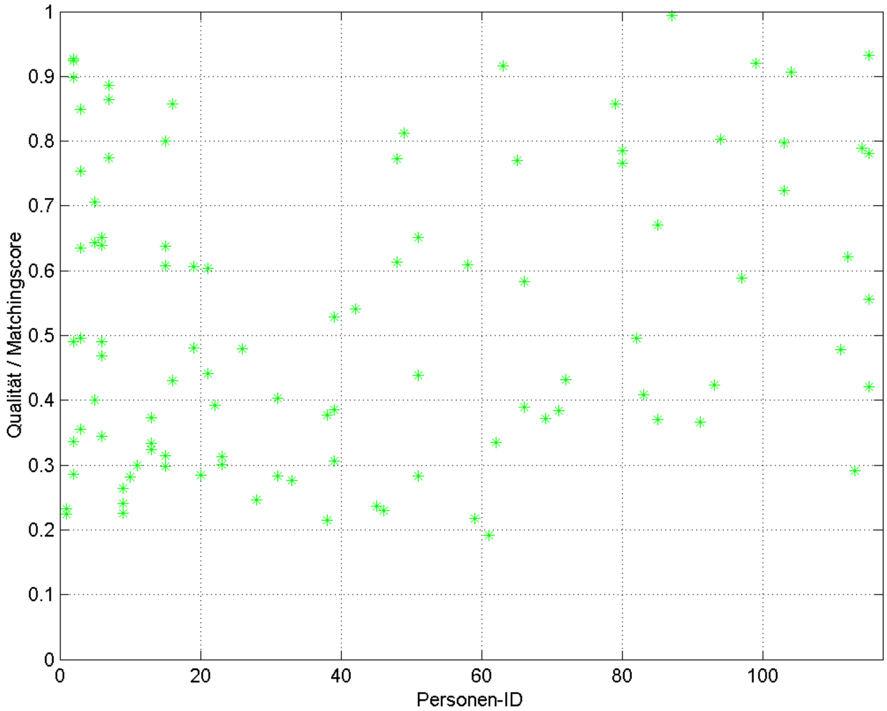


Figure 87: 4-3 Algorithm 3 correct hits

Qualität/Matchingscore	Quality/matching score
Personen-ID	Person ID

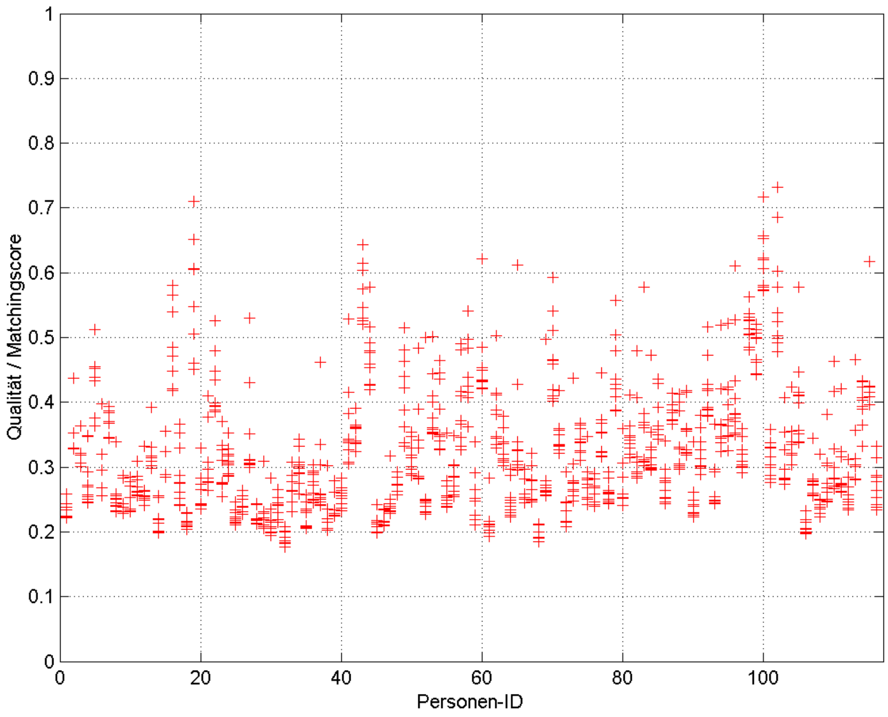


Figure 88: 4-3 Algorithm 3 false hits

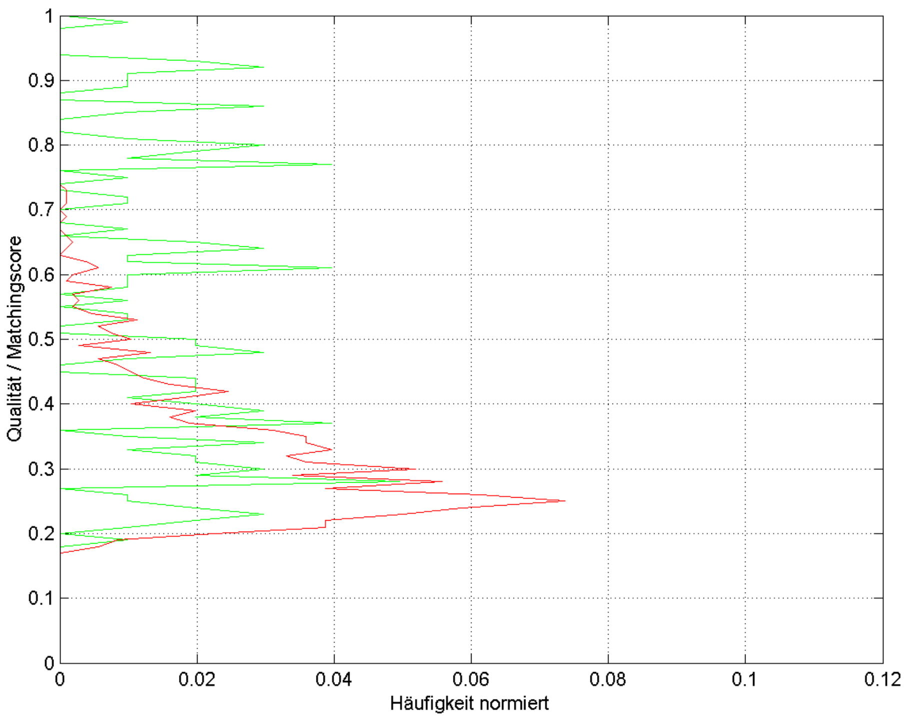


Figure 89: 4-3 Algorithm 3 total correct and false hits

Qualität/Matchingscore	Quality/matching score
Personen-ID	Person ID
Häufigkeit normiert	Normalised frequency

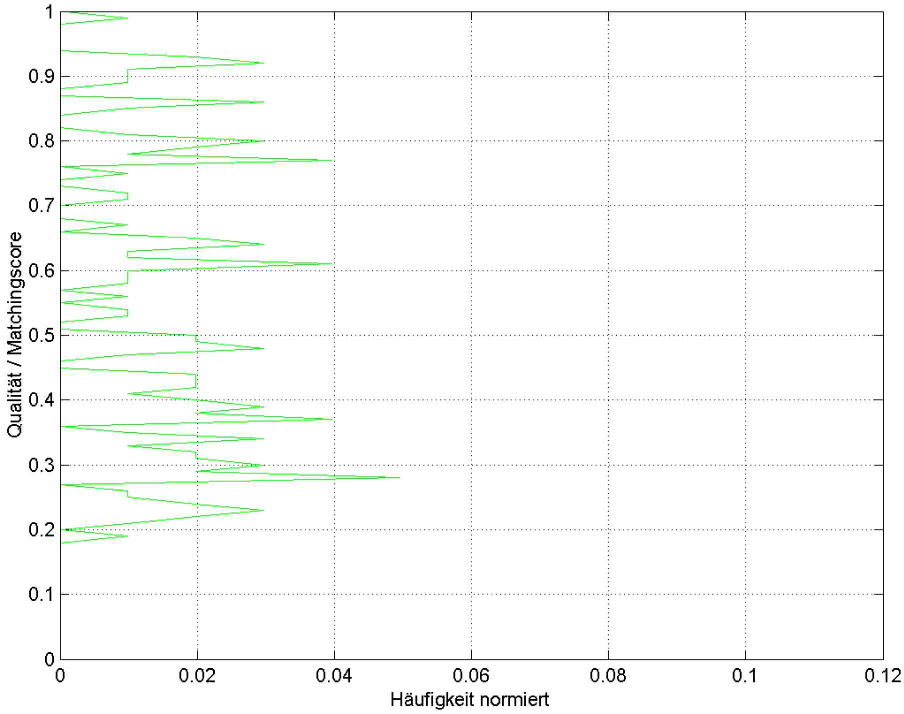


Figure 90: 4-3 Algorithm 3 total correct hits

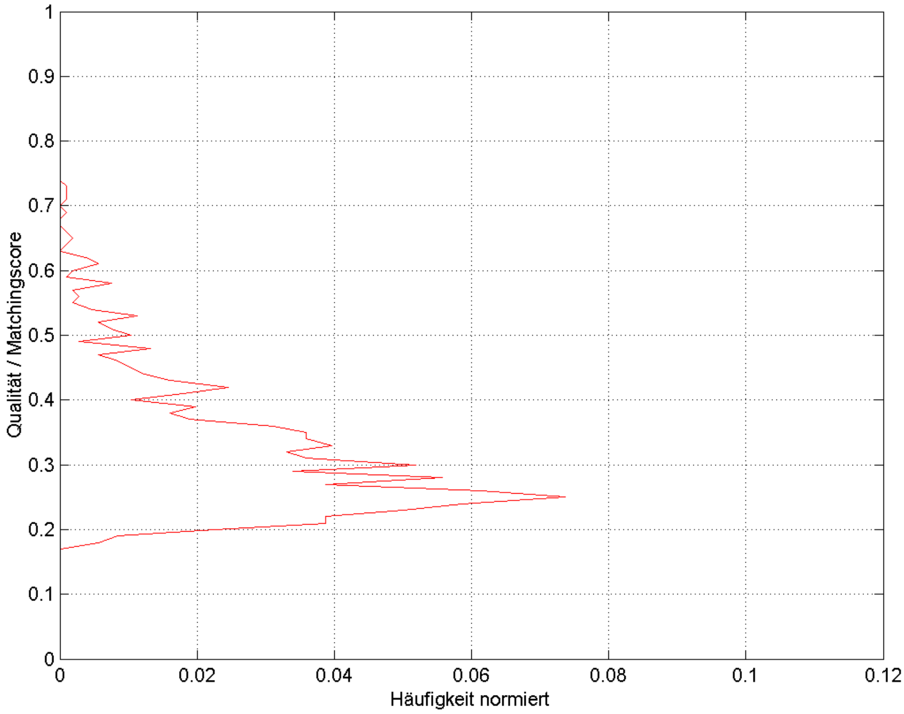


Figure 91: 4-3 Algorithm 3 total false hits

Qualität/Matchingscore	Quality/matching score
Häufigkeit normiert	Normalised frequency

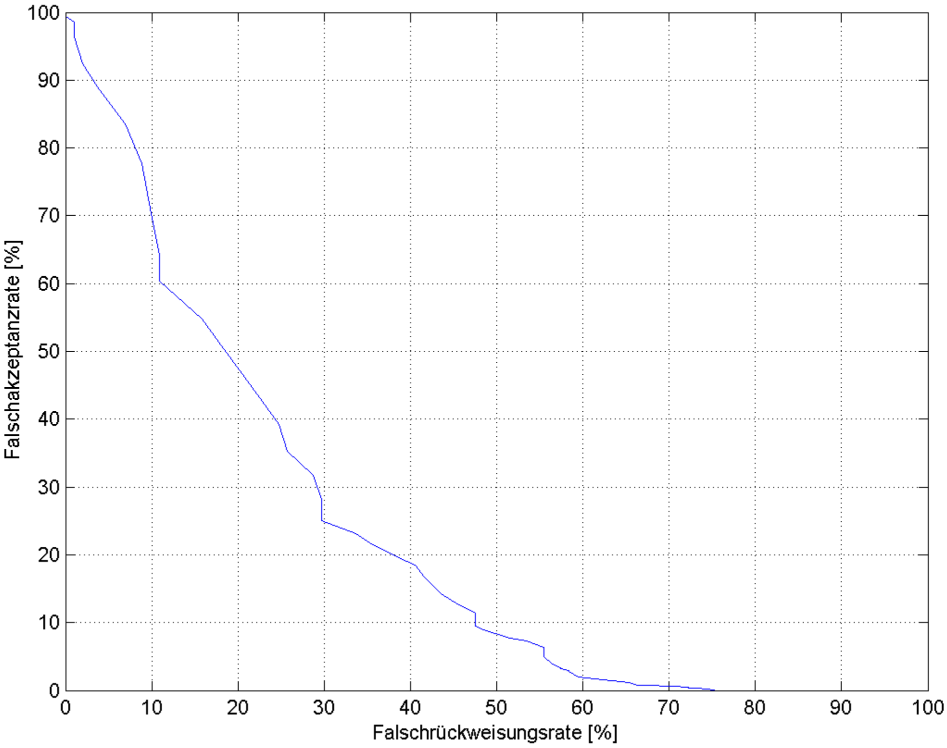


Figure 92: 4-3 Algorithm 3 receiver operating characteristics

Falschakzeptanzrate [%]	False acceptance rate [%]
Falschrückweisungsrate [%]	False rejection rate [%]

5.7.4.15 4-4-Algorithm 3

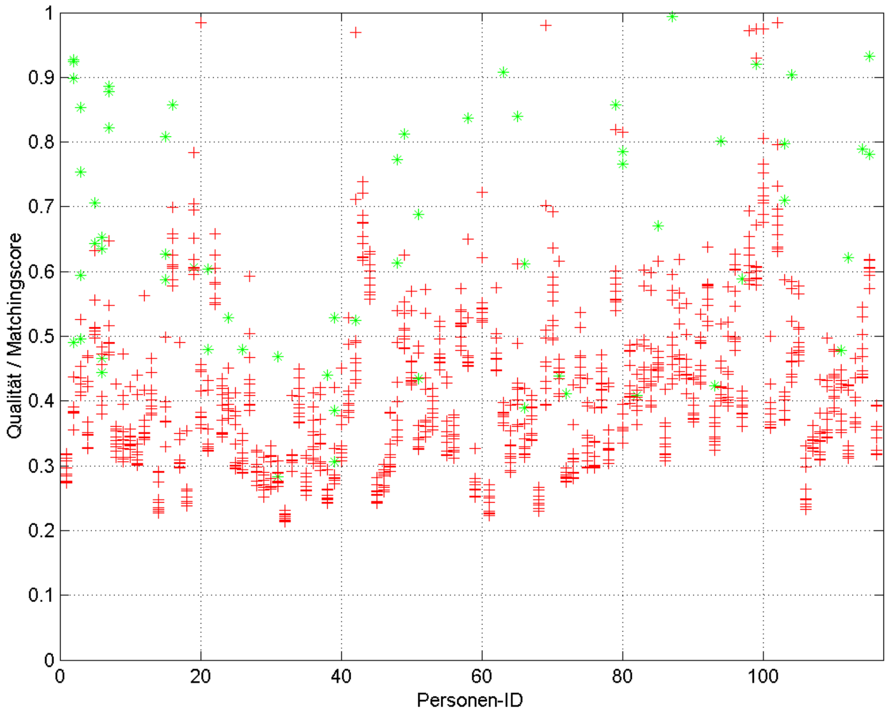


Figure 93: 4-4 Algorithm 3 correct and false hits

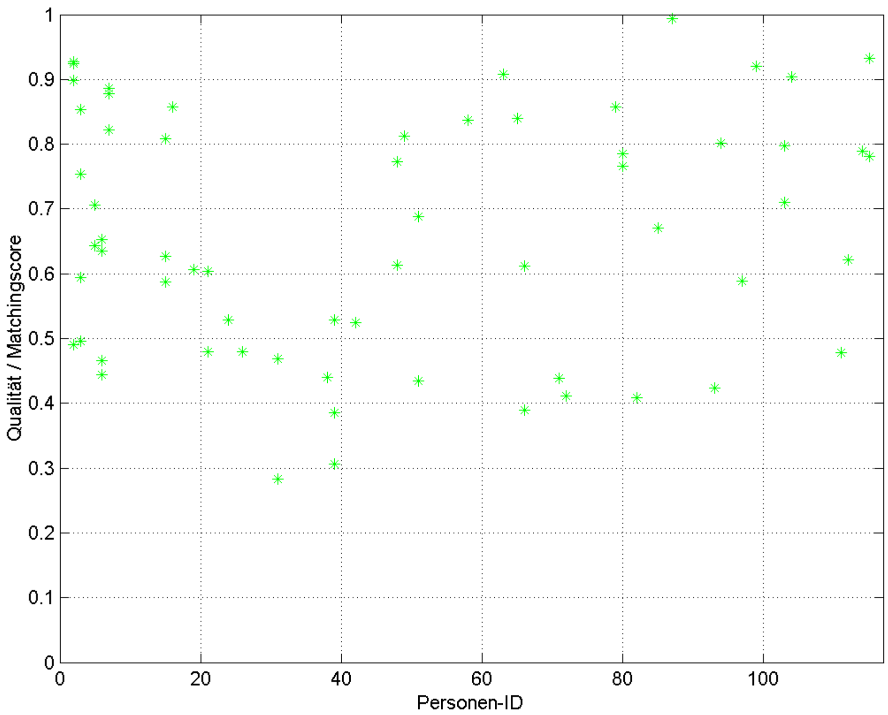


Figure 94: 4-4 Algorithm 3 correct hits

Qualität/Matchingscore	Quality/matching score
Personen-ID	Person ID

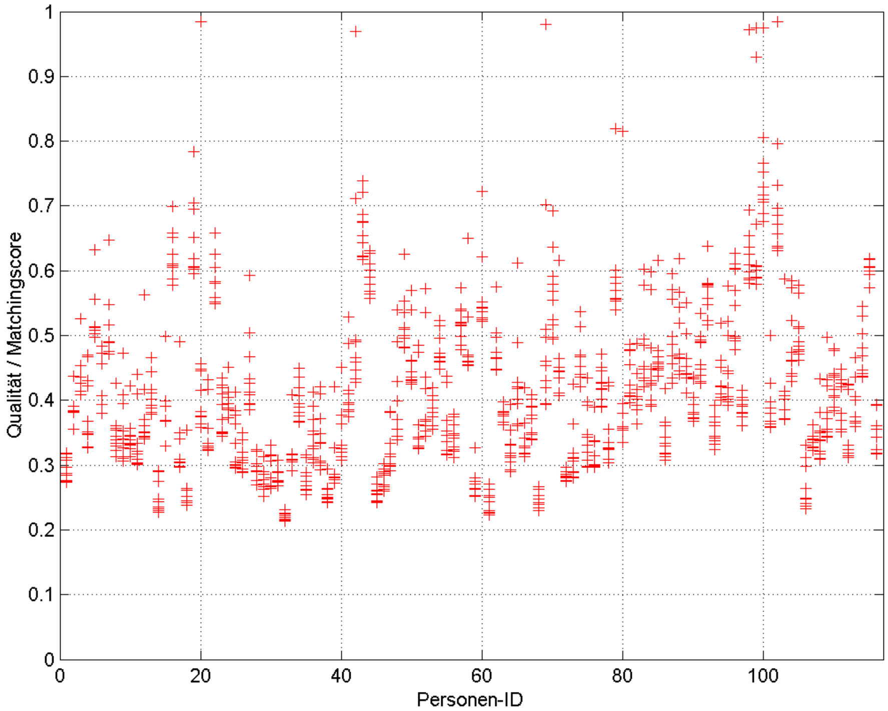


Figure 95: 4-4 Algorithm 3 false hits

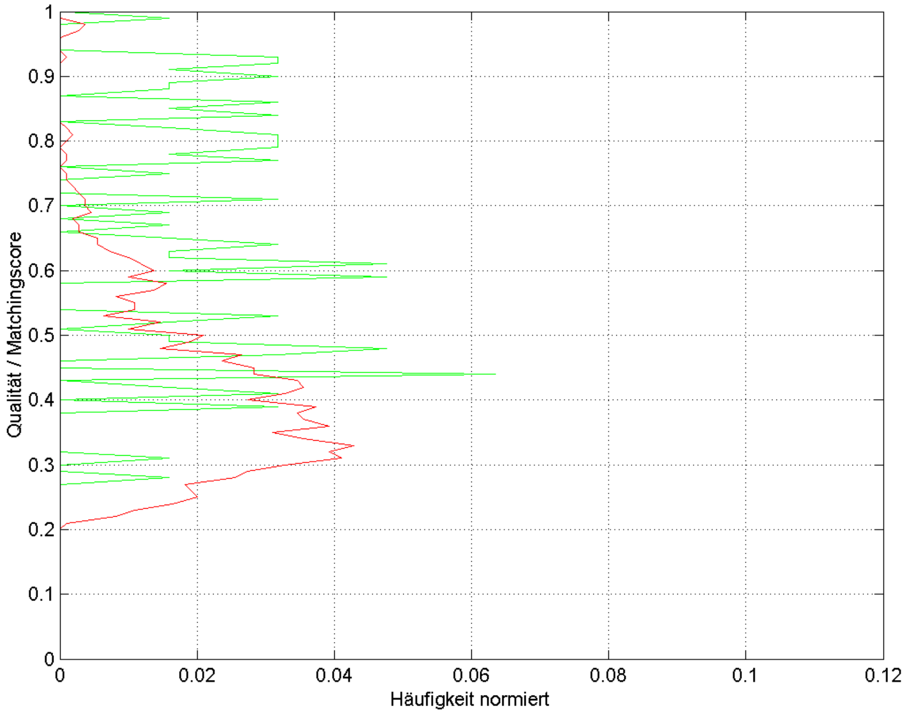


Figure 96: 4-4 Algorithm 3 total correct and false hits

Qualität/Matchingscore	Quality/matching score
Personen-ID	Person ID
Häufigkeit normiert	Normalised frequency

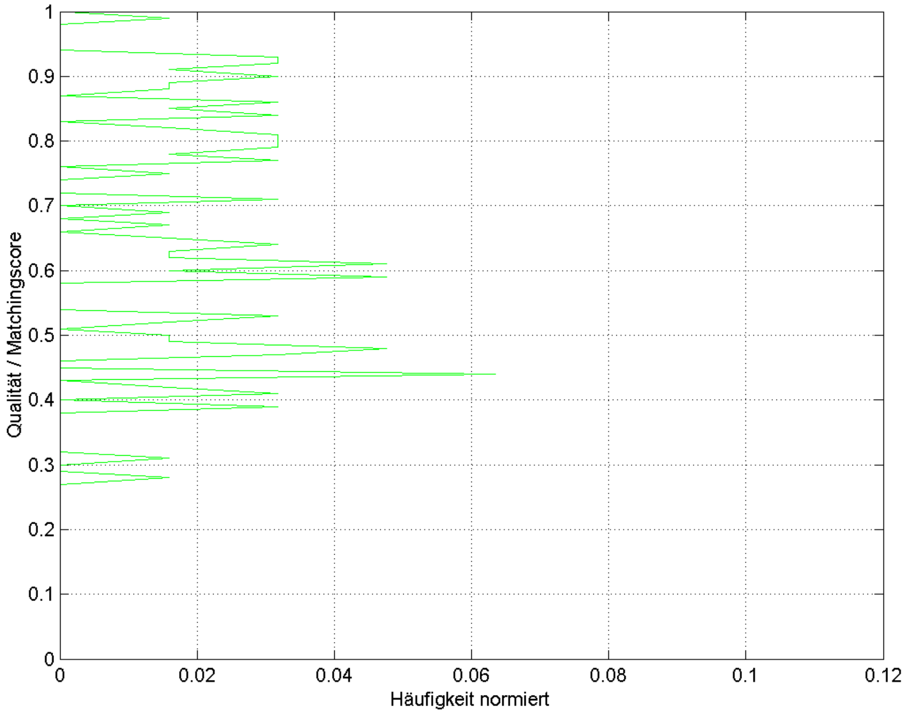


Figure 97: 4-4 Algorithm 3 total correct hits

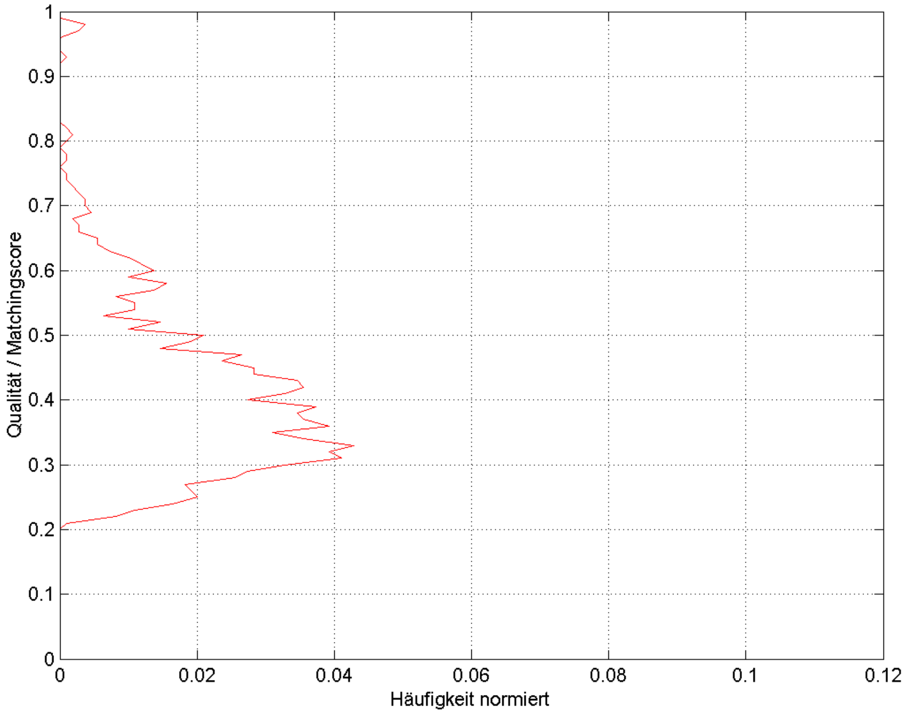


Figure 98: 4-4 Algorithm 3 total false hits

Qualität/Matchingscore	Quality/matching score
Häufigkeit normiert	Normalised frequency

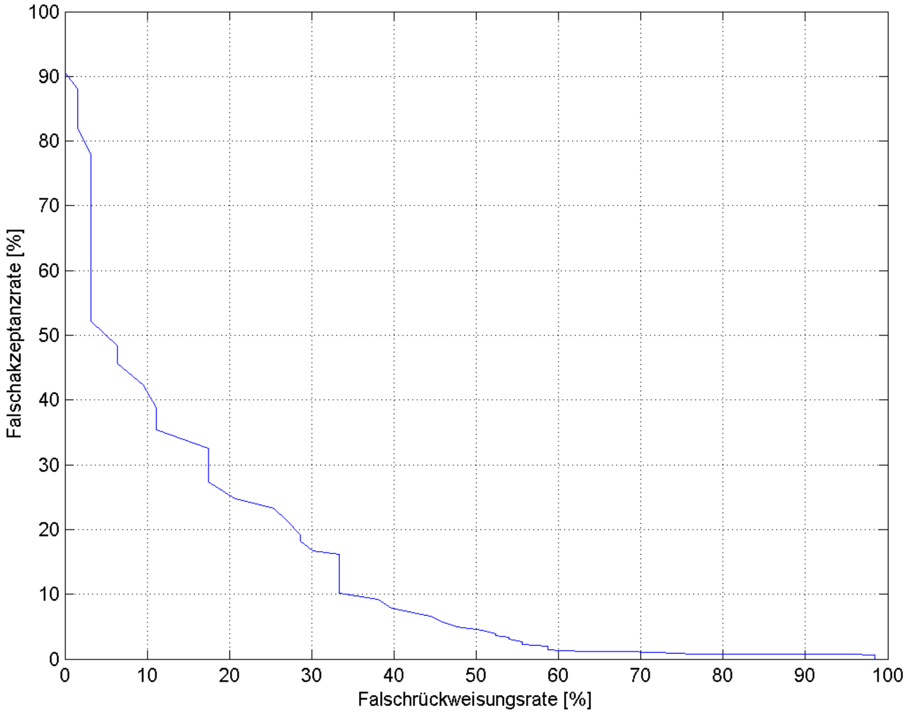


Figure 99: 4-4 Algorithm 3 receiver operating characteristics

Falschakzeptanzrate [%]	False acceptance rate [%]
Falschrückweisungsrate [%]	False rejection rate [%]

5.7.4.16 Interpretation of results

Unlike Test Group 1, in this case the results of all possible comparisons were not deliberately established though verification, but with the aid of identification the ten best matching scores for a search image were ascertained and recorded. This data record of ten results could thus be composed out of both hits, i.e. other images of the same person, and also non-hits, i.e. images of other persons. What these ten images have in common is simply the fact that they are the ten images with the best matching scores for the relevant search image.

Another important difference compared with Test Group 1 is that the image set that was used in some cases contained more than one image of a person and the results of these matches were also evaluated. There are therefore also diagrams that constitute the matching scores of correct matches (hits).

On the other hand, the three possible effects on the matching scores that an increase in the number of images in the reference set could have are unchanged: the matching scores could increase, remain the same or decline. Since, however, hits and non-hits were considered separately in this test group, nine possible results could be generated through combination.

If we consider the effects on the matches first of all, we can see from the total curves that the matching scores do not change noticeably. If, however, one compares the number of hits, it is striking that this declines as the size of the database increases. The reason for this is clear from the curve showing the total non-hits: the non-hits exhibit a clear upward shift for each incremental addition to the database size. This suggests that the number of non-hits with good matching scores had risen, so that they “jumped” above the matching score for a hit, which remained unchanged. In this way more and more hits were displaced by non-hits from the result data records containing 10 values. As a result, the number of hits fell, as is illustrated in the relevant graphs.

5.7.5 Test Group 5 long-term comparison

5.7.5.1 Enrolment and verifications

Test	Number of reference images	Algorithm 1	Algorithm 2	Algorithm 3
5-1	196	162	163	196
5-2	2	2	0	2
5-3	7	4	7	7
5-4	6	5	5	6
5-5	11	6	6	11
5-6	5	3	4	5
5-7	4	4	2	4
5-8	7	6	5	7
5-9	5	3	5	5
5-10	6	3	4	6
5-11	4	4	4	4
5-12	7	6	7	7
5-13	9	8	9	9
5-14	8	5	6	8
5-15	11	10	9	11
5-16	7	6	5	7
5-17	4	3	3	4
5-18	8	8	8	8
5-19	7	6	6	7
5-20	4	4	3	4
5-21	7	6	6	7
5-22	3	1	2	3
5-23	6	4	4	6
5-24	12	12	11	12
5-25	15	14	13	15
5-26	7	6	7	7
5-27	5	5	5	5
5-28	6	5	5	6
5-29	6	6	6	6
5-30	7	7	6	7

Table 27: Enrolments Test Group 5

Test	Algorithm 1	Algorithm 2	Algorithm 3
5-1	26,244	15,161[14]	38,416
5-2	0[15]	0[15]	0[15]
5-3	0[15]	0[15]	0[15]
5-4	0[15]	0[15]	0[15]
5-5	0[15]	0[15]	0[15]
5-6	6	9	20
5-7	12	2	12
5-8	16	4	26
5-9	6	10	20
5-10	6	11	30
5-11	10	3	10
5-12	28	21	40
5-13	56	39	72
5-14	20	7	50
5-15	68	46	86
5-16	30	11	42
5-17	6	4	12
5-18	54	45	54
5-19	24	18	36
5-20	2	0[16]	2
5-21	30	25	42
5-22	0[17]	0[18]	4
5-23	6	4	12
5-24	128	86	128
5-25	170	105	196
5-26	26	30	38
5-27	16	16	16
5-28	14	114	24
5-29	26	25	26
5-30	42	23	42

Table 28: verifications Test Group 5

[14] The low number of analysable comparisons is due to the fact that not all the results were reported.

[15] No analysis could be formed in the tests, as no accurate age information was available for the images.

[16] In this subset, age information was available for only two images, neither of which was accepted as a comparison image.

[17] None of the images was accepted as a comparison image.

[18] In this case, only two images were accepted, taken 16 years apart. Hence there were no analysable results.

5.7.5.2 Notes on the diagrams[19]

For Tests 5-1, in each case two diagrams were prepared. The first of these is a graph whose x-axis presents the difference between the two images compared in years, with the relevant average matching score shown on the y-axis. The x values always constitute the difference between comparison image and search image. Negative values mean here that the comparison image was older than the search image (e.g. comparison image 1990, search image 1998 difference = 1990-1999 = -9). The axial symmetry of the images is partly due to the fact that each image was used each as search image and as comparison image. Thus, first of all image A was compared with image B and then image B was compared with image A. As the matching scores were normally very similar, whereas the age differences were symmetric about the difference line of 0 years (e.g. A B difference 9 years, difference B A -9 years), the result is a diagram that is virtually symmetric about the line.

The second graph shows how many corresponding comparisons could be carried out. It is clear that the smaller the difference in time, the higher the number: relevant photographs are more frequent.

The diagrams for tests 5-2 to 5-30 (to the extent that analysable results are available) are contained in the annex to make the main body of the report more digestible.

[19] See also 5.7.2 Notes on the interpretation of the diagrams.

5.7.5.3 5-1 Algorithm 1

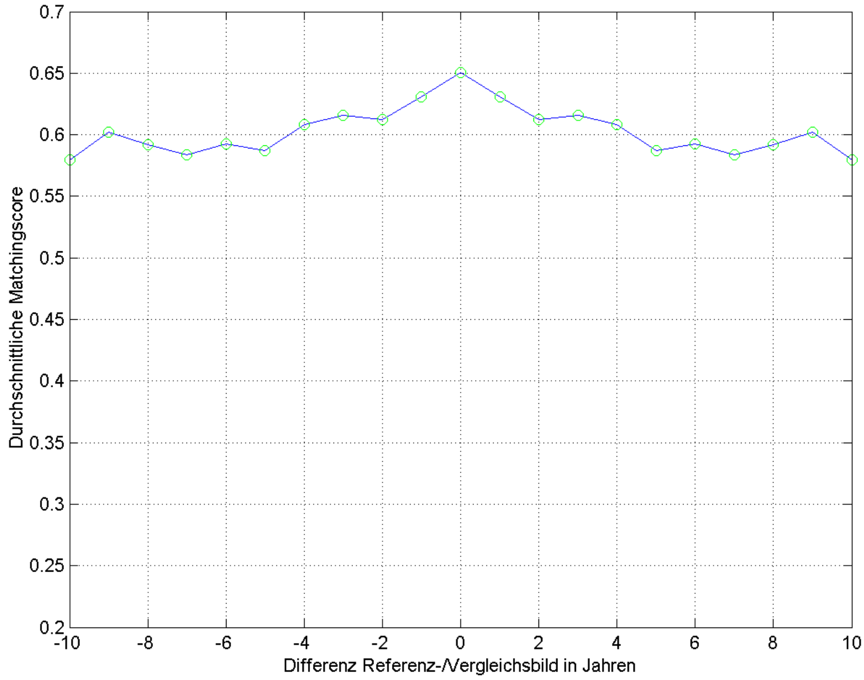


Figure 100: 5-1 Algorithm 1

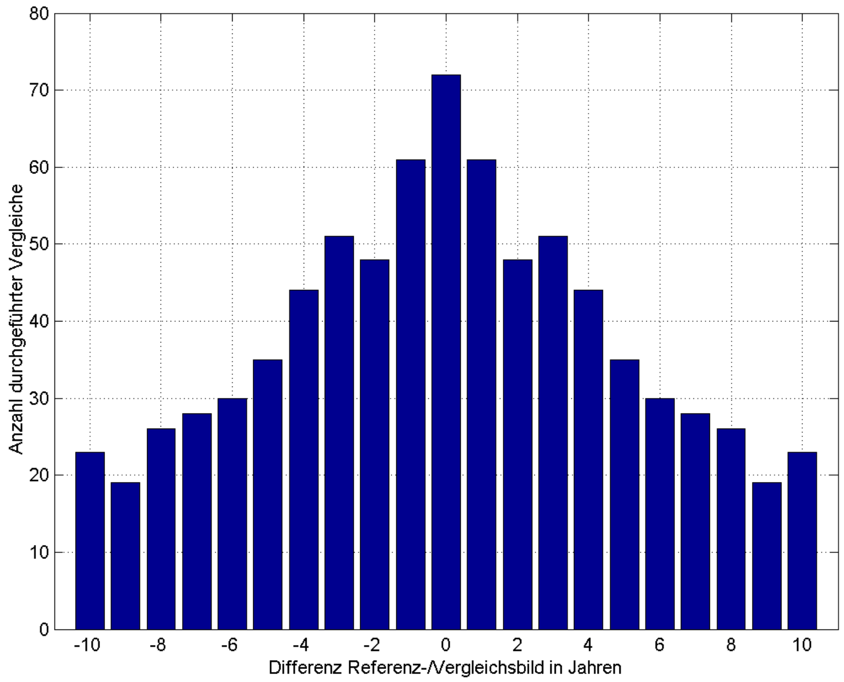


Figure 101: 5-1 Algorithm 1 number of comparisons carried out

Durchschnittliche Matchingscore	Average matching score
Differenz Referenz-/ Vergleichsbild in Jahren	Difference between reference and comparison images in years
Anzahl durchgeführter Vergleiche	Number of comparisons carried out

5.7.5.4 5-1 Algorithm 2

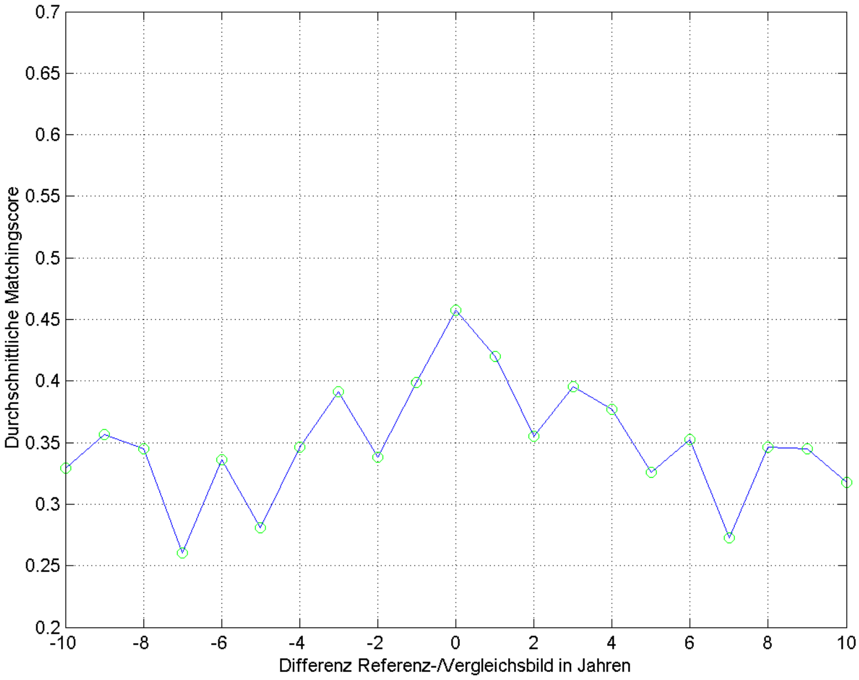


Figure 102: 5-1 Algorithm 2

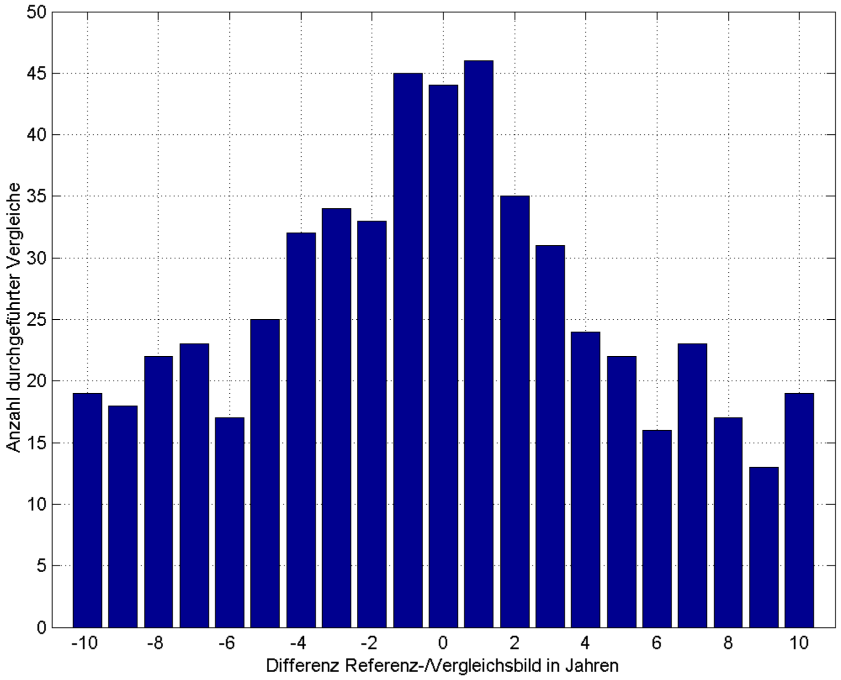


Figure 103: 5-1 Algorithm 2 number of comparisons carried out

Durchschnittliche Matchingscore	Average matching score
Differenz Referenz-/ Vergleichsbild in Jahren	Difference between reference and comparison images in years
Anzahl durchgeführter Vergleiche	Number of comparisons carried out

5.7.5.5 5-1 Algorithm 3

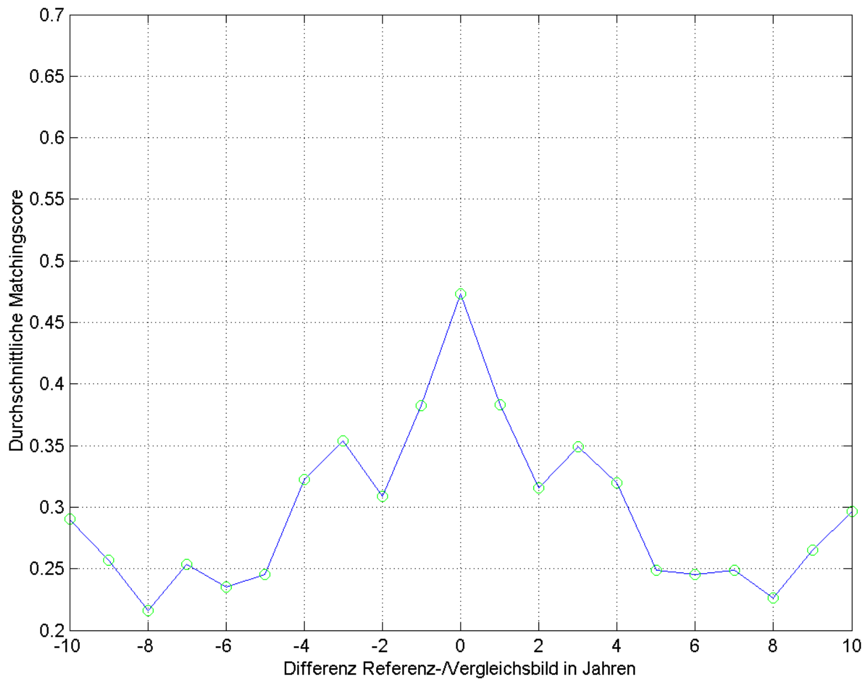


Figure 104: 5-1 Algorithm 3

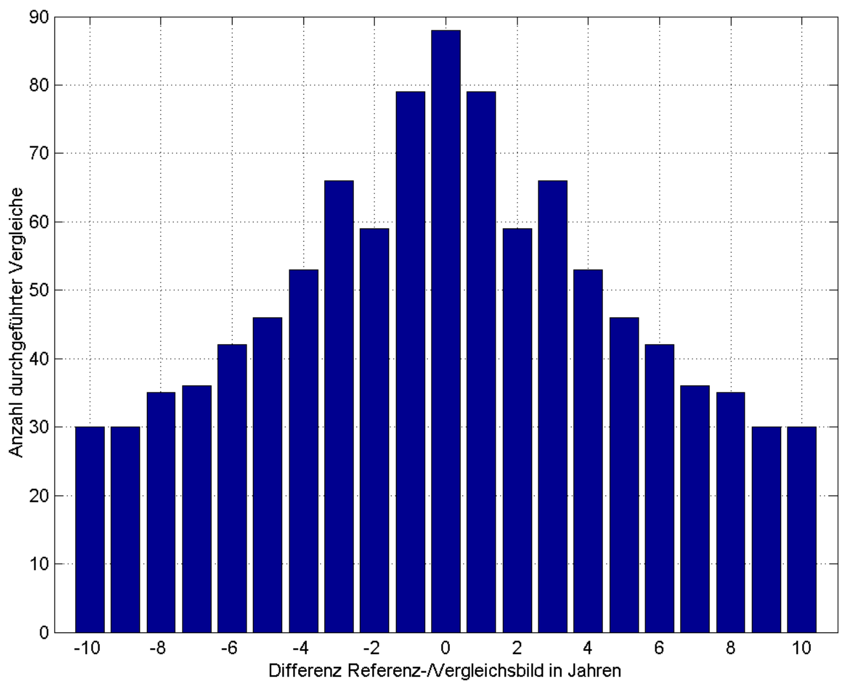


Figure 105: 5-1 Algorithm 3 number of comparisons carried out

Durchschnittliche Matchingscore	Average matching score
Differenz Referenz-/ Vergleichsbild in Jahren	Difference between reference and comparison images in years
Anzahl durchgeführter Vergleiche	Number of comparisons carried out

5.7.5.6 Interpretation of results

The diagrams for the number of comparisons carried out provide a clear and logical picture: the lower the interval in years, the more images are contained in the database. Looked at in another way, this means that the wider the interval, the more difficult it becomes to find matching image material and accordingly the more unusual it is as well.

The results are what one would expect. Viewed from an abstract point of view, the further apart in time that the photographs were taken, the lower the matching score. The increase that is to be observed in the case of comparisons a long way apart (nine or ten years) is probably due to the fact that by chance better image material was available for that time than in comparisons with shorter (7 or 8 years) time intervals. This is likely to also be the reason for the irregularities in the graphs. A larger number of corresponding images would probably reduce the variance here and hence smooth out the curve.

It is interesting to note that, as one would expect (see 5.7.5.2 Notes on the diagrams), the axial symmetry is spoilt in some diagrams (Figure 102: 5-1 Algorithm 2, Figure 104: 5-1 Algorithm 3). Evidently in those cases the images were treated differently, depending on whether they were being used as reference or comparison images. This suggests that an image was scored differently, depending on whether the image served as reference image or as comparison image.

5.7.6 Test Group 6 long-term comparison in large reference set

Due to the less linear results, the sponsor decided not to carry out the optional Test Group 6 5.6.5 Test Group 6 long-term comparison in large reference set).

5.7.7 Test Group 7 other tests

This group of tests served to perform various plausibility checks.

5.7.7.1 Notes on the diagrams[20]

For each investigation carried out in this test group, in each case verifications of images for the same and different persons were considered separately. The matching scores achieved were rounded to two decimal places and summed.

[20] See also 5.7.2 Notes on the interpretation of the diagrams.

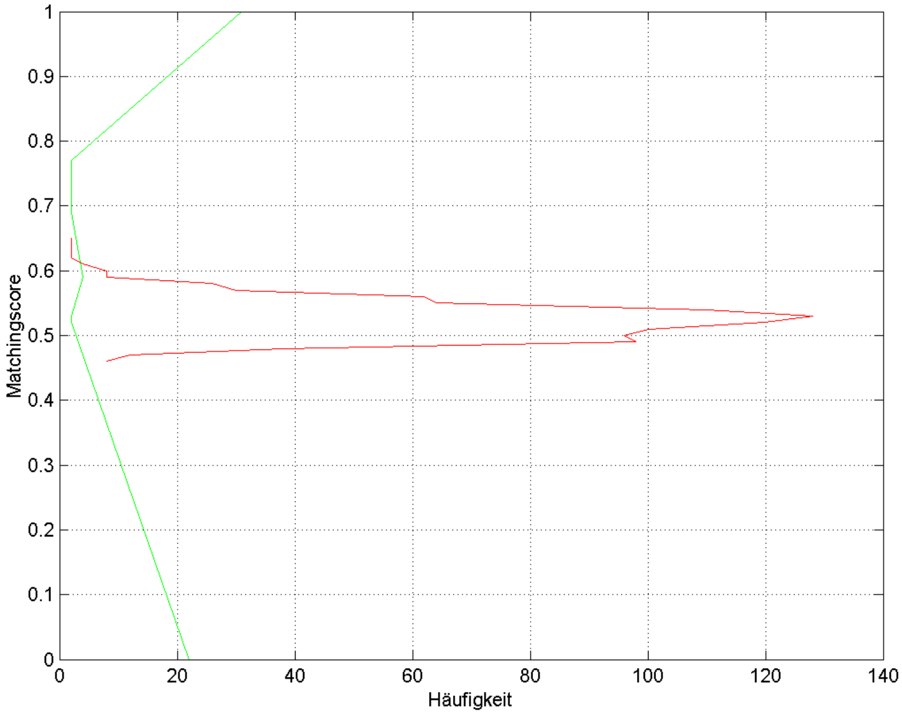


Figure 106: 7-1 Algorithm 1

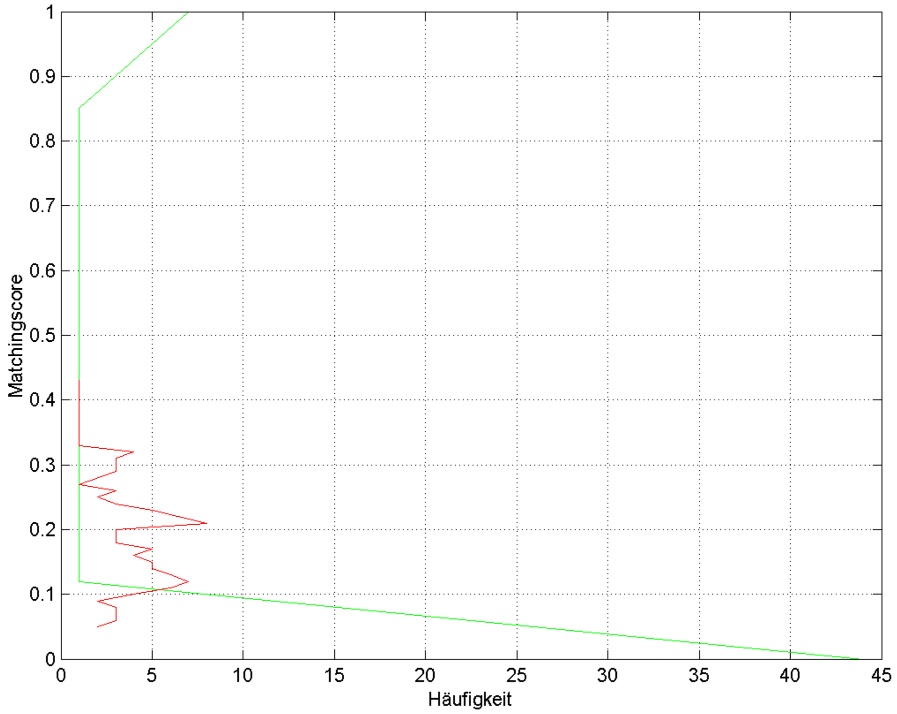


Figure 107: 7-1 Algorithm 2

Matchingscore	Matching score
Häufigkeit	Frequency

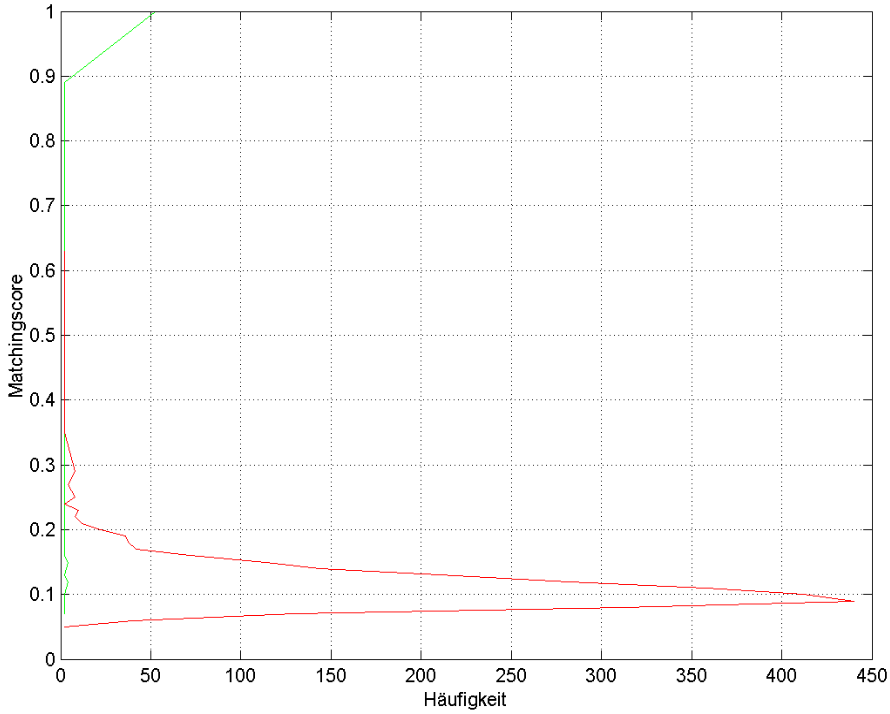


Figure 108: 7-1 Algorithm 3

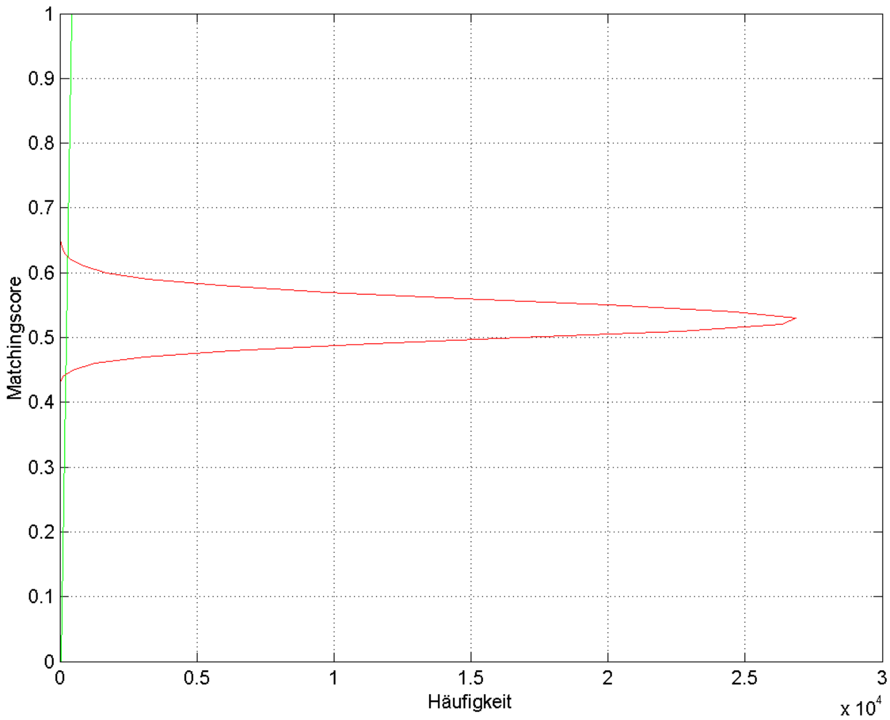


Figure 109: 7-2 Algorithm 1

Matchingscore	Matching score
Häufigkeit	Frequency

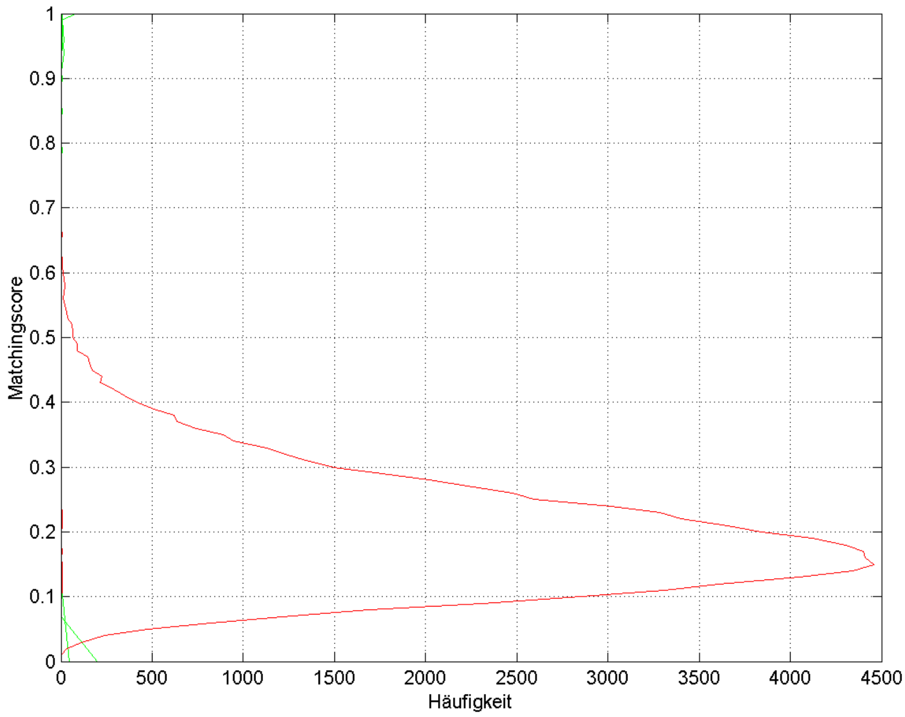


Figure 110: 7-2 Algorithm 2

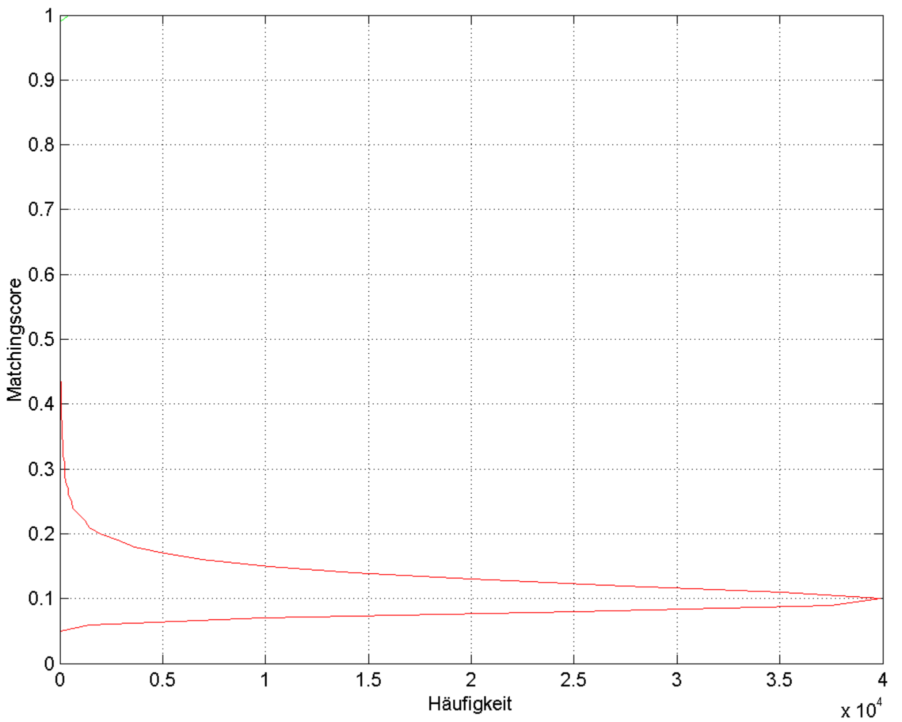


Figure 111: 7-2 Algorithm 3

Matchingscore	Matching score
Häufigkeit	Frequency

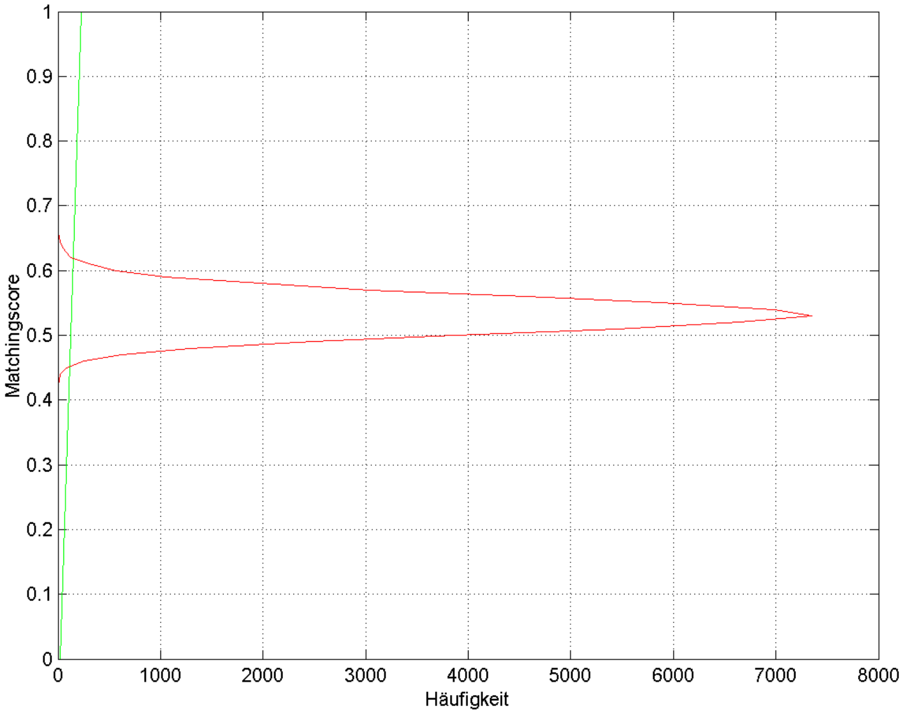


Figure 112: 7-3 Algorithm 1

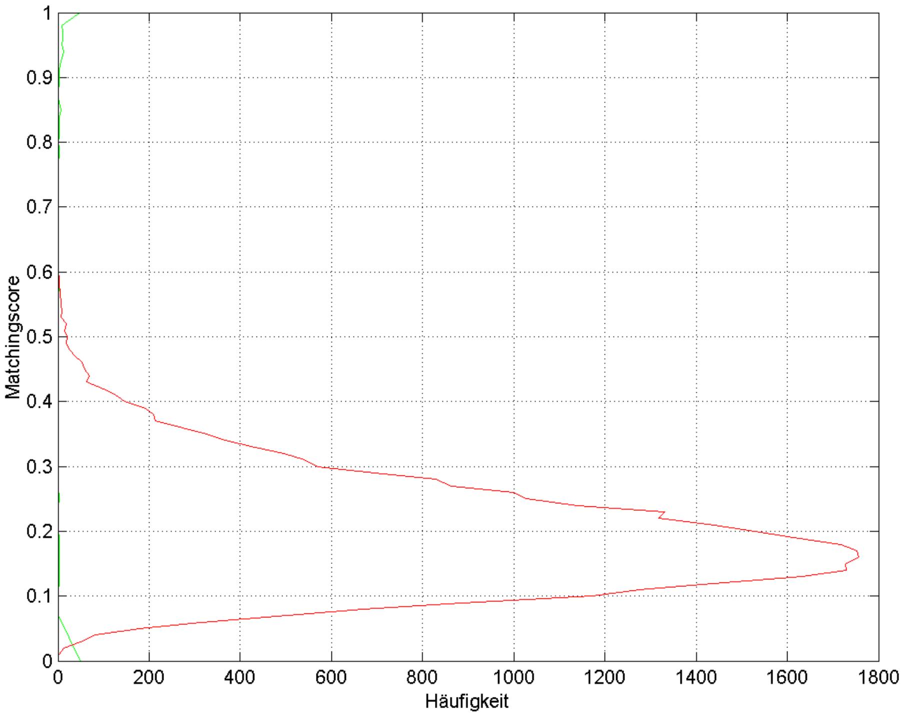


Figure 113: 7-3 Algorithm 2

Matchingscore	Matching score
Häufigkeit	Frequency

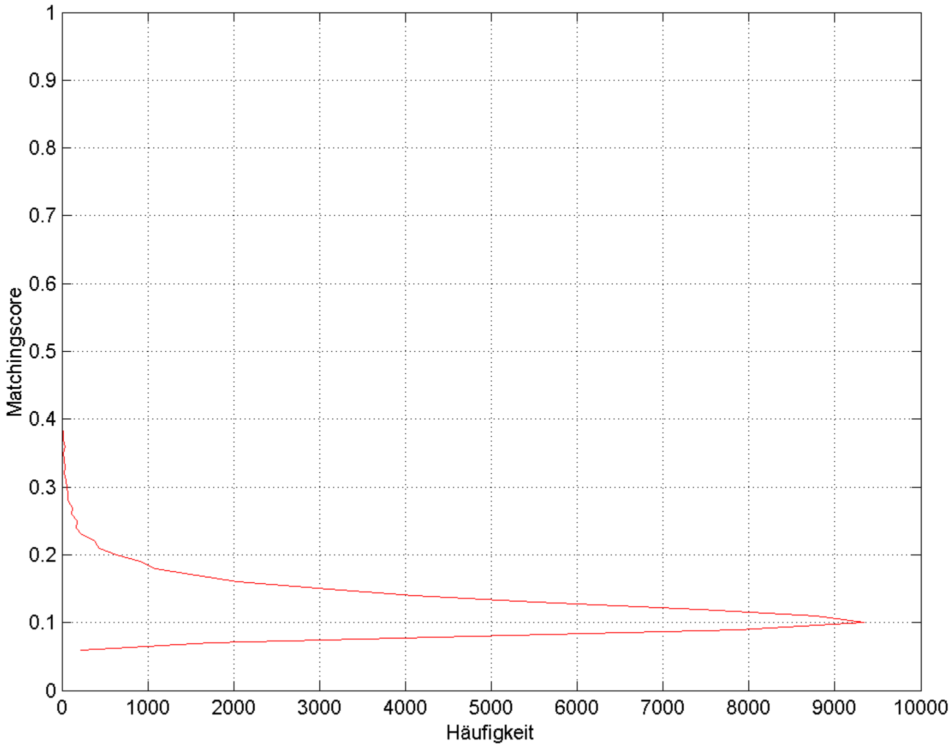


Figure 114: 7-3 Algorithm 3

5.7.7.2 Interpretation of results

Given that the experimental set-up was similar to that used for Test Group 1, the results obtained for the non-hits in Test 7-1 can be compared the results for Test Group 1. As the matches were irrelevant in Test Group 1, naturally these must be ignored.

The difference compared with Test Group 1 lies in the image material used: whereas Test Group 1 made use of the entire image stock, in the 7-1 tests only images from the FICGR image collection were used. In Test 7-1, using exclusively images from the FICGR image collection, a check was carried out as to whether mixing the test images in the main tests would have a material effect on the results.

After comparing the curves for non-hits in Test Group 1 and those from the 7-1 tests, it is clear that no such effect occurred.

Tests 7-2 and 7-3 were intended to check in a very abstract way whether random false classifications and the thus noisy image material that was incorrectly used in the tests could render the results unusable. But once again, the changes were very small. However, the proportion of noisy images was significantly higher than in the other test groups and the size of the reference set was small. The influence of image material incorrectly classified as good has had a negligible impact on the test results.

Matchingscore	Matching score
Häufigkeit	Frequency

6 BioFace II Phase 2: system test

As well as the algorithm test carried out in Phase 1, BioFace II included a comparison of independently working facial recognition systems. Several installations consisting of camera and computer hardware were installed and operated in the entrance area to the Federal Office of Criminal Investigation.

6.1 The test environment

The tests took place at the Federal Office of Criminal Investigation's Wiesbaden premises, Äppelallee 45, 65203 Wiesbaden. Twenty employees (subjects) of the BKA had volunteered to take part in the system test for two periods of six weeks.

The system test was integrated into the normal access control process for employees. On arrival at the BKA site access control point at the Äppelallee entrance, the subjects passed several turnstiles. They then registered themselves on a time registration system. During the test phase, the route from the central turnstile to the time registration terminal was also monitored by the facial recognition systems (see Figure 116: schematic representation of the entrance area). For the duration of the tests, the time registration system provided information on whether and when the subjects had entered the building.

Measuring point (see Figure 117)	4.45am. Dark, fluorescent tubes above the entrance are in operation	5.40am. Dawn, fluorescent tubes above the entrance are in operation	6.15am. Dawn, fluorescent tubes above the entrance are in operation	7.15am. Bright, fluorescent tubes above the entrance are switched off	7.50am. Bright, fluorescent tubes above the entrance are switched off
0	End 120[21]	End 120	End 120	Centre 30[22]	End 30
1	End 250	End 250	End 250	End 30	Centre 60
Betw. 1 & 2	Centre 500	Centre 500	Centre 500	End 60	End 60
2	End 250	End 250	End 250	Centre 120	Centre 120
3	Centre 30	Centre 30	Centre 30	End 120[23]	Centre 120[24]

Table 29: Lighting conditions at the entrance to Äppelallee 45, 19 August 2002. Weather conditions: light cloud, 21ºC. Measurements were taken with an FGL photometer (hand-held measuring instrument)

In parallel to this, each facial recognition system generated a log which recorded when which subject was recognised. By matching the reliable data from the time registration system with the log produced by the facial recognition systems, it was possible to detect and quantify correct and incorrect recognitions.

As the area monitored was not a closed indoor area, it was important to check the lighting conditions, as too little or too much illumination of faces could impair recognition performance. Therefore the lighting conditions in the installation area were measured in advance of installing the equipment (Table 29).

[21] "End 120" means that the pointer of the instrument was in the position where field 120 transitions to field 250.

[22] "Centre 30" means that the pointer of the instrument stood at the centre of the field labelled 30.

[23] Rising sun was shining into the measurement area.

[24] No direct sunshine any longer.

It transpired in the course of a preliminary pilot phase that the existing lighting was inadequate as it illuminated persons precisely in the catchment area directly from above. Additional light fittings were therefore installed above the camera installations (see Figure 116: schematic representation of the entrance area, description “LIGHT NEW”). To prevent any negative effects or interactions between the different vendors resulting from their own lighting installations, this change was agreed with and accepted by all the vendors.

In order to be able to carry out a more detailed analysis of errors, the camera data (video signals) for one facial recognition system was recorded. In this way, a parallel analysis was carried out for all persons as to whether recognition was possible at all. If recognition was not possible for the systems, this was logged.

Figure 115: Entrance area to the Federal Office of Criminal Investigation

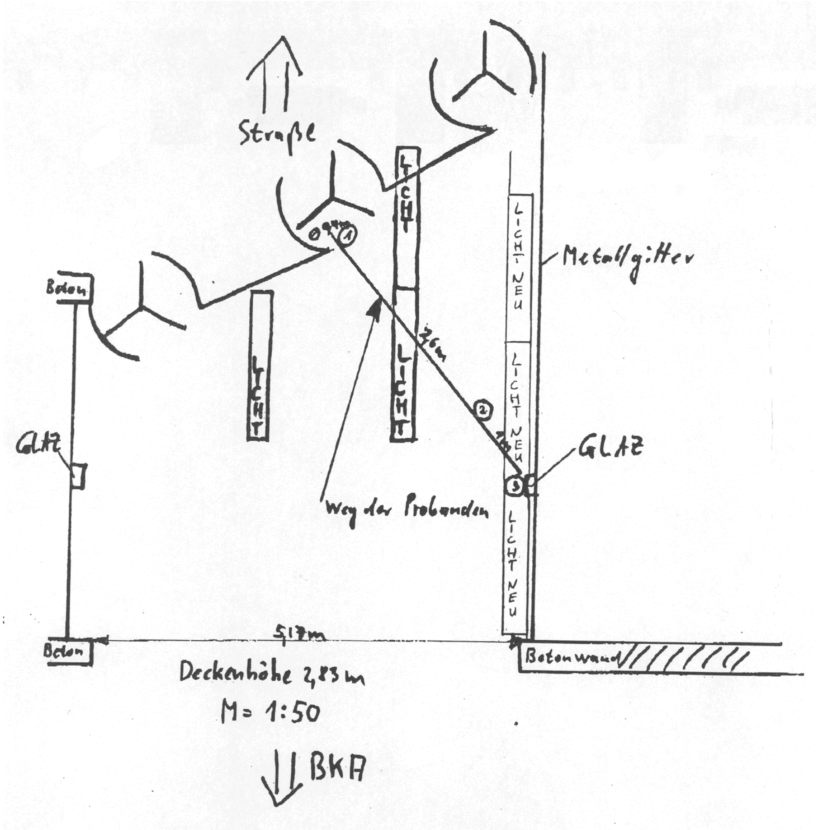


Figure 116: schematic representation of the entrance area

Straße	Street	5,17m	5.17m
LICHT	LIGHT	Deckenhöhe 2,83m	Ceiling height 2.83 m
GLAZ	GLAZ time registration system	M = 1:50	Scale = 1:50
Beton	Concrete	BKA	BKA
Weg der Probanden	Path taken by subjects	Betonwand	Concrete wall
LICHT NEU	LIGHT NEW	Metallgitter	Metal grid

Figure 117: overview of the system test installation

Figure 118: the camera installation as viewed by the subjects

Figure 119: the camera installation including the additional lighting

Figure 120: computer installation in the booth to the right

Figure 121: computer installation in the booth to the left

Error code	Description
1	No error detectable for the person
10	Turnstile not used
11	Position of head
12	Direction of gaze
13	Headgear
14	Facial expression (e.g. eyes closed)
15	Outside hours during which the video control recording collected data
16	Camera covered
50	Absence (holiday, business trip)
51	Time logging not used

Table 30: error codes used for video control recording

To simulate the behaviour that would occur with a large number of enrolled users, as well as the 20 actual subjects, images of another 500 imaginary users were also enrolled into the user database of the facial recognition systems. These had been picked at random from the test images available (see section 4.3). After six weeks, the images from live enrolment (BioFace II Phase 2.1) were replaced by digital photographs of the subjects (BioFace II Phase 2.2).

6.2 Procedure

The subjects entered the BKA grounds in the morning between 6am and 9am. They used the centre turnstile and registered themselves on the time registration system. Each day, at the end of the evaluation period, the log files from the facial recognition systems were exported. In parallel to this, the time registration logs were also extracted. The parallel video recording was analysed and the error codes entered into the time registration table.

Test	Start date	End date
BioFace II.1	1 November 2002	12 December 2002
BioFace II.2	3 February 2003	14 March 2003

Table 31: testing periods for test phase II

The recognition logs and the time registration table with the error codes from the evaluation of the video recording were e-mailed to the Fraunhofer Institute for Computer Graphics Research in pseudonymised form each day. To ensure the confidentiality of the data, it was sent encrypted, using PGP.

This scenario normally required no additional actions on the part of the subjects.

With regard to the collection of data, care had to be taken to ensure that where a subject passed the monitored areas more than once during a surveillance period, only the first occasion was logged in the time recording. To avoid the possibility that in this case, due to missing entries in the logged time registration data, a non-recognition would wrongly be recorded, steps had to be taken to ensure that the subjects used an unsupervised entry or exit point.

6.3 The test database

An SQL database (PostgreSQL) was used to store the recognition events and time registration data.

Figure 122 shows the database design that was used to store the data from the system test (Phase 2).

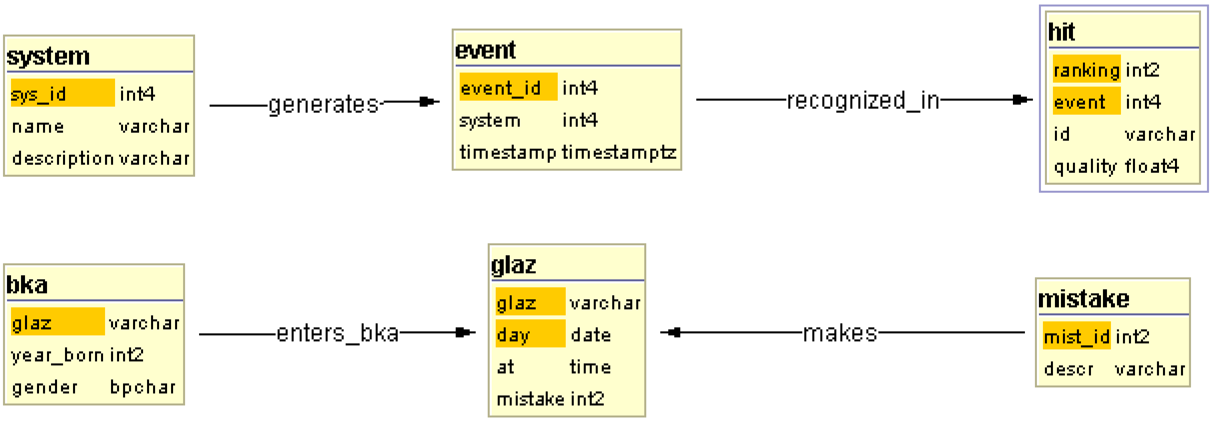


Figure 122: database structure BioFace II Phase 2

The database for the second phase of BioFace II was structured in two parts. The first part contained the verified log data from the time registration in the “glaz” table, which contained the GLAZ number (“glaz”), the day, the time and any relevant error code (see Table 30: error codes used for video control recording). The meaning of the error code was stored in “mistake”, details relating to the subject in “bka”. These two tables were both referenced from the “glaz” table.

The second part of the database containing the recognition events. For each recognition event there was a corresponding entry in the “event” table, whereby the time and the relevant system were also recorded. The hits returned during the event were stored in “hit” and linked with the event via the unique “event_id”. Here the ranking of the hit data record (“ranking”), the matching score (“quality”) and the ID ostensibly found (“id”) were returned. There was also an assignment of system IDs to manufacturers with a corresponding description in the “system” table.

6.4 Results

6.4.1 Recording of results

Because the subjects’ passage through the monitored area was independently recorded by parallel capture of times for flexitime purposes, it was possible to determine precisely when which subject had passed the camera installation. However, the fact that the flexitime time recording system did not record seconds, but only the hour and minute, complicated matters. As the computer clocks and the flexitime recordings were not automatically synchronised (e.g. via the DCF77 signal[25]), a longer period than planned had to be used for the evaluation. Under optimal conditions (with synchronised clocks), recognition would have occurred in the period 0-10 seconds prior to the flexitime time recording. Recognitions before or after this window would then have been false acceptances. Since, however, the clocks were synchronised manually once a week, to be on the safe side a deviation of up to 1 minute was assumed. Moreover, the flexitime recordings did not provide any second data, so that the period to be evaluated was determined at ±2 minutes. The following example explains the problem. If the flexitime recording was one minute earlier than the computer clock, the time

[25] DCF77 is a time signal that is broadcast as time information on frequency 77.5 kHz (see http://www.dcf77.de).

between registration by the facial recognition system and flexitime recording is set at 3 seconds and the flexitime registration is assumed to have taken place at 8:12:01 hrs., then for the purposes of the evaluation we get “8:12 hrs.” from the flexitime data, whereas the corresponding log entry in the system would be “8:10 hrs.”. This is arrived at as follows: 8:12:01 hrs. – 3 seconds’ walk – 1 minute deviation = 8:10:58 hrs.

The evaluation therefore proceeded as follows: for each event recorded in the flexitime recordings, the log files for the facial recognition systems were searched for a suitable recognition in the period ±2 minutes. If a recognition event that matched the person was found, then this was saved in the database table for correct recognitions. All other events within the data recording hours of 5.45am to 9.15am[26] were stored as false recognitions. As the best five events were expected for each recognition event, the number of false scores naturally predominated (as can be seen in Figure 126: results for BioFace II.2.1 System D on page 123, for example).

If, despite registering on the flexitime system, a person was not recognised by any of the installed systems, checks were carried out each day as to whether recognition was in fact possible or whether, for example, the person had unintentionally used the wrong turnstile. These events in the flexitime recordings were ignored.

6.4.2 Correct identifications

Algorithm	Evaluated test days	Hits in position 1	Hits in position 2	Hits in position 3	Hits in position 4	Hits in position 5
System A	29	64	29	34	-	-
System B	28	8	-	-	-	-
System C	29	90	26	16	6	6
System D	29	35	2	-	5	-

Table 32: correct identifications in BioFace II.2.1

Algorithm	Evaluated test days	Hits in position 1	Hits in position 2	Hits in position 3	Hits in position 4	Hits in position 5
System A	28	31	26	34	-	-
System B	8	1	-	-	-	-
System C	28	29	19	19	6	10
System D	27	1	-	-	1	-

Table 33: correct identifications in BioFace II.2.2

26 The period deviated here, with the sponsor’s agreement, from the period specified for the subjects of 6am to 9pm in each case by 15 minutes, as during the evaluation it turned out that a material number of persons passed the installation either slightly early or slightly late.

Algorithm	False rejection rate (FRR) BioFace II.2.1	False rejection rate (FRR) BioFace II.2.2
Total number of possible recognitions	402	335
System A	68%	73%
System B	98%	(99.7%)[27]
System C	64%	75%
System D	90%	99%

Table 34: false rejection rate (FRR) for possible recognitions

Person ID	Flexitime events	System A	System B	System C	System D
1	13	38% (5)	8% (1)	15% (2)	23% (3)
2	19	47% (9)		68% (13)	5% (1)
3	22	50% (11)		27% (6)	14% (3)
4	16	56% (9)	6% (1)	31% (5)	
5	16	44% (7)			19% (3)
6	16	19% (3)	6% (1)	88% (14)	19% (3)
7	20	40% (8)	10% (2)	45% (9)	5% (1)
8	17	35% (6)		65% (11)	12% (2)
9	17	29% (5)		53% (9)	29% (5)
10	15	33% (5)		80% (12)	7% (1)
11	19	37% (7)	5% (1)	74% (14)	5% (1)
12	19	37% (7)	11% (2)	37% (7)	16% (3)
13	20	25% (5)		10% (2)	10% (2)
14	9	11% (1)		33% (3)	44% (4)
15	12	83% (10)			8% (1)
16	17	35% (6)		76% (13)	6% (1)
17	2			50% (1)	50% (1)
18	16	50% (8)		63% (10)	6% (1)
19	20	40% (8)		65% (13)	10% (2)
20	14	50% (7)			29% (4)

Table 35: correct identifications in BioFace II.2.1

[27] The system did not run on all the test days (see Table 33: correct identifications in BioFace II.2.2).

Person ID	Flexitime events	System A	System B	System C	System D
1	9	33% (3)		11% (1)	
2	23	22% (5)			
3	27	33% (9)		37% (10)	
4	7	14% (1)		29% (2)	
5	13	15% (2)		15% (2)	8% (1)
6	22	9% (2)		32% (7)	
7	20	35% (7)		45% (9)	
8	17	24% (4)		24% (4)	
9	7	57% (4)		29% (2)	
10	15	33% (5)			
11	26	31% (8)		8% (2)	
12	20	35% (7)		45% (9)	
13	17	35% (6)		6% (1)	
14	7	29% (2)		14% (1)	
15	15	27% (4)		13% (2)	7% (1)
16	13	38% (5)		23% (3)	
17	18	33% (6)		6% (1)	
18	12	33% (4)	8% (1)	25% (3)	
19	24	8% (2)		67% (16)	
20	23	22% (5)		35% (8)	

Table 36: correct identifications in BioFace II.2.2

Day	System A	System B	System C	System D
1 Nov. 2002[28]				
4 Nov. 2002	4	1	7	2
5 Nov. 2002	10		2	3
6 Nov. 2002	9		5	4
7 Nov. 2002	2	1	5	6
8 Nov. 2002	5		4	3
11 Nov. 2002	5		4	2
12 Nov. 2002	7		4	1
13 Nov. 2002	3		4	1
14 Nov. 2002	3	1	5	2
15 Nov. 2002	1	1	2	3
18 Nov. 2002	2		6	4
19 Nov. 2002	1		6	1
20 Nov. 2002	4		3	2
21 Nov. 2002	4		4	1
22 Nov. 2002	4		3	1
25 Nov. 2002	2		5	
26 Nov. 2002	4	1	6	
27 Nov. 2002	4		5	1
28 Nov. 2002	5		7	
29 Nov. 2002	2		6	
2 Dec. 2002	1		8	2
3 Dec. 2002	4	1	9	
4 Dec. 2002	2		5	
5 Dec. 2002	1		10	
6 Dec. 2002	7		6	1
9 Dec. 2002	7	1	6	
10 Dec. 2002	9		2	
11 Dec. 2002	3	1	5	2
12 Dec. 2002	12			

Table 37: correct identifications in BioFace II.2.1

[28] Due to failure of the flexitime recording system, not all the systems were evaluated on 1 November 2002.

Day	System A	System B	System C	System D
3 Feb. 2003	9			
4 Feb. 2003	4		2	
5 Feb. 2003	5		2	
6 Feb. 2003	4		2	
7 Feb. 2003	2		3	
10 Feb. 2003	4		3	1
11 Feb. 2003	5		1	
12 Feb. 2003	3		5	1
13 Feb. 2003	5		3	
14 Feb. 2003	1		3	
17 Feb. 2003	2		3	
18 Feb. 2003			4	
19 Feb. 2003	2			
20 Feb. 2003	5		3	
21 Feb. 2003	3		3	
24 Feb. 2003	4		5	
25 Feb. 2003	5		4	
26 Feb. 2003	6		3	
27 Feb. 2003	2		3	
28 Feb. 2003	3		1	
4 Mar. 2003	2		1	
5 Mar. 2003	4		4	
6 Mar. 2003	3		1	
7 Mar. 2003	1		3	
10 Mar. 2003			6	
11 Mar. 2003	3	1	5	
12 Mar. 2003	1		3	
13 Mar. 2003	1		4	
14 Mar. 2003	2		3	

Table 38: correct identifications in BioFace II.2.2

6.4.3 Notes to the diagrams

To make them easier to read, the diagrams which follow are not uniformly scaled. If the same scale had been used on all of them they would have been virtually unusable. Since on the diagrams for System C (Figure 125: results for BioFace II.2.1 System C and Figure 129: results for BioFace II.2.2 System C) very low matching scores occur very frequently, these were excluded and the diagram was scaled so as to make the course of the curve over the remaining part of the matching score range easy to follow.

6.4.4 Diagrams

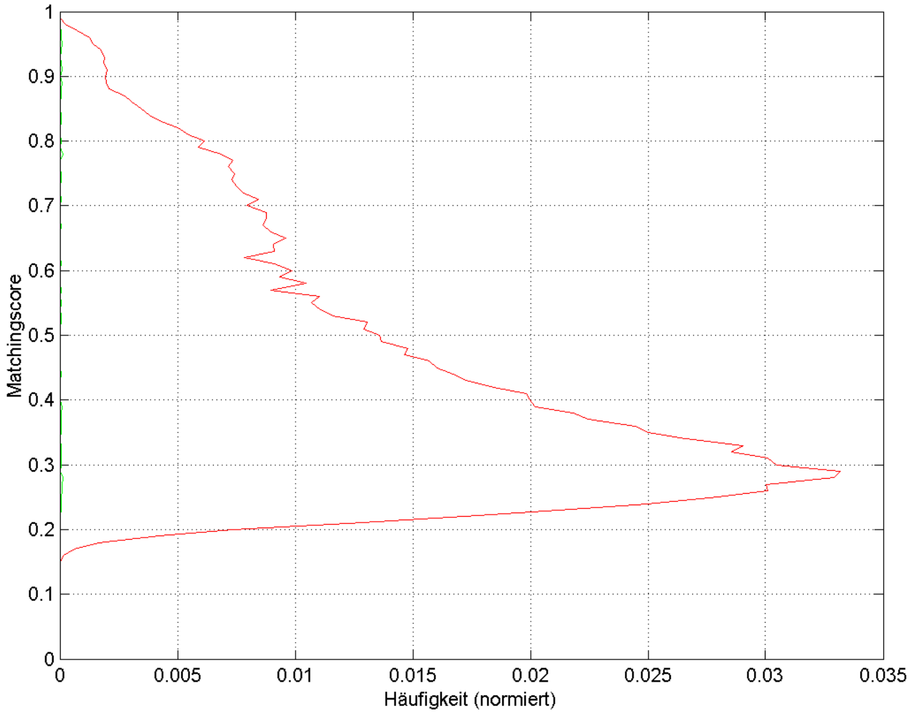


Figure 123: results for BioFace II.2.1 System A

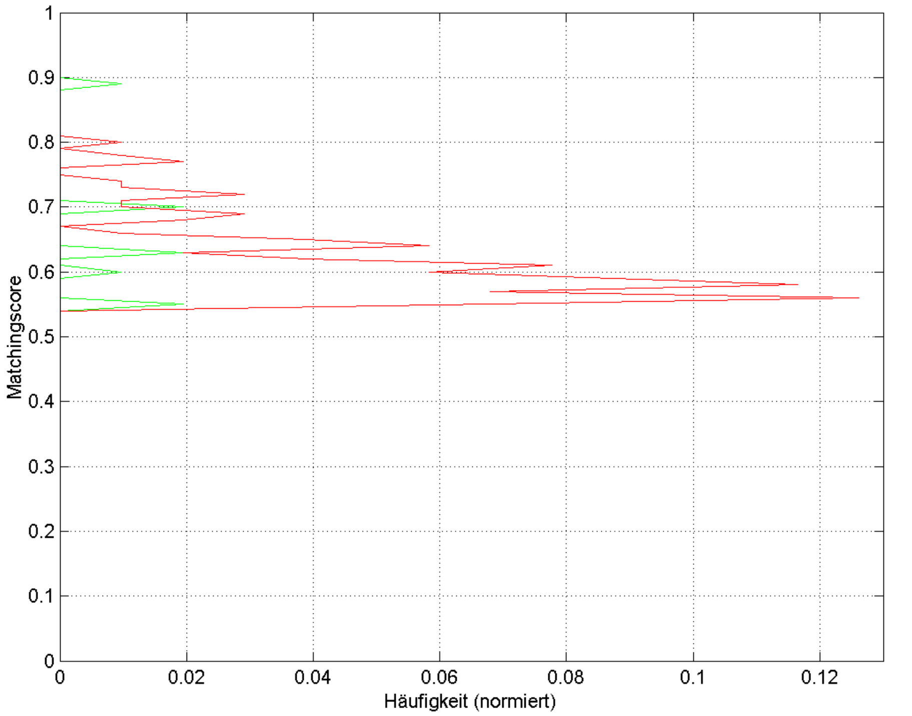


Figure 124: results for BioFace II.2.1 System B

Qualität/Matchingscore	Quality/matching score
Häufigkeit (normiert)	Frequency (normalised)

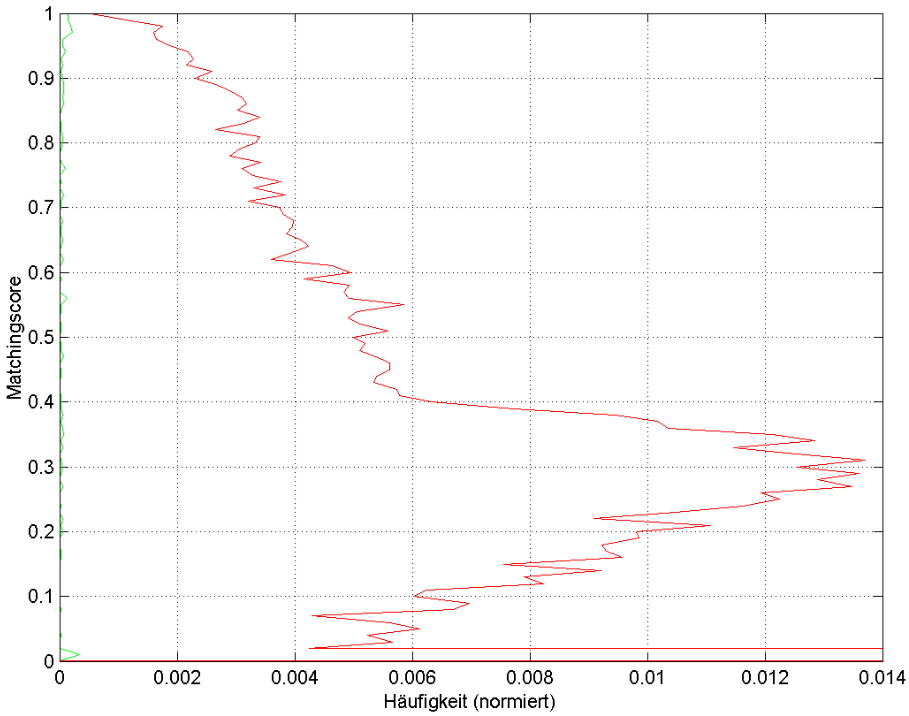


Figure 125: results for BioFace II.2.1 System C

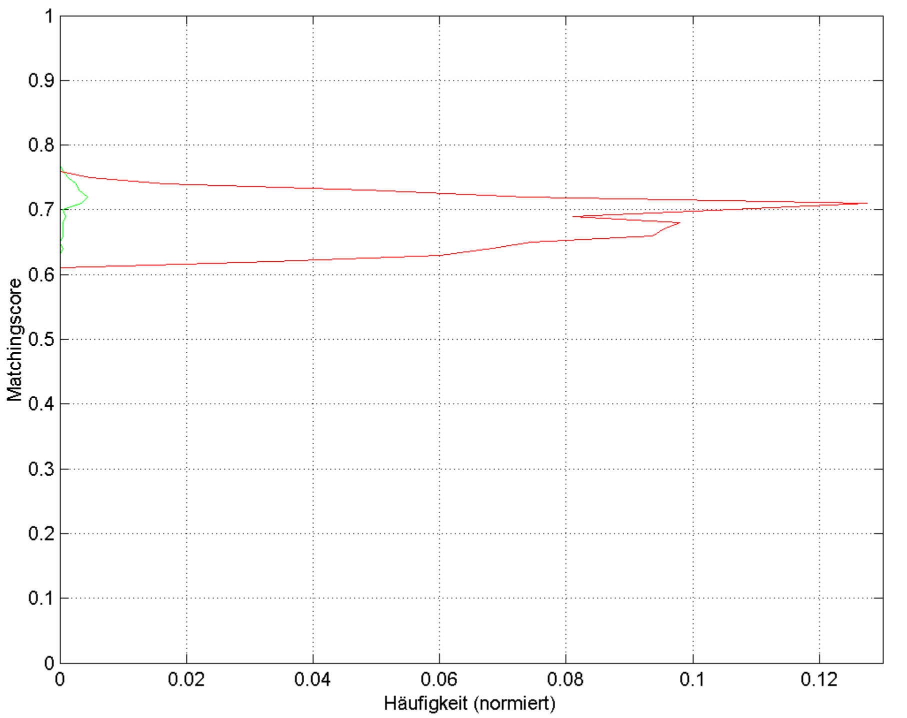


Figure 126: results for BioFace II.2.1 System D

Qualität/Matchingscore	Quality/matching score
Häufigkeit (normiert)	Frequency (normalised)

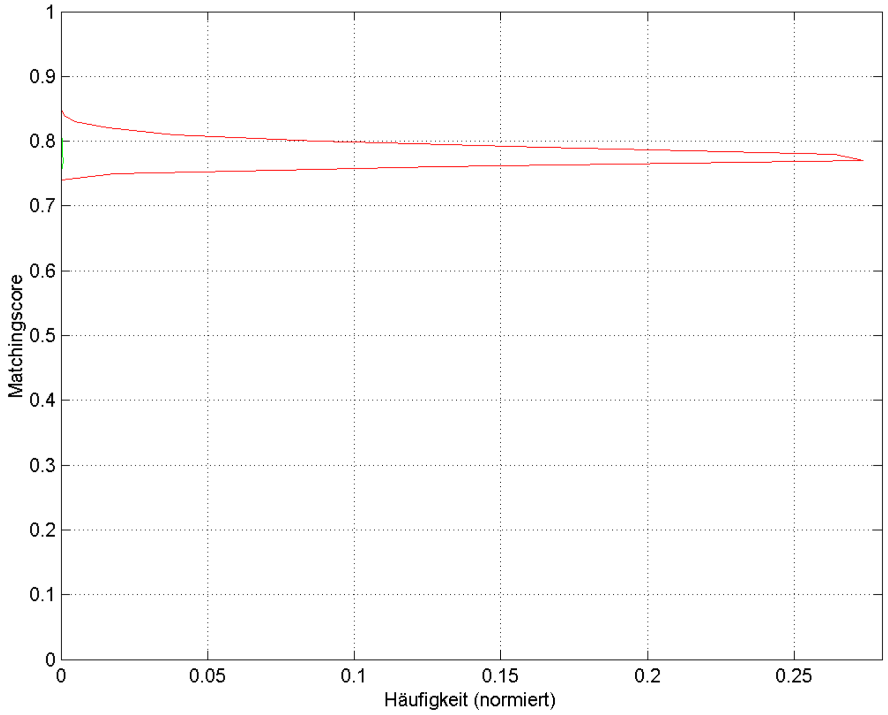


Figure 127: results for BioFace II.2.2 System A

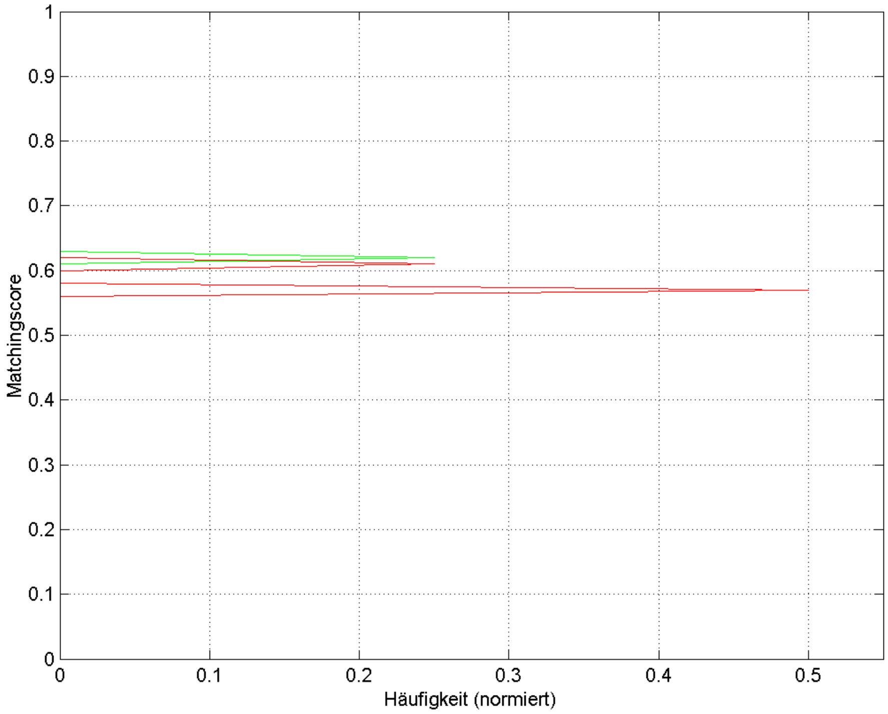


Figure 128: results for BioFace II.2.2 System B

Qualität/Matchingscore	Quality/matching score
Häufigkeit (normiert)	Frequency (normalised)

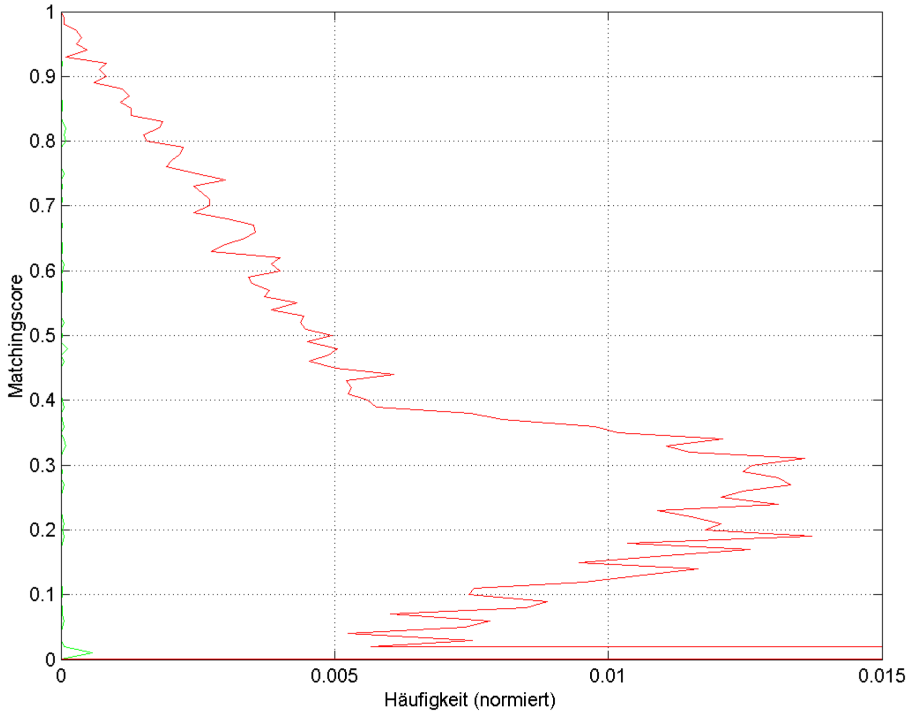


Figure 129: results for BioFace II.2.2 System C

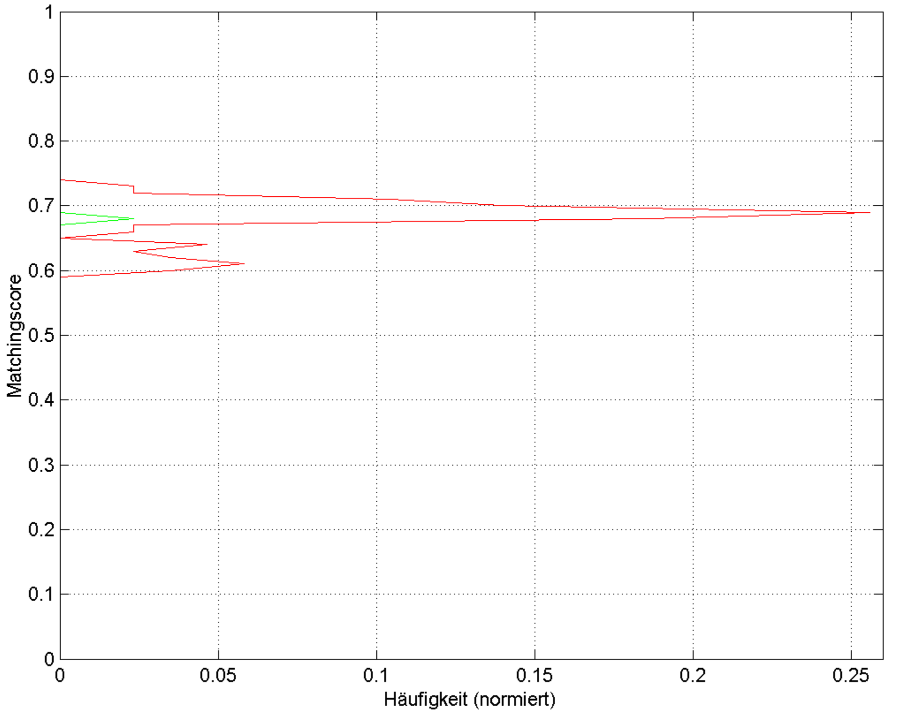


Figure 130: results for BioFace II.2.2 System D

Qualität/Matchingscore	Quality/matching score
Häufigkeit (normiert)	Frequency (normalised)

6.4.5 Interpretation of results

The recognition results obtained can be read directly in Table 34: false rejection rate (FRR) for possible recognitions on page 116. It is clear from the figures that systems A and C performed the best in both test phases. They are followed by systems D and B.

If one examines the results for Phase 1 more closely (Table 32: correct identifications in BioFace II.2.1 on page 115), however, one can discern a clear difference between systems A and C: with System C, the hits are spread more strongly across the results data record, whereas with System A, they occur only in the first three positions.

After a promising start, System D declined steadily over the test period (see Table 37: correct identifications in BioFace II.2.1 on page 118).

On moving from Phase 1 with live enrolment to Phase 2 with enrolment of digital images, a deterioration in performance can be discerned for all the systems (Table 33: correct identifications in BioFace II.2.2 on page 115), with System D experiencing the most dramatic decline. Once again, there is a more pronounced distribution of results over the five positions under System C than under System A.

System B cannot be included in this comparison, as that system was not available for extensive periods of the second test face.

7 Summary and looking ahead

7.1 Summary of the results

The results achieved in Phase 1 of the BioFace II project appear at first sight to contradict each other. In Test Group 1/noise-free comparison, it is clear that the average matching score of non-hits was not materially influenced by the size of the database (see 5.7.3.17). This stands in contrast with Test Group 4/hunt, in which the number of hits declined as the size of the test groups increased (see 5.7.4.16). If one examines these results more closely, the reason for this discrepancy becomes apparent: during Test Group 1 all the results were ascertained and considered using verification, whereas in Test Group 4 the first ten hits in an identification were used for the evaluation.

As already analysed in Test Group 1, the number of good and poor matching scores rises in proportion to its distribution over the value spectrum, i.e. the distribution remains the same. However, if one keeps the number and set of images for the same persons constant and in parallel to this allows the number of images of different persons to rise, as occurred in Test Group 4, the extra non-hits with high scores that occur gradually replace the hits in the identification result sets. This is the reason for the apparent decline in the number of correct hits.

Since during identification the low-scoring results – including non-hits – were excluded, the high-scoring non-hits appear to grow because we are now looking only at the absolute growth without comparison.

During the system test, systems A and C proved the better systems. After a promising start, System D declined steadily over the test period (see Table 37: correct identifications in BioFace II.2.1 on page 118). In the first part of the test, System B did not achieve a recognition rate of any consequence and for various reasons could not be operated for extensive parts of the second phase.

7.2 Looking ahead

The extremely interesting results from BioFace II show that it is necessary to analyse precisely where the causes for changes in the behaviour and recognition performance of biometric facial recognition systems lie. Experience gained, amongst other things, from talking to the vendors suggests that, specifically, the image material has a large influence on the test results. Thus it could be that a European system was predominantly tested and developed with European facial images. The use of image sets containing other population groups, for example, from the African or Asian regions, could possibly produce a different picture.

Again, the quality of the image itself can have a material influence on the recognition performance of facial recognition systems. However, it is necessary for corresponding scoring to be able to access measurably scaled, noisy image material, which generally can only be obtained by creating it accordingly. Neither for the use nor for the development of corresponding systems is it helpful to make any statements about the influence of noise factors without also stating from what point these have a material influence. The subject of project BioFace III will therefore be to examine the measurable influence of noise factors.

Annex A Additional result s for Test Group 4

The tables below present Key Figures 1 and 2 (see section 5.7.4.2 Scoring according to key figures on page 41) for Test Group 4.

Person ID	Number of reference images	Algorithm 1	Algorithm 2	Algorithm 3
1	5	0.60	0.00	0.60
2	10	0.80	0.60	0.60
3	10	0.60	0.30	0.60
4	6	0.00	0.00	0.00
5	4	1.00	0.00	1.00
6	7	1.00	0.00	0.86
7	6	0.50	0.17	0.83
8	3	0.67	0.00	0.00
9	6	0.00	0.00	0.50
10	2	0.00	0.00	0.50
11	5	0.20	0.40	0.20
12	1	0.00	0.00	0.00
13	8	0.88	0.38	0.62
14	7	0.00	0.14	0.14
15	10	0.80	0.00	0.80
16	6	0.33	0.17	0.67
17	3	0.00	0.00	0.33
18	2	0.00	0.00	0.00
19	6	0.50	0.00	0.67
20	4	0.00	0.00	0.50
21	5	0.00	0.00	0.40
22	3	1.00	0.33	1.00
23	6	0.67	0.00	0.83
24	3	1.00	0.00	0.67
25	3	0.00	0.00	0.00
26	2	1.00	0.00	0.50
27	1	0.00	0.00	0.00
28	2	0.50	0.00	1.00
29	1	1.00	0.00	0.00
30	2	0.50	0.50	0.00
31	3	1.00	0.33	0.67

32	1	0.00	0.00	0.00
33	2	1.00	0.00	1.00
34	3	0.33	0.33	0.33
35	1	0.00	0.00	0.00
36	4	0.50	0.00	0.25
37	3	0.33	0.00	0.33
38	6	1.00	0.00	0.83
39	8	0.38	0.00	0.62
40	1	1.00	0.00	0.00
41	1	0.00	0.00	0.00
42	1	1.00	1.00	1.00
43	1	1.00	1.00	1.00
44	1	1.00	0.00	1.00
45	1	1.00	0.00	1.00
46	2	0.50	0.00	1.00
47	2	0.00	0.00	0.00
48	2	0.00	0.00	1.00
49	1	1.00	1.00	1.00
50	2	0.00	0.00	0.00
51	4	1.00	0.00	1.00
52	2	0.50	0.00	0.50
53	1	0.00	0.00	0.00
54	0	0.00	0.00	0.00
55	4	0.75	0.00	0.25
56	1	0.00	0.00	0.00
57	1	0.00	0.00	0.00
58	2	0.00	0.50	1.00
59	10	0.00	0.00	0.30
60	1	0.00	0.00	0.00
61	2	0.00	0.00	0.50
62	3	0.67	0.33	0.67
63	3	0.67	0.33	0.33
64	2	0.50	0.00	0.50
65	6	0.50	0.00	0.33
66	5	0.60	0.60	0.80
67	2	0.00	0.00	0.00
68	1	1.00	0.00	0.00

69	1	1.00	0.00	1.00
70	2	0.50	0.50	0.50
71	1	1.00	0.00	1.00
72	1	1.00	0.00	1.00
73	1	1.00	0.00	0.00
74	1	0.00	1.00	1.00
75	1	0.00	0.00	0.00
76	1	1.00	0.00	1.00
77	2	1.00	0.00	0.50
78	1	1.00	0.00	0.00
79	1	0.00	0.00	1.00
80	2	0.00	0.50	1.00
81	2	0.50	0.00	0.50
82	1	1.00	0.00	1.00
83	1	0.00	0.00	1.00
84	1	1.00	0.00	1.00
85	2	0.50	0.00	1.00
86	1	1.00	0.00	0.00
87	1	1.00	1.00	1.00
88	1	1.00	0.00	0.00
89	1	1.00	0.00	0.00
90	1	0.00	0.00	0.00
91	1	1.00	0.00	1.00
92	1	0.00	0.00	1.00
93	1	1.00	1.00	1.00
94	1	1.00	0.00	1.00
95	1	1.00	0.00	0.00
96	1	0.00	0.00	0.00
97	1	0.00	1.00	1.00
98	1	0.00	0.00	0.00
99	1	1.00	0.00	1.00
100	1	1.00	0.00	1.00
101	1	1.00	0.00	1.00
102	1	1.00	0.00	1.00
103	2	1.00	0.00	1.00
104	1	1.00	0.00	1.00
105	1	1.00	0.00	1.00

106	1	1.00	0.00	1.00
107	1	0.00	0.00	1.00
108	1	0.00	0.00	0.00
109	1	0.00	0.00	0.00
110	2	1.00	0.00	0.50
111	2	1.00	0.00	0.50
112	1	1.00	0.00	1.00
113	1	1.00	1.00	1.00
114	1	1.00	1.00	1.00
115	5	1.00	0.20	0.80
116	6	0.50	0.00	0.00

Table 39: Key Figure 1 by person ID for Test 4-1

Person ID	Number of reference images	Algorithm 1	Algorithm 2	Algorithm 3
1	5	0.40	0.00	0.60
2	10	0.70	0.60	0.60
3	10	0.60	0.30	0.50
4	6	0.00	0.00	0.00
5	4	1.00	0.00	1.00
6	7	0.71	0.00	0.71
7	6	0.50	0.17	0.67
8	3	0.33	0.00	0.33
9	6	0.00	0.00	0.50
10	2	0.00	0.00	0.50
11	5	0.20	0.40	0.20
12	1	0.00	0.00	0.00
13	8	0.75	0.38	0.50
14	7	0.00	0.14	0.00
15	10	0.70	0.00	0.80
16	6	0.33	0.17	0.33
17	3	0.00	0.00	0.33
18	2	0.00	0.00	0.00
19	6	0.33	0.00	0.50
20	4	0.00	0.00	0.25
21	5	0.00	0.00	0.60
22	3	0.67	0.33	0.67

23	6	0.50	0.00	0.50
24	3	1.00	0.00	0.67
25	3	0.00	0.00	0.00
26	2	1.00	0.00	0.50
27	1	0.00	0.00	0.00
28	2	0.50	0.00	1.00
29	1	0.00	0.00	0.00
30	2	0.50	0.50	0.00
31	3	1.00	0.33	0.67
32	1	0.00	0.00	0.00
33	2	1.00	0.00	0.50
34	3	0.00	0.33	0.00
35	1	0.00	0.00	0.00
36	4	0.25	0.00	0.25
37	3	0.00	0.00	0.00
38	6	0.67	0.00	0.50
39	8	0.25	0.00	0.50
40	1	1.00	0.00	0.00
41	1	0.00	0.00	0.00
42	1	1.00	1.00	1.00
43	1	1.00	1.00	1.00
44	1	1.00	0.00	1.00
45	1	0.00	0.00	1.00
46	2	0.50	0.00	1.00
47	2	0.00	0.00	0.00
48	2	0.00	0.00	1.00
49	1	1.00	1.00	1.00
50	2	0.00	0.00	0.00
51	4	1.00	0.00	1.00
52	2	0.00	0.00	0.50
53	1	0.00	0.00	0.00
54	0	0.00	0.00	0.00
55	4	0.75	0.00	0.00
56	1	0.00	0.00	0.00
57	1	0.00	0.00	0.00
58	2	0.00	0.50	0.50
59	10	0.00	0.00	0.10

60	1	0.00	0.00	0.00
61	2	0.00	0.00	0.50
62	3	0.67	0.33	0.67
63	3	0.33	0.33	0.33
64	2	0.50	0.00	0.50
65	6	0.17	0.00	0.33
66	5	0.60	0.40	0.40
67	2	0.00	0.00	0.00
68	1	1.00	0.00	0.00
69	1	1.00	0.00	1.00
70	2	0.50	0.50	0.50
71	1	1.00	0.00	1.00
72	1	0.00	0.00	1.00
73	1	1.00	0.00	1.00
74	1	0.00	1.00	0.00
75	1	0.00	0.00	0.00
76	1	1.00	0.00	1.00
77	2	0.50	0.00	0.00
78	1	1.00	0.00	0.00
79	1	0.00	0.00	1.00
80	2	0.00	0.50	1.00
81	2	0.50	0.00	0.50
82	1	1.00	0.00	1.00
83	1	0.00	0.00	1.00
84	1	1.00	0.00	1.00
85	2	0.50	0.00	1.00
86	1	0.00	0.00	0.00
87	1	1.00	1.00	1.00
88	1	1.00	0.00	0.00
89	1	1.00	0.00	0.00
90	1	0.00	0.00	0.00
91	1	1.00	0.00	1.00
92	1	0.00	0.00	1.00
93	1	1.00	1.00	1.00
94	1	1.00	0.00	1.00
95	1	0.00	0.00	0.00
96	1	0.00	0.00	0.00

97	1	0.00	1.00	1.00
98	1	0.00	0.00	0.00
99	1	1.00	0.00	1.00
100	1	1.00	0.00	0.00
101	1	1.00	0.00	1.00
102	1	1.00	0.00	0.00
103	2	1.00	0.00	1.00
104	1	1.00	0.00	1.00
105	1	1.00	0.00	1.00
106	1	0.00	0.00	1.00
107	1	0.00	0.00	0.00
108	1	0.00	0.00	0.00
109	1	0.00	0.00	0.00
110	2	1.00	0.00	0.50
111	2	1.00	0.00	0.50
112	1	1.00	0.00	1.00
113	1	1.00	1.00	1.00
114	1	1.00	1.00	1.00
115	5	1.00	0.20	0.80
116	6	0.50	0.00	0.00

Table 40: Key Figure 1 by person ID for Test 4-2

Person ID	Number of reference images	Algorithm 1	Algorithm 2	Algorithm 3
1	5	0.20	0.00	0.40
2	10	0.70	0.60	0.60
3	10	0.60	0.30	0.50
4	6	0.00	0.00	0.00
5	4	0.75	0.00	0.75
6	7	0.71	0.00	0.71
7	6	0.50	0.17	0.50
8	3	0.33	0.00	0.00
9	6	0.00	0.00	0.50
10	2	0.00	0.00	0.50
11	5	0.20	0.20	0.20
12	1	0.00	0.00	0.00
13	8	0.75	0.38	0.38

14	7	0.00	0.14	0.00
15	10	0.60	0.00	0.50
16	6	0.33	0.17	0.33
17	3	0.00	0.00	0.00
18	2	0.00	0.00	0.00
19	6	0.17	0.00	0.33
20	4	0.00	0.00	0.25
21	5	0.00	0.00	0.40
22	3	0.67	0.33	0.33
23	6	0.33	0.00	0.33
24	3	1.00	0.00	0.00
25	3	0.00	0.00	0.00
26	2	1.00	0.00	0.50
27	1	0.00	0.00	0.00
28	2	0.50	0.00	0.50
29	1	0.00	0.00	0.00
30	2	0.50	0.50	0.00
31	3	1.00	0.33	0.67
32	1	0.00	0.00	0.00
33	2	0.50	0.00	0.50
34	3	0.00	0.33	0.00
35	1	0.00	0.00	0.00
36	4	0.25	0.00	0.00
37	3	0.00	0.00	0.00
38	6	0.50	0.00	0.33
39	8	0.25	0.00	0.38
40	1	1.00	0.00	0.00
41	1	0.00	0.00	0.00
42	1	1.00	1.00	1.00
43	1	1.00	1.00	0.00
44	1	0.00	0.00	0.00
45	1	0.00	0.00	1.00
46	2	0.00	0.00	0.50
47	2	0.00	0.00	0.00
48	2	0.00	0.00	1.00
49	1	1.00	1.00	1.00
50	2	0.00	0.00	0.00

51	4	1.00	0.00	0.75
52	2	0.00	0.00	0.00
53	1	0.00	0.00	0.00
54	0	0.00	0.00	0.00
55	4	0.75	0.00	0.00
56	1	0.00	0.00	0.00
57	1	0.00	0.00	0.00
58	2	0.00	0.50	0.50
59	10	0.00	0.00	0.10
60	1	0.00	0.00	0.00
61	2	0.00	0.00	0.50
62	3	0.67	0.00	0.33
63	3	0.33	0.00	0.33
64	2	0.50	0.00	0.00
65	6	0.17	0.00	0.17
66	5	0.40	0.40	0.40
67	2	0.00	0.00	0.00
68	1	1.00	0.00	0.00
69	1	1.00	0.00	1.00
70	2	0.00	0.50	0.00
71	1	0.00	0.00	1.00
72	1	0.00	0.00	1.00
73	1	1.00	0.00	0.00
74	1	0.00	1.00	0.00
75	1	0.00	0.00	0.00
76	1	1.00	0.00	0.00
77	2	0.50	0.00	0.00
78	1	0.00	0.00	0.00
79	1	0.00	0.00	1.00
80	2	0.00	0.50	1.00
81	2	0.50	0.00	0.00
82	1	1.00	0.00	1.00
83	1	0.00	0.00	1.00
84	1	1.00	0.00	0.00
85	2	0.50	0.00	1.00
86	1	0.00	0.00	0.00
87	1	1.00	1.00	1.00

88	1	1.00	0.00	0.00
89	1	0.00	0.00	0.00
90	1	0.00	0.00	0.00
91	1	1.00	0.00	1.00
92	1	0.00	0.00	0.00
93	1	1.00	1.00	1.00
94	1	1.00	0.00	1.00
95	1	0.00	0.00	0.00
96	1	0.00	0.00	0.00
97	1	0.00	1.00	1.00
98	1	0.00	0.00	0.00
99	1	1.00	0.00	1.00
100	1	0.00	0.00	0.00
101	1	0.00	0.00	0.00
102	1	0.00	0.00	0.00
103	2	1.00	0.00	1.00
104	1	1.00	0.00	1.00
105	1	1.00	0.00	0.00
106	1	0.00	0.00	0.00
107	1	0.00	0.00	0.00
108	1	0.00	0.00	0.00
109	1	0.00	0.00	0.00
110	2	1.00	0.00	0.00
111	2	1.00	0.00	0.50
112	1	1.00	0.00	1.00
113	1	1.00	1.00	1.00
114	1	1.00	1.00	1.00
115	5	0.60	0.20	0.80
116	6	0.33	0.00	0.00

Table 41: Key Figure 1 by person ID for Test 4-3

Person ID	Number of reference images	Algorithm 1	Algorithm 2	Algorithm 3
1	5	0.00	0.00	0.00
2	10	0.50	0.50	0.40
3	10	0.40	0.30	0.40
4	6	0.00	0.00	0.00

5	4	0.50	0.00	0.50
6	7	0.71	0.00	0.57
7	6	0.50	0.17	0.50
8	3	0.33	0.00	0.00
9	6	0.00	0.00	0.00
10	2	0.00	0.00	0.00
11	5	0.20	0.20	0.00
12	1	0.00	0.00	0.00
13	8	0.75	0.25	0.00
14	7	0.00	0.14	0.00
15	10	0.40	0.00	0.30
16	6	0.33	0.17	0.17
17	3	0.00	0.00	0.00
18	2	0.00	0.00	0.00
19	6	0.00	0.00	0.17
20	4	0.00	0.00	0.00
21	5	0.00	0.00	0.40
22	3	0.67	0.33	0.00
23	6	0.33	0.00	0.00
24	3	1.00	0.00	0.33
25	3	0.00	0.00	0.00
26	2	1.00	0.00	0.50
27	1	0.00	0.00	0.00
28	2	0.50	0.00	0.00
29	1	0.00	0.00	0.00
30	2	0.00	0.00	0.00
31	3	1.00	0.00	0.67
32	1	0.00	0.00	0.00
33	2	0.00	0.00	0.00
34	3	0.00	0.33	0.00
35	1	0.00	0.00	0.00
36	4	0.00	0.00	0.00
37	3	0.00	0.00	0.00
38	6	0.33	0.00	0.17
39	8	0.25	0.00	0.38
40	1	1.00	0.00	0.00
41	1	0.00	0.00	0.00

42	1	1.00	1.00	1.00
43	1	1.00	1.00	0.00
44	1	0.00	0.00	0.00
45	1	0.00	0.00	0.00
46	2	0.00	0.00	0.00
47	2	0.00	0.00	0.00
48	2	0.00	0.00	1.00
49	1	1.00	1.00	1.00
50	2	0.00	0.00	0.00
51	4	0.75	0.00	0.50
52	2	0.00	0.00	0.00
53	1	0.00	0.00	0.00
54	0	0.00	0.00	0.00
55	4	0.25	0.00	0.00
56	1	0.00	0.00	0.00
57	1	0.00	0.00	0.00
58	2	0.00	0.50	0.50
59	10	0.00	0.00	0.00
60	1	0.00	0.00	0.00
61	2	0.00	0.00	0.00
62	3	0.67	0.00	0.00
63	3	0.33	0.00	0.33
64	2	0.50	0.00	0.00
65	6	0.17	0.00	0.17
66	5	0.40	0.40	0.40
67	2	0.00	0.00	0.00
68	1	1.00	0.00	0.00
69	1	1.00	0.00	0.00
70	2	0.00	0.00	0.00
71	1	0.00	0.00	1.00
72	1	0.00	0.00	1.00
73	1	1.00	0.00	0.00
74	1	0.00	1.00	0.00
75	1	0.00	0.00	0.00
76	1	0.00	0.00	0.00
77	2	0.00	0.00	0.00
78	1	0.00	0.00	0.00

79	1	0.00	0.00	1.00
80	2	0.00	0.50	1.00
81	2	0.00	0.00	0.00
82	1	1.00	0.00	1.00
83	1	0.00	0.00	0.00
84	1	0.00	0.00	0.00
85	2	0.50	0.00	0.50
86	1	0.00	0.00	0.00
87	1	1.00	1.00	1.00
88	1	0.00	0.00	0.00
89	1	0.00	0.00	0.00
90	1	0.00	0.00	0.00
91	1	1.00	0.00	0.00
92	1	0.00	0.00	0.00
93	1	1.00	1.00	1.00
94	1	1.00	0.00	1.00
95	1	0.00	0.00	0.00
96	1	0.00	0.00	0.00
97	1	0.00	1.00	1.00
98	1	0.00	0.00	0.00
99	1	1.00	0.00	1.00
100	1	0.00	0.00	0.00
101	1	0.00	0.00	0.00
102	1	0.00	0.00	0.00
103	2	1.00	0.00	1.00
104	1	1.00	0.00	1.00
105	1	1.00	0.00	0.00
106	1	0.00	0.00	0.00
107	1	0.00	0.00	0.00
108	1	0.00	0.00	0.00
109	1	0.00	0.00	0.00
110	2	0.50	0.00	0.00
111	2	0.50	0.00	0.50
112	1	1.00	0.00	1.00
113	1	1.00	1.00	0.00
114	1	1.00	1.00	1.00
115	5	0.40	0.20	0.00

116	6	0.00	0.00	0.00

Table 42: Key Figure 1 by person ID for Test 4-4

Person ID	Number of reference images	Algorithm 1	Algorithm 2	Algorithm 3
1	5	15.60	0.00	14.40
2	10	40.80	26.40	25.20
3	10	27.00	8.10	25.20
4	6	0.00	0.00	0.00
5	4	33.00	0.00	34.00
6	7	48.00	0.00	38.57
7	6	13.50	1.67	28.33
8	3	10.67	0.00	0.00
9	6	0.00	0.00	12.50
10	2	0.00	0.00	4.50
11	5	2.00	7.20	2.00
12	1	0.00	0.00	0.00
13	8	41.12	10.12	22.50
14	7	0.00	1.43	0.57
15	10	41.60	0.00	40.00
16	6	6.33	1.67	20.67
17	3	0.00	0.00	2.67
18	2	0.00	0.00	0.00
19	6	12.50	0.00	18.67
20	4	0.00	0.00	7.50
21	5	0.00	0.00	6.80
22	3	25.00	3.33	27.00
23	6	22.67	0.00	27.50
24	3	27.00	0.00	12.00
25	3	0.00	0.00	0.00
26	2	19.00	0.00	5.00
27	1	0.00	0.00	0.00
28	2	5.00	0.00	18.00
29	1	2.00	0.00	0.00
30	2	5.00	5.00	0.00
31	3	27.00	3.33	12.67
32	1	0.00	0.00	0.00

33	2	19.00	0.00	18.00
34	3	3.00	3.00	3.00
35	1	0.00	0.00	0.00
36	4	8.50	0.00	2.00
37	3	2.00	0.00	1.67
38	6	39.00	0.00	29.17
39	8	9.38	0.00	23.12
40	1	10.00	0.00	0.00
41	1	0.00	0.00	0.00
42	1	10.00	10.00	10.00
43	1	10.00	10.00	10.00
44	1	9.00	0.00	10.00
45	1	5.00	0.00	9.00
46	2	4.00	0.00	17.00
47	2	0.00	0.00	0.00
48	2	0.00	0.00	19.00
49	1	10.00	10.00	10.00
50	2	0.00	0.00	0.00
51	4	34.00	0.00	33.00
52	2	0.50	0.00	1.50
53	1	0.00	0.00	0.00
54	0	0.00	0.00	0.00
55	4	20.25	0.00	2.00
56	1	0.00	0.00	0.00
57	1	0.00	0.00	0.00
58	2	0.00	5.00	17.00
59	10	0.00	0.00	3.90
60	1	0.00	0.00	0.00
61	2	0.00	0.00	4.00
62	3	12.67	3.00	11.33
63	3	9.33	2.67	3.33
64	2	5.00	0.00	5.00
65	6	10.50	0.00	5.67
66	5	15.60	12.00	19.20
67	2	0.00	0.00	0.00
68	1	10.00	0.00	0.00
69	1	10.00	0.00	10.00

70	2	4.00	5.00	4.50
71	1	10.00	0.00	10.00
72	1	2.00	0.00	10.00
73	1	10.00	0.00	0.00
74	1	0.00	10.00	10.00
75	1	0.00	0.00	0.00
76	1	9.00	0.00	7.00
77	2	15.00	0.00	1.00
78	1	9.00	0.00	0.00
79	1	0.00	0.00	10.00
80	2	0.00	5.00	19.00
81	2	4.50	0.00	4.00
82	1	10.00	0.00	10.00
83	1	0.00	0.00	10.00
84	1	10.00	0.00	8.00
85	2	5.00	0.00	19.00
86	1	7.00	0.00	0.00
87	1	10.00	10.00	10.00
88	1	10.00	0.00	0.00
89	1	9.00	0.00	0.00
90	1	0.00	0.00	0.00
91	1	10.00	0.00	10.00
92	1	0.00	0.00	10.00
93	1	10.00	10.00	10.00
94	1	10.00	0.00	10.00
95	1	4.00	0.00	0.00
96	1	0.00	0.00	0.00
97	1	0.00	10.00	10.00
98	1	0.00	0.00	0.00
99	1	10.00	0.00	10.00
100	1	10.00	0.00	8.00
101	1	10.00	0.00	8.00
102	1	9.00	0.00	10.00
103	2	19.00	0.00	19.00
104	1	10.00	0.00	10.00
105	1	10.00	0.00	9.00
106	1	2.00	0.00	10.00

107	1	0.00	0.00	1.00
108	1	0.00	0.00	0.00
109	1	0.00	0.00	0.00
110	2	18.00	0.00	4.50
111	2	18.00	0.00	5.00
112	1	10.00	0.00	10.00
113	1	10.00	10.00	10.00
114	1	10.00	10.00	10.00
115	5	38.00	2.00	25.60
116	6	13.00	0.00	0.00

Table 43: Key Figure 2 by person ID for Test 4-1

Person ID	Number of reference images	Algorithm 1	Algorithm 2	Algorithm 3
1	5	6.80	0.00	9.60
2	10	31.50	26.40	24.60
3	10	27.00	8.10	20.00
4	6	0.00	0.00	0.00
5	4	31.00	0.00	30.00
6	7	28.57	0.00	28.57
7	6	13.50	1.67	21.33
8	3	3.33	0.00	0.33
9	6	0.00	0.00	12.50
10	2	0.00	0.00	3.50
11	5	2.00	4.80	1.80
12	1	0.00	0.00	0.00
13	8	33.75	9.38	15.00
14	7	0.00	1.43	0.00
15	10	32.90	0.00	40.80
16	6	6.33	1.67	5.67
17	3	0.00	0.00	1.67
18	2	0.00	0.00	0.00
19	6	4.33	0.00	9.50
20	4	0.00	0.00	2.50
21	5	0.00	0.00	16.20
22	3	12.67	3.33	12.67
23	6	12.50	0.00	9.00

24	3	27.00	0.00	10.00
25	3	0.00	0.00	0.00
26	2	19.00	0.00	5.00
27	1	0.00	0.00	0.00
28	2	5.00	0.00	15.00
29	1	0.00	0.00	0.00
30	2	4.50	5.00	0.00
31	3	27.00	3.00	12.67
32	1	0.00	0.00	0.00
33	2	16.00	0.00	5.00
34	3	0.00	3.00	0.00
35	1	0.00	0.00	0.00
36	4	2.50	0.00	1.25
37	3	0.00	0.00	0.00
38	6	22.00	0.00	10.50
39	8	4.75	0.00	15.00
40	1	10.00	0.00	0.00
41	1	0.00	0.00	0.00
42	1	10.00	10.00	10.00
43	1	10.00	10.00	6.00
44	1	2.00	0.00	9.00
45	1	0.00	0.00	9.00
46	2	2.50	0.00	11.00
47	2	0.00	0.00	0.00
48	2	0.00	0.00	19.00
49	1	10.00	10.00	10.00
50	2	0.00	0.00	0.00
51	4	34.00	0.00	32.00
52	2	0.00	0.00	0.50
53	1	0.00	0.00	0.00
54	0	0.00	0.00	0.00
55	4	20.25	0.00	0.00
56	1	0.00	0.00	0.00
57	1	0.00	0.00	0.00
58	2	0.00	5.00	5.00
59	10	0.00	0.00	0.80
60	1	0.00	0.00	0.00

61	2	0.00	0.00	4.00
62	3	12.67	1.67	12.00
63	3	3.33	2.00	3.33
64	2	5.00	0.00	4.00
65	6	1.67	0.00	4.67
66	5	13.80	7.60	7.60
67	2	0.00	0.00	0.00
68	1	10.00	0.00	0.00
69	1	10.00	0.00	10.00
70	2	1.00	5.00	2.00
71	1	7.00	0.00	10.00
72	1	0.00	0.00	10.00
73	1	10.00	0.00	5.00
74	1	0.00	10.00	0.00
75	1	0.00	0.00	0.00
76	1	8.00	0.00	6.00
77	2	3.00	0.00	0.00
78	1	3.00	0.00	0.00
79	1	0.00	0.00	10.00
80	2	0.00	5.00	19.00
81	2	3.50	0.00	2.50
82	1	10.00	0.00	10.00
83	1	0.00	0.00	9.00
84	1	10.00	0.00	4.00
85	2	5.00	0.00	19.00
86	1	0.00	0.00	0.00
87	1	10.00	10.00	10.00
88	1	8.00	0.00	0.00
89	1	6.00	0.00	0.00
90	1	0.00	0.00	0.00
91	1	10.00	0.00	9.00
92	1	0.00	0.00	5.00
93	1	10.00	10.00	10.00
94	1	10.00	0.00	10.00
95	1	0.00	0.00	0.00
96	1	0.00	0.00	0.00
97	1	0.00	10.00	10.00

98	1	0.00	0.00	0.00
99	1	10.00	0.00	10.00
100	1	6.00	0.00	0.00
101	1	7.00	0.00	6.00
102	1	5.00	0.00	0.00
103	2	19.00	0.00	19.00
104	1	10.00	0.00	10.00
105	1	10.00	0.00	8.00
106	1	0.00	0.00	6.00
107	1	0.00	0.00	0.00
108	1	0.00	0.00	0.00
109	1	0.00	0.00	0.00
110	2	17.00	0.00	1.00
111	2	17.00	0.00	5.00
112	1	10.00	0.00	10.00
113	1	9.00	10.00	10.00
114	1	10.00	10.00	10.00
115	5	33.00	2.00	24.80
116	6	11.00	0.00	0.00

Table 44: Key Figure 2 by person ID for Test 4-2

Person ID	Number of reference images	Algorithm 1	Algorithm 2	Algorithm 3
1	5	1.00	0.00	4.00
2	10	31.50	26.40	23.40
3	10	25.80	8.10	19.50
4	6	0.00	0.00	0.00
5	4	18.75	0.00	16.50
6	7	28.57	0.00	27.86
7	6	13.50	1.67	13.50
8	3	3.33	0.00	0.00
9	6	0.00	0.00	6.50
10	2	0.00	0.00	4.00
11	5	2.00	2.00	1.80
12	1	0.00	0.00	0.00
13	8	32.25	9.00	8.62
14	7	0.00	1.43	0.00

15	10	26.40	0.00	18.00
16	6	6.33	1.67	4.33
17	3	0.00	0.00	0.00
18	2	0.00	0.00	0.00
19	6	1.00	0.00	3.33
20	4	0.00	0.00	2.25
21	5	0.00	0.00	7.60
22	3	12.67	3.33	1.00
23	6	6.33	0.00	4.00
24	3	27.00	0.00	0.00
25	3	0.00	0.00	0.00
26	2	19.00	0.00	5.00
27	1	0.00	0.00	0.00
28	2	4.50	0.00	5.00
29	1	0.00	0.00	0.00
30	2	1.00	4.50	0.00
31	3	27.00	3.00	12.67
32	1	0.00	0.00	0.00
33	2	4.00	0.00	3.00
34	3	0.00	3.00	0.00
35	1	0.00	0.00	0.00
36	4	1.50	0.00	0.00
37	3	0.00	0.00	0.00
38	6	13.50	0.00	4.67
39	8	4.75	0.00	10.12
40	1	10.00	0.00	0.00
41	1	0.00	0.00	0.00
42	1	10.00	10.00	10.00
43	1	10.00	10.00	0.00
44	1	0.00	0.00	0.00
45	1	0.00	0.00	9.00
46	2	0.00	0.00	3.50
47	2	0.00	0.00	0.00
48	2	0.00	0.00	19.00
49	1	10.00	10.00	10.00
50	2	0.00	0.00	0.00
51	4	32.00	0.00	15.75

52	2	0.00	0.00	0.00
53	1	0.00	0.00	0.00
54	0	0.00	0.00	0.00
55	4	12.00	0.00	0.00
56	1	0.00	0.00	0.00
57	1	0.00	0.00	0.00
58	2	0.00	5.00	5.00
59	10	0.00	0.00	0.20
60	1	0.00	0.00	0.00
61	2	0.00	0.00	0.50
62	3	12.67	0.00	1.00
63	3	3.33	0.00	3.33
64	2	5.00	0.00	0.00
65	6	1.67	0.00	1.67
66	5	7.60	7.60	7.60
67	2	0.00	0.00	0.00
68	1	10.00	0.00	0.00
69	1	10.00	0.00	9.00
70	2	0.00	3.00	0.00
71	1	0.00	0.00	9.00
72	1	0.00	0.00	10.00
73	1	10.00	0.00	0.00
74	1	0.00	9.00	0.00
75	1	0.00	0.00	0.00
76	1	2.00	0.00	0.00
77	2	2.00	0.00	0.00
78	1	0.00	0.00	0.00
79	1	0.00	0.00	10.00
80	2	0.00	5.00	19.00
81	2	1.00	0.00	0.00
82	1	10.00	0.00	10.00
83	1	0.00	0.00	9.00
84	1	6.00	0.00	0.00
85	2	5.00	0.00	16.00
86	1	0.00	0.00	0.00
87	1	10.00	10.00	10.00
88	1	4.00	0.00	0.00

89	1	0.00	0.00	0.00
90	1	0.00	0.00	0.00
91	1	10.00	0.00	9.00
92	1	0.00	0.00	0.00
93	1	10.00	10.00	10.00
94	1	10.00	0.00	10.00
95	1	0.00	0.00	0.00
96	1	0.00	0.00	0.00
97	1	0.00	10.00	10.00
98	1	0.00	0.00	0.00
99	1	10.00	0.00	10.00
100	1	0.00	0.00	0.00
101	1	0.00	0.00	0.00
102	1	0.00	0.00	0.00
103	2	19.00	0.00	19.00
104	1	10.00	0.00	10.00
105	1	10.00	0.00	0.00
106	1	0.00	0.00	0.00
107	1	0.00	0.00	0.00
108	1	0.00	0.00	0.00
109	1	0.00	0.00	0.00
110	2	12.00	0.00	0.00
111	2	15.00	0.00	5.00
112	1	10.00	0.00	10.00
113	1	10.00	10.00	3.00
114	1	10.00	10.00	10.00
115	5	13.80	2.00	24.00
116	6	3.33	0.00	0.00

Table 45: Key Figure 2 by person ID for Test 4-3

Person ID	Number of reference images	Algorithm 1	Algorithm 2	Algorithm 3
1	5	0.00	0.00	0.00
2	10	17.50	19.00	13.60
3	10	13.60	8.10	13.20
4	6	0.00	0.00	0.00
5	4	9.50	0.00	9.50

6	7	26.43	0.00	16.57
7	6	13.50	1.67	13.50
8	3	2.67	0.00	0.00
9	6	0.00	0.00	0.00
10	2	0.00	0.00	0.00
11	5	2.00	1.80	0.00
12	1	0.00	0.00	0.00
13	8	29.25	2.75	0.00
14	7	0.00	1.43	0.00
15	10	11.20	0.00	8.10
16	6	6.33	1.67	1.67
17	3	0.00	0.00	0.00
18	2	0.00	0.00	0.00
19	6	0.00	0.00	0.67
20	4	0.00	0.00	0.00
21	5	0.00	0.00	7.60
22	3	12.00	3.33	0.00
23	6	4.33	0.00	0.00
24	3	26.00	0.00	3.33
25	3	0.00	0.00	0.00
26	2	19.00	0.00	5.00
27	1	0.00	0.00	0.00
28	2	4.50	0.00	0.00
29	1	0.00	0.00	0.00
30	2	0.00	0.00	0.00
31	3	21.00	0.00	10.67
32	1	0.00	0.00	0.00
33	2	0.00	0.00	0.00
34	3	0.00	3.00	0.00
35	1	0.00	0.00	0.00
36	4	0.00	0.00	0.00
37	3	0.00	0.00	0.00
38	6	6.33	0.00	1.67
39	8	4.25	0.00	9.00
40	1	10.00	0.00	0.00
41	1	0.00	0.00	0.00
42	1	9.00	10.00	8.00

43	1	10.00	10.00	0.00
44	1	0.00	0.00	0.00
45	1	0.00	0.00	0.00
46	2	0.00	0.00	0.00
47	2	0.00	0.00	0.00
48	2	0.00	0.00	19.00
49	1	10.00	10.00	10.00
50	2	0.00	0.00	0.00
51	4	19.50	0.00	8.50
52	2	0.00	0.00	0.00
53	1	0.00	0.00	0.00
54	0	0.00	0.00	0.00
55	4	1.50	0.00	0.00
56	1	0.00	0.00	0.00
57	1	0.00	0.00	0.00
58	2	0.00	5.00	5.00
59	10	0.00	0.00	0.00
60	1	0.00	0.00	0.00
61	2	0.00	0.00	0.00
62	3	12.67	0.00	0.00
63	3	3.33	0.00	3.33
64	2	5.00	0.00	0.00
65	6	1.67	0.00	1.67
66	5	5.20	4.40	7.20
67	2	0.00	0.00	0.00
68	1	8.00	0.00	0.00
69	1	9.00	0.00	0.00
70	2	0.00	0.00	0.00
71	1	0.00	0.00	5.00
72	1	0.00	0.00	10.00
73	1	10.00	0.00	0.00
74	1	0.00	5.00	0.00
75	1	0.00	0.00	0.00
76	1	0.00	0.00	0.00
77	2	0.00	0.00	0.00
78	1	0.00	0.00	0.00
79	1	0.00	0.00	10.00

80	2	0.00	5.00	17.00
81	2	0.00	0.00	0.00
82	1	10.00	0.00	6.00
83	1	0.00	0.00	0.00
84	1	0.00	0.00	0.00
85	2	5.00	0.00	5.00
86	1	0.00	0.00	0.00
87	1	10.00	10.00	10.00
88	1	0.00	0.00	0.00
89	1	0.00	0.00	0.00
90	1	0.00	0.00	0.00
91	1	10.00	0.00	0.00
92	1	0.00	0.00	0.00
93	1	10.00	10.00	10.00
94	1	10.00	0.00	10.00
95	1	0.00	0.00	0.00
96	1	0.00	0.00	0.00
97	1	0.00	10.00	10.00
98	1	0.00	0.00	0.00
99	1	9.00	0.00	8.00
100	1	0.00	0.00	0.00
101	1	0.00	0.00	0.00
102	1	0.00	0.00	0.00
103	2	19.00	0.00	19.00
104	1	10.00	0.00	10.00
105	1	10.00	0.00	0.00
106	1	0.00	0.00	0.00
107	1	0.00	0.00	0.00
108	1	0.00	0.00	0.00
109	1	0.00	0.00	0.00
110	2	5.00	0.00	0.00
111	2	5.00	0.00	5.00
112	1	10.00	0.00	10.00
113	1	9.00	4.00	0.00
114	1	10.00	10.00	10.00
115	5	7.60	2.00	0.00
116	6	0.00	0.00	0.00

Table 46: Key Figure 2 by person ID for Test 4-4

Annex B Additional results for Test Group 5

B.1 Notes to the diagrams[29]

For many tests no results could be obtained (see also 5.7.5.1 Enrolment and verifications). The tests concerned are therefore missing from the sections below.

[29] See also 5.7.2 Notes on the interpretation of the diagrams.

B.2 Algorithm 1

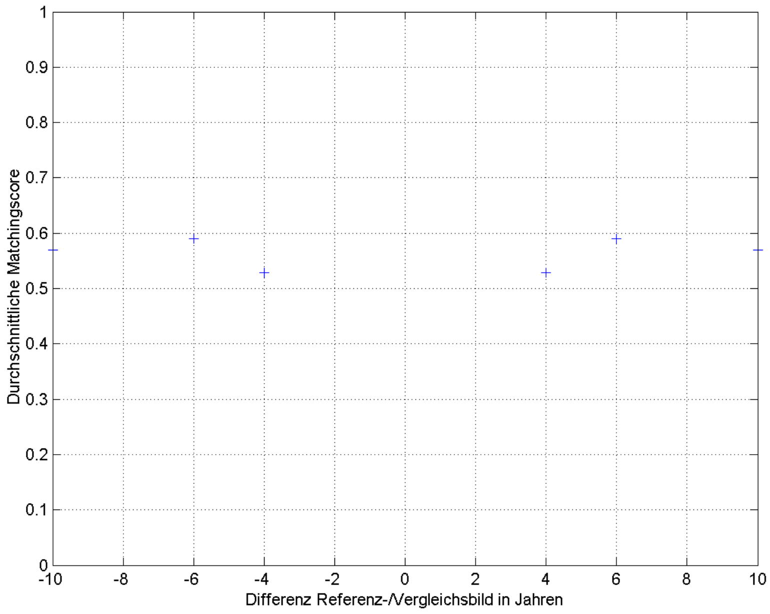


Figure 131: 5-6 Algorithm 1

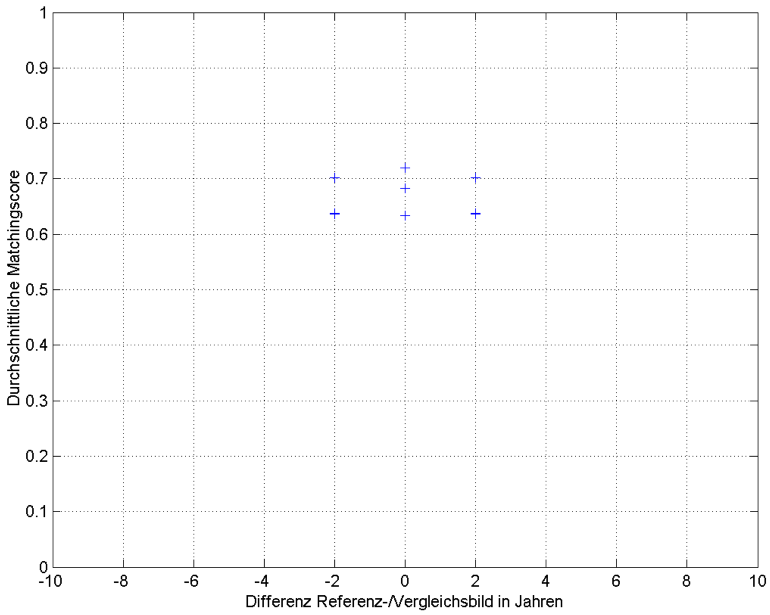


Figure 132: 5-7 Algorithm 1

Durchschnittliche Matchingscore	Average matching score
Differenz Referenz-/ Vergleichsbild in Jahren	Difference between reference and comparison images in years

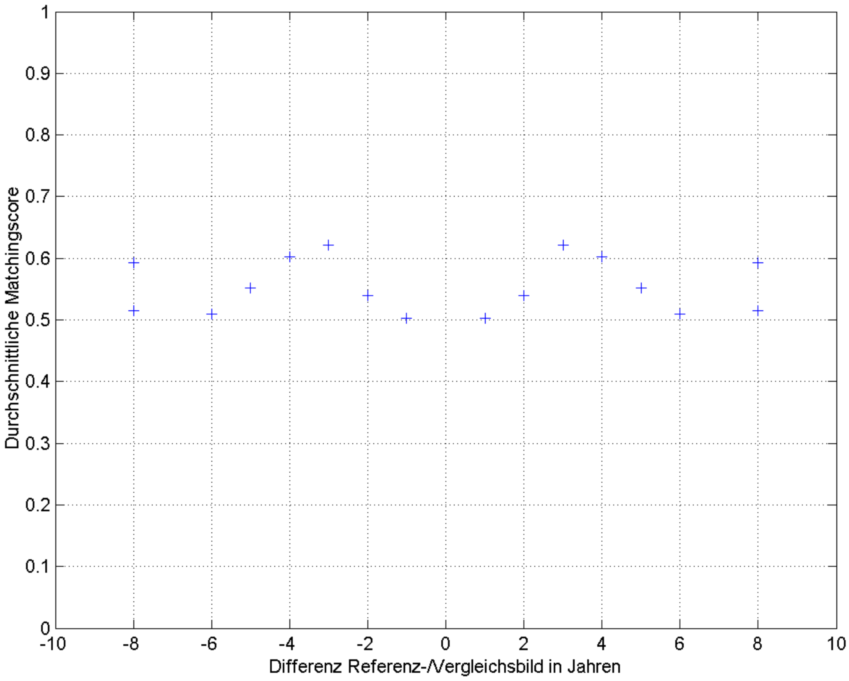


Figure 133: 5-8 Algorithm 1

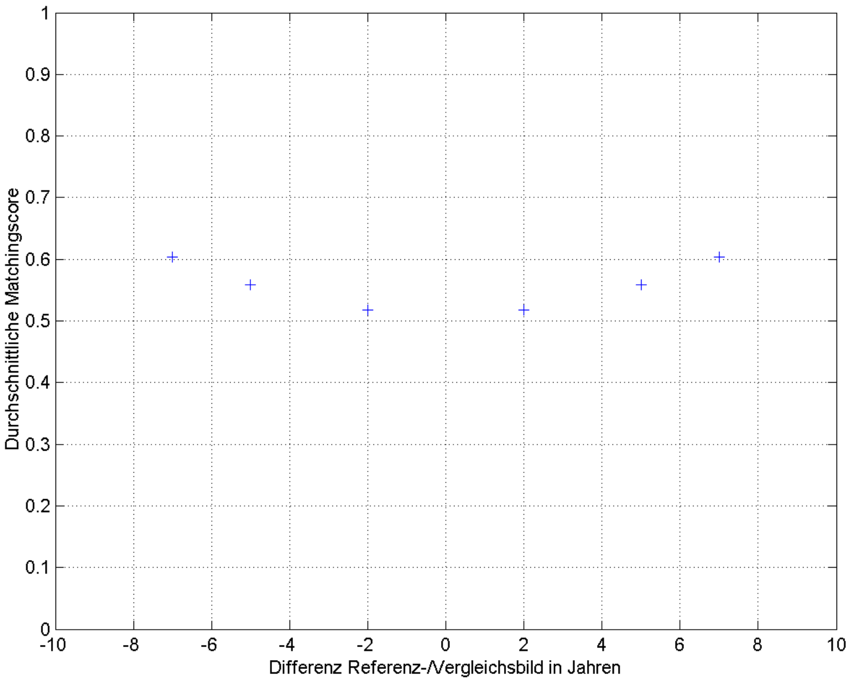


Figure 134: 5-9 Algorithm 1

Durchschnittliche Matchingscore	Average matching score
Differenz Referenz-/ Vergleichsbild in Jahren	Difference between reference and comparison images in years

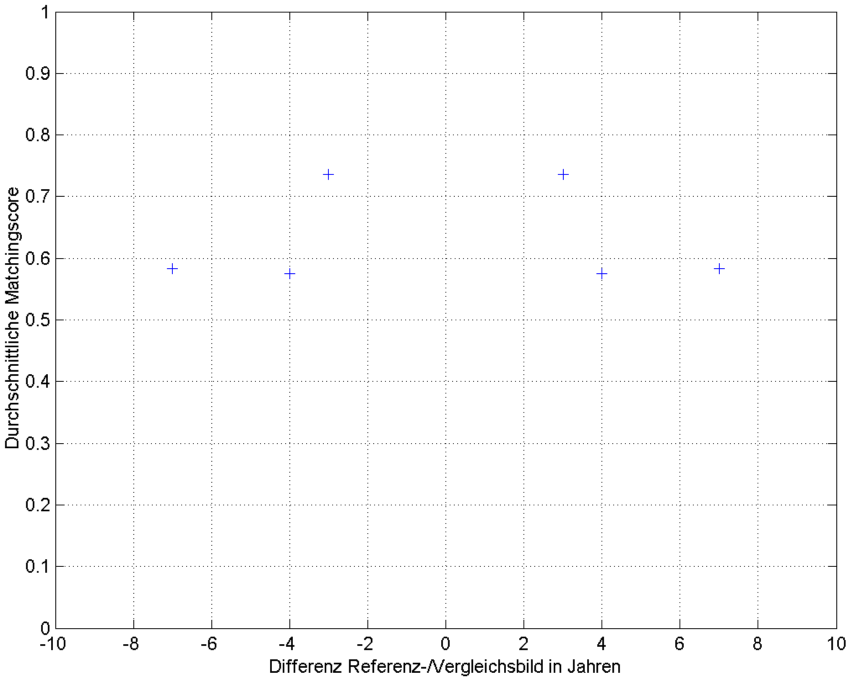


Figure 135: 5-10 Algorithm 1

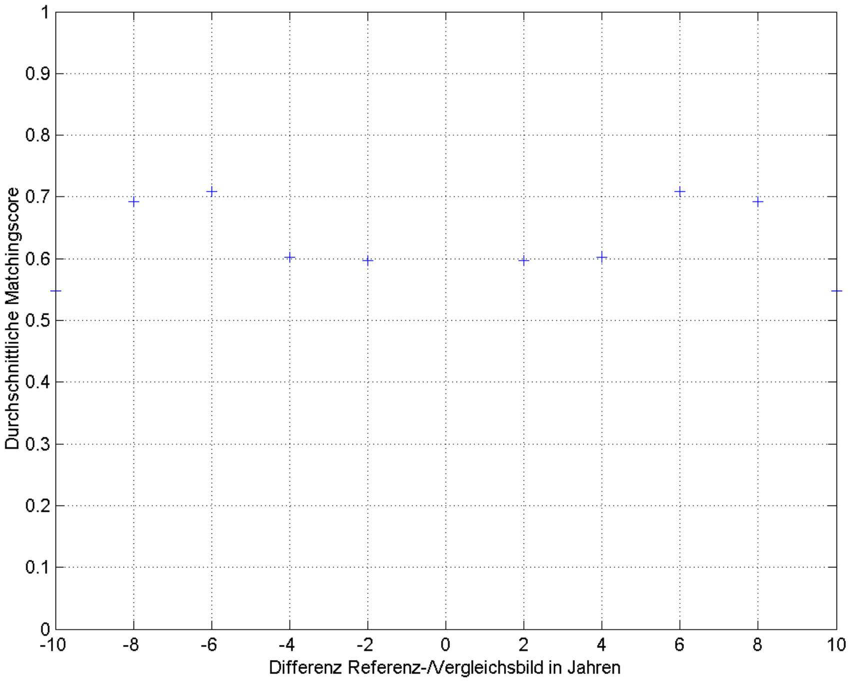


Figure 136: 5-11 Algorithm 1

Durchschnittliche Matchingscore	Average matching score
Differenz Referenz-/ Vergleichsbild in Jahren	Difference between reference and comparison images in years

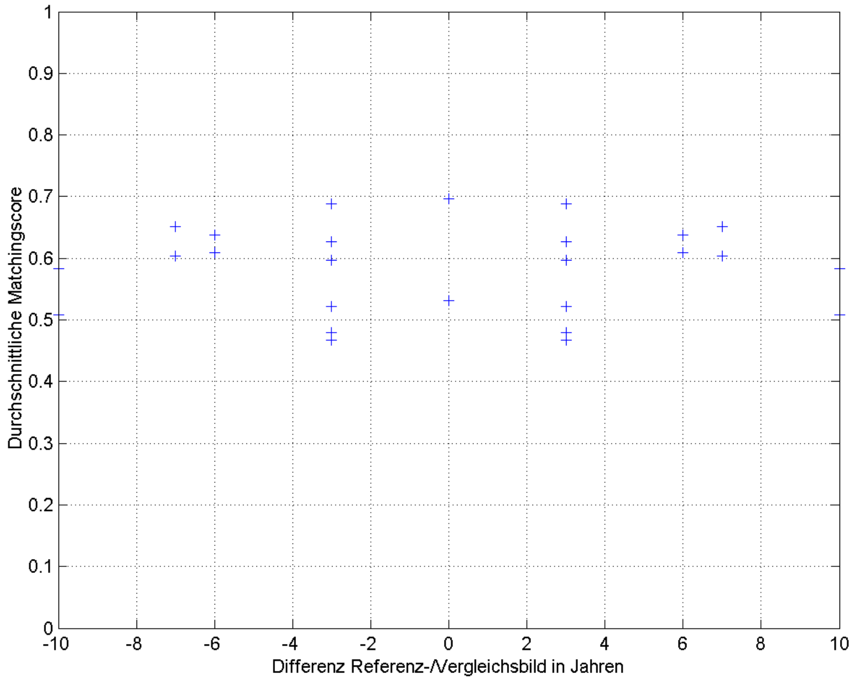


Figure 137: 5-12 Algorithm 1

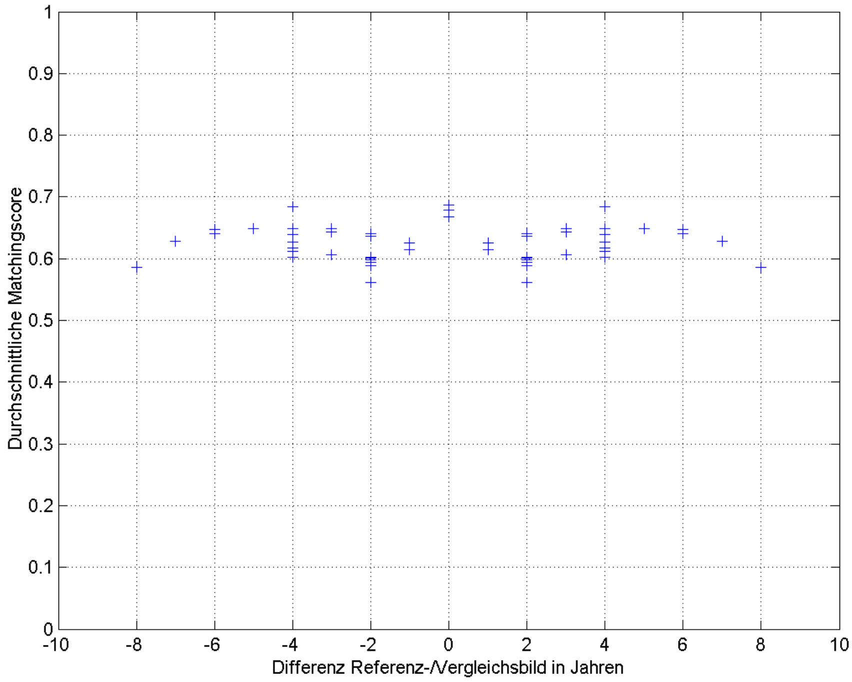


Figure 138: 5-13 Algorithm 1

Durchschnittliche Matchingscore	Average matching score
Differenz Referenz-/ Vergleichsbild in Jahren	Difference between reference and comparison images in years

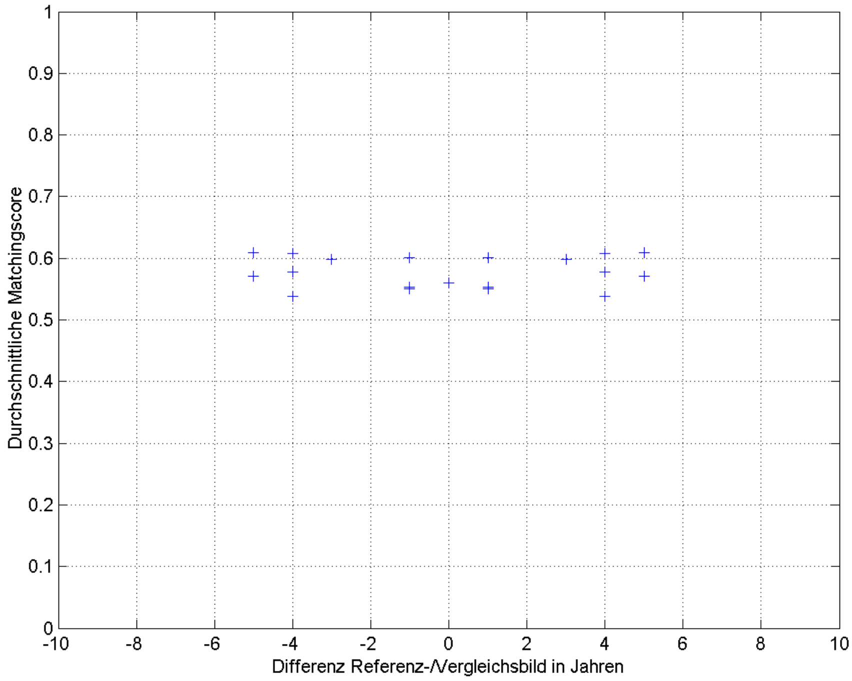


Figure 139: 5-14 Algorithm 1

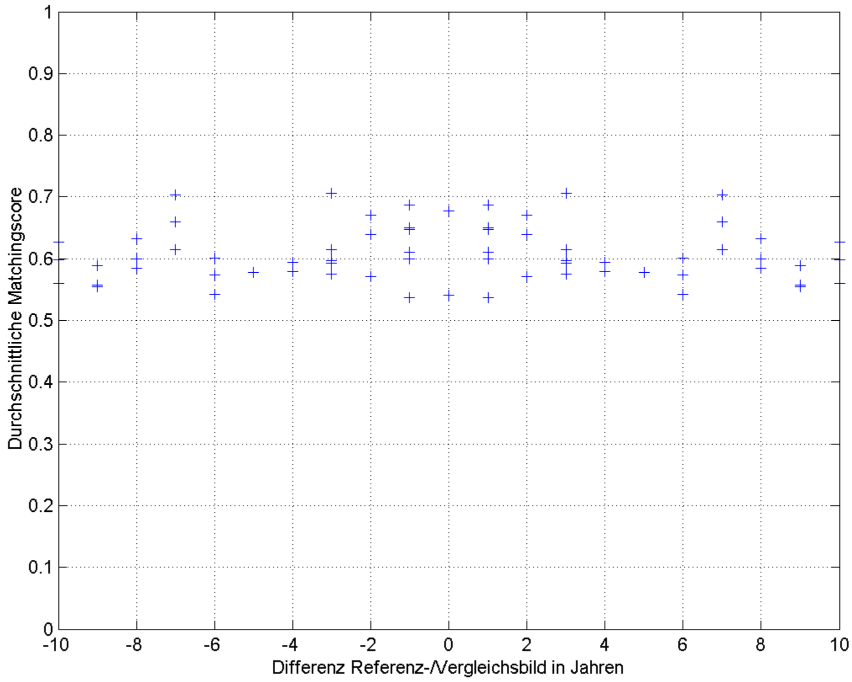


Figure 140: 5-15 Algorithm 1

Durchschnittliche Matchingscore	Average matching score
Differenz Referenz-/ Vergleichsbild in Jahren	Difference between reference and comparison images in years

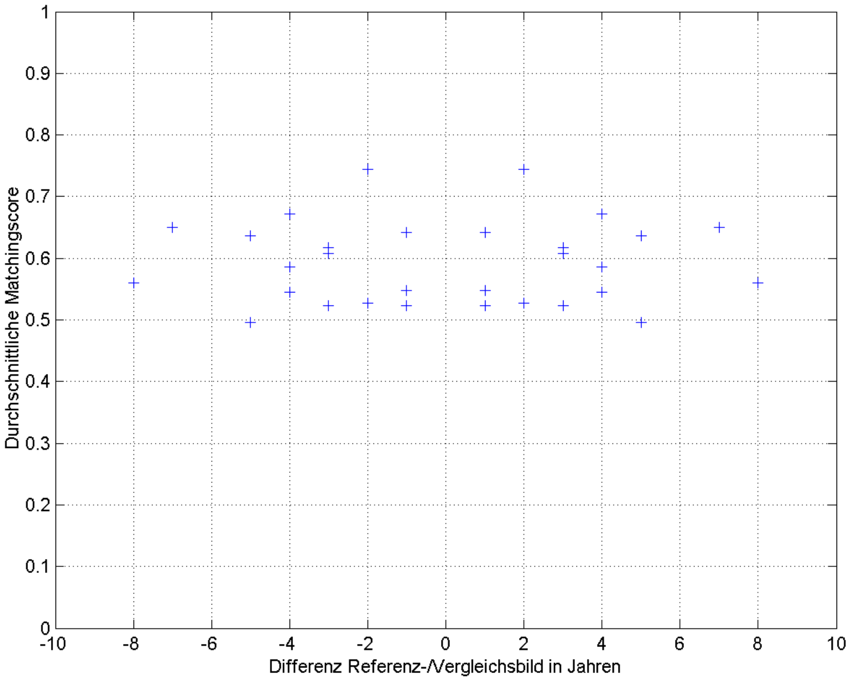


Figure 141: 5-16 Algorithm 1

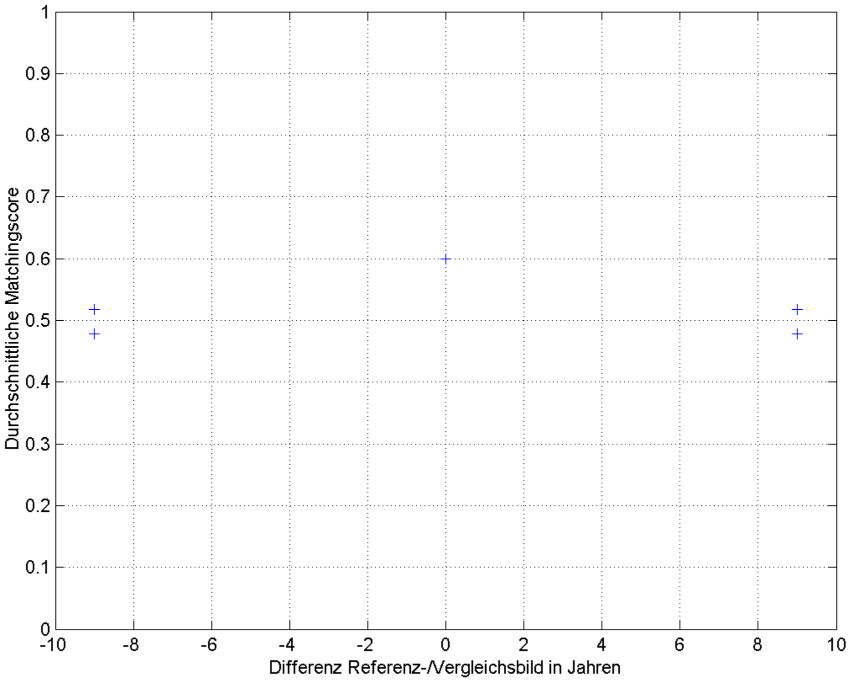


Figure 142: 5-17 Algorithm 1

Durchschnittliche Matchingscore	Average matching score
Differenz Referenz-/ Vergleichsbild in Jahren	Difference between reference and comparison images in years

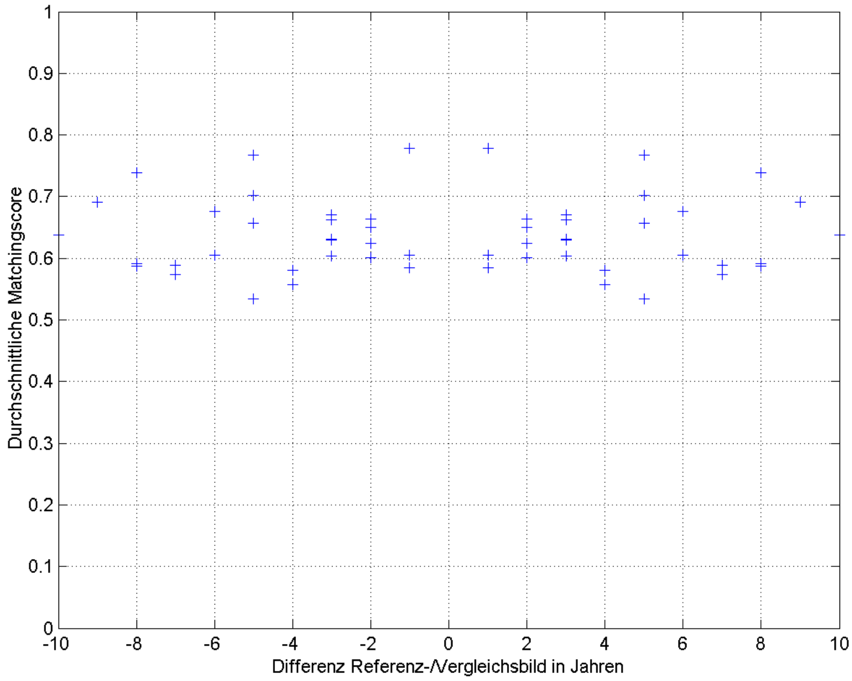


Figure 143: 5-18 Algorithm 1

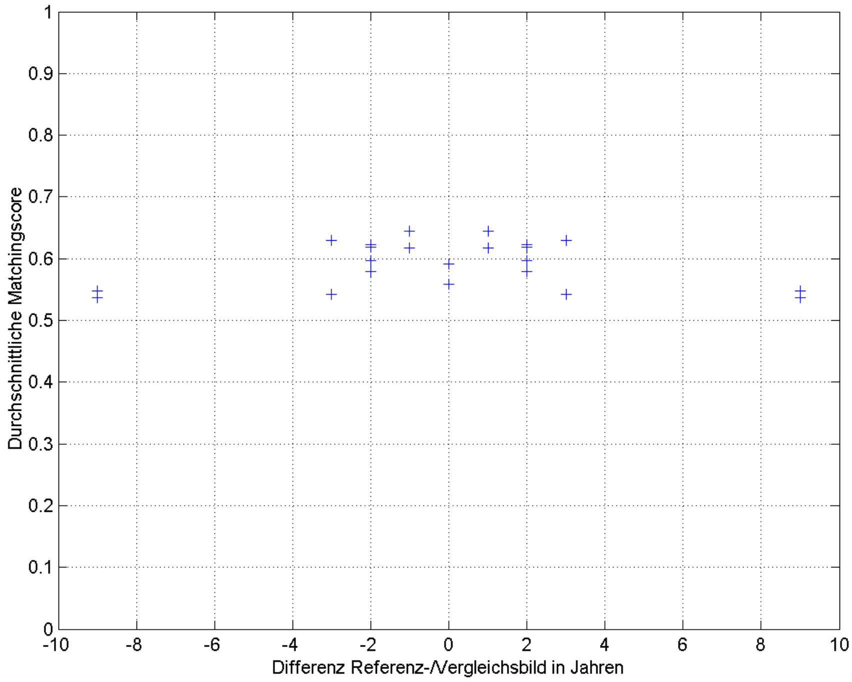


Figure 144: 5-19 Algorithm 1

Durchschnittliche Matchingscore	Average matching score
Differenz Referenz-/ Vergleichsbild in Jahren	Difference between reference and comparison images in years

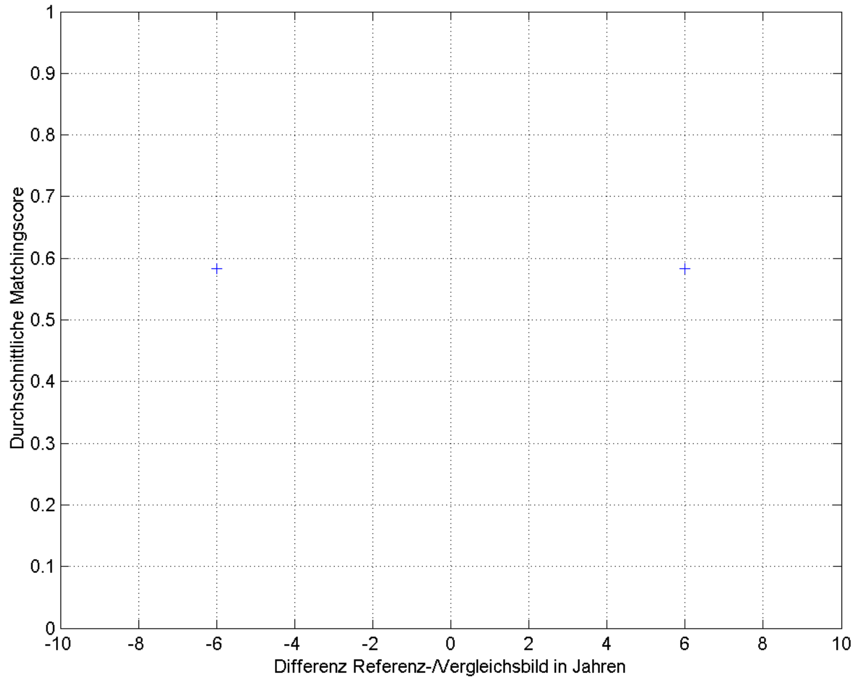


Figure 145: 5-20 Algorithm 1

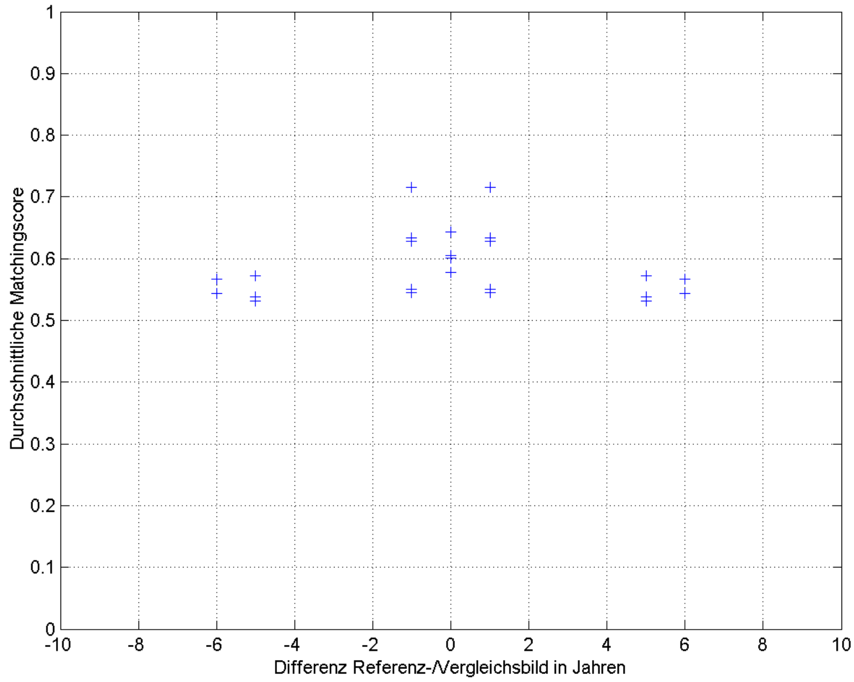


Figure 146: 5-21 Algorithm 1

Durchschnittliche Matchingscore	Average matching score
Differenz Referenz-/ Vergleichsbild in Jahren	Difference between reference and comparison images in years

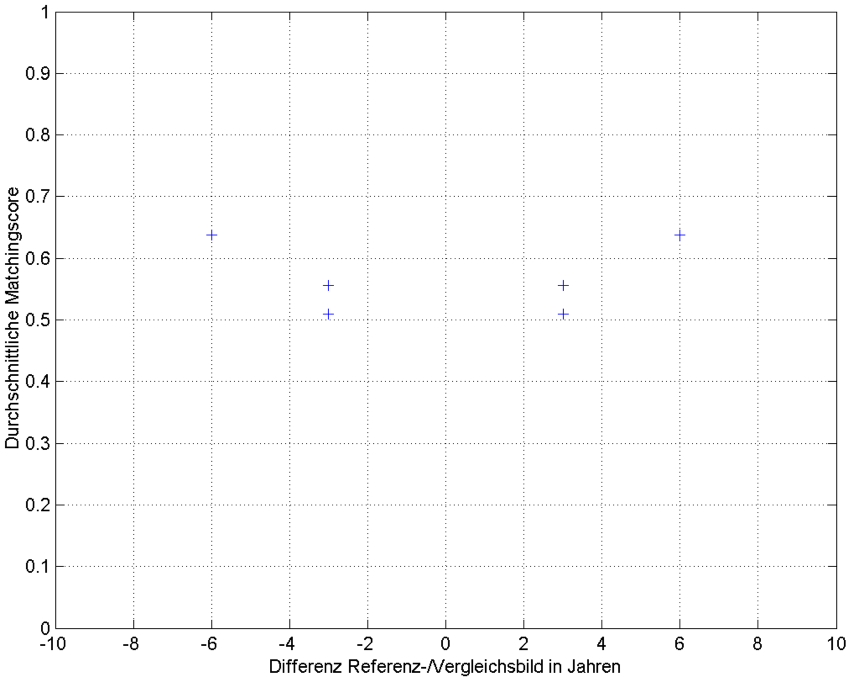


Figure 147: 5-23 Algorithm 1

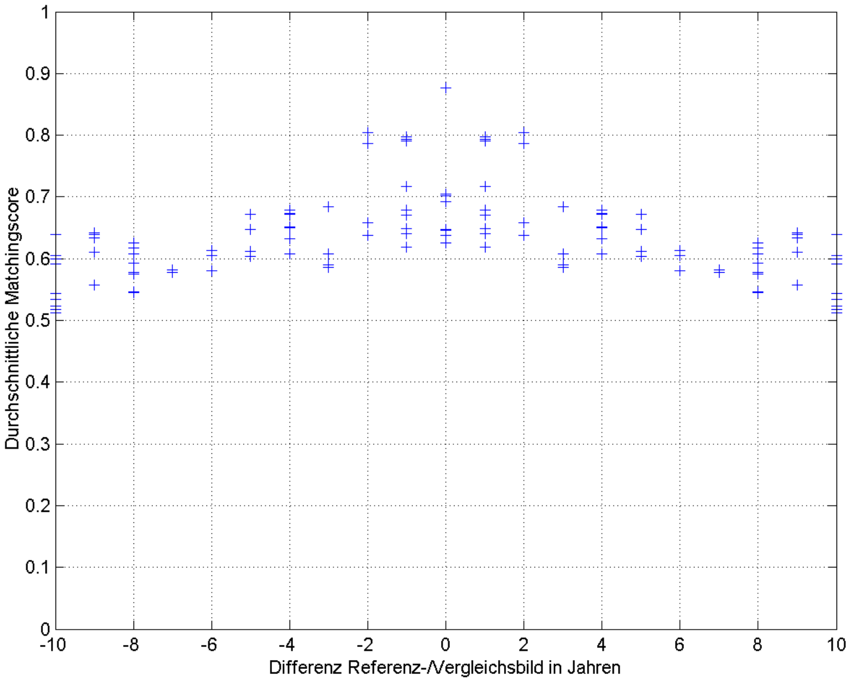


Figure 148: 5-24 Algorithm 1

Durchschnittliche Matchingscore	Average matching score
Differenz Referenz-/ Vergleichsbild in Jahren	Difference between reference and comparison images in years

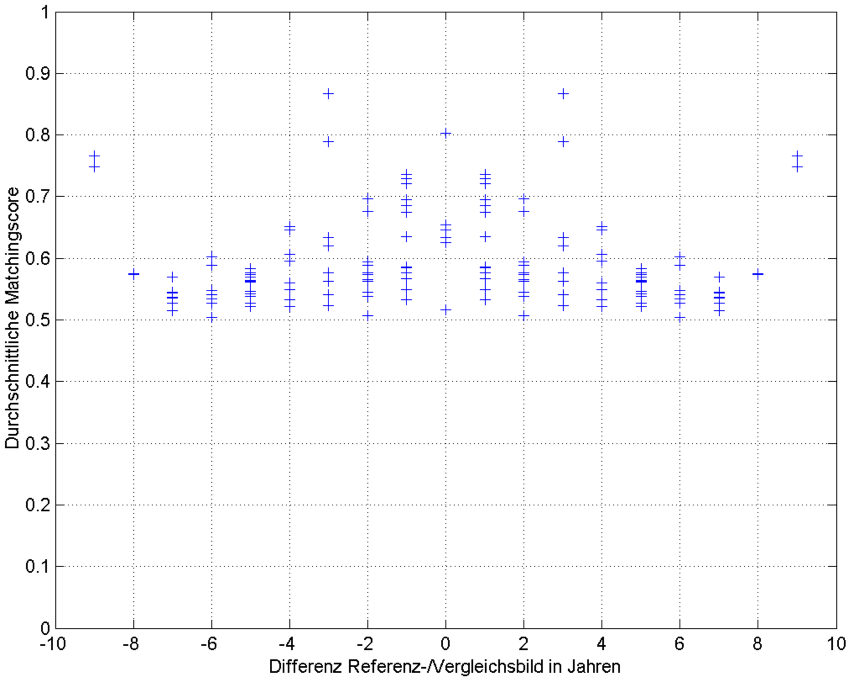


Figure 149: 5-25 Algorithm 1

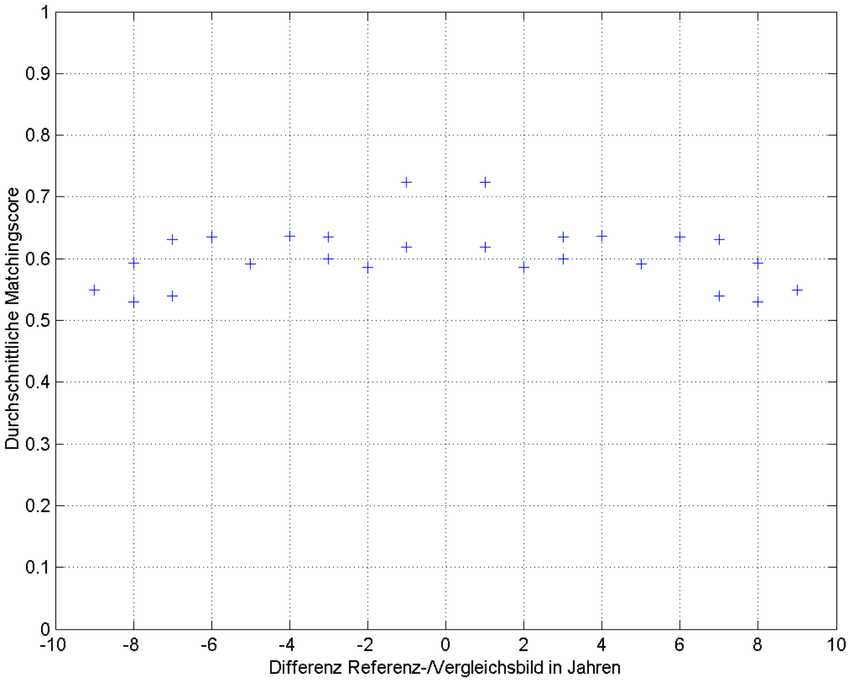


Figure 150: 5-26 Algorithm 1

Durchschnittliche Matchingscore	Average matching score
Differenz Referenz-/ Vergleichsbild in Jahren	Difference between reference and comparison images in years

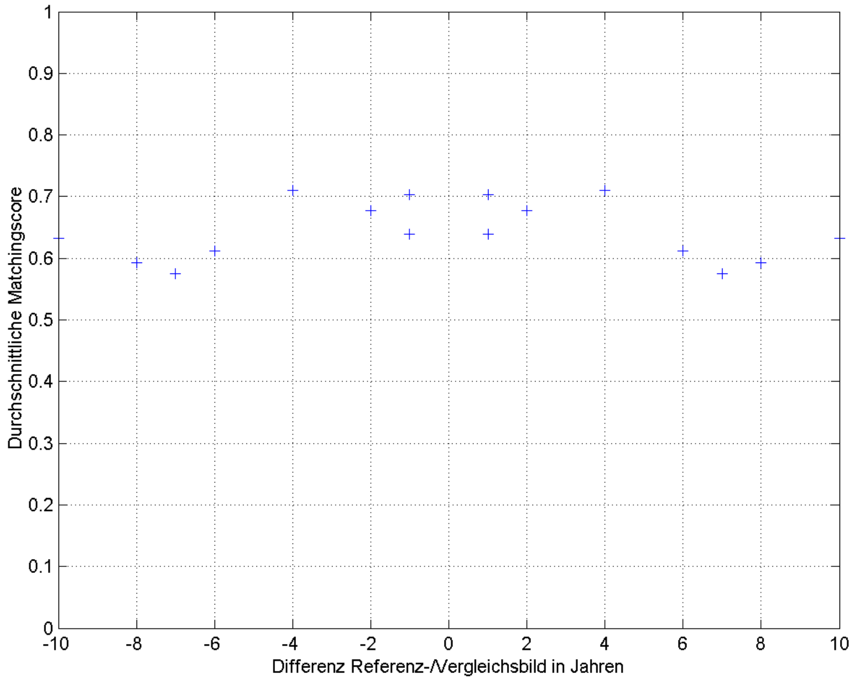


Figure 151: 5-27 Algorithm 1

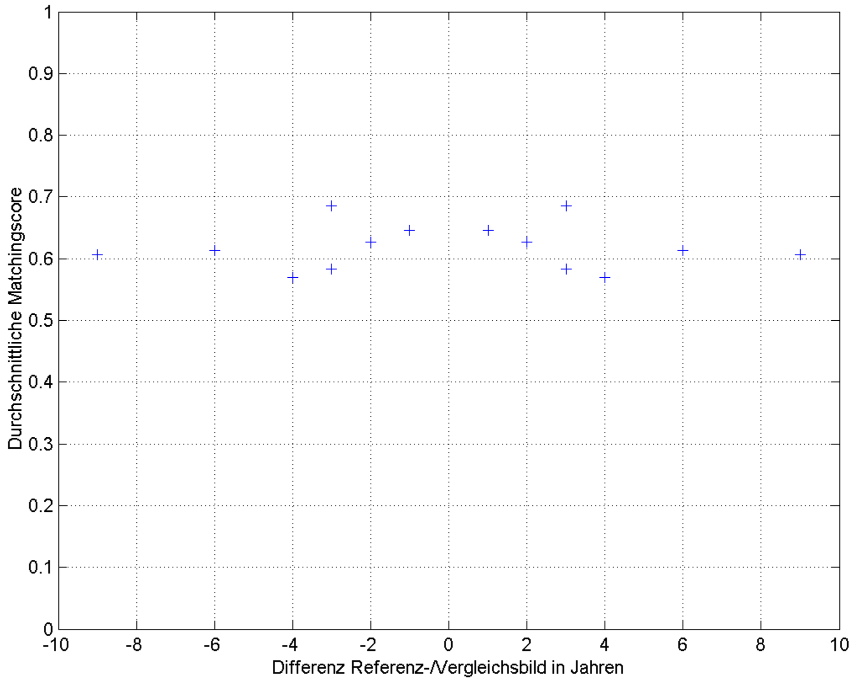


Figure 152: 5-28 Algorithm 1

Durchschnittliche Matchingscore	Average matching score
Differenz Referenz-/ Vergleichsbild in Jahren	Difference between reference and comparison images in years

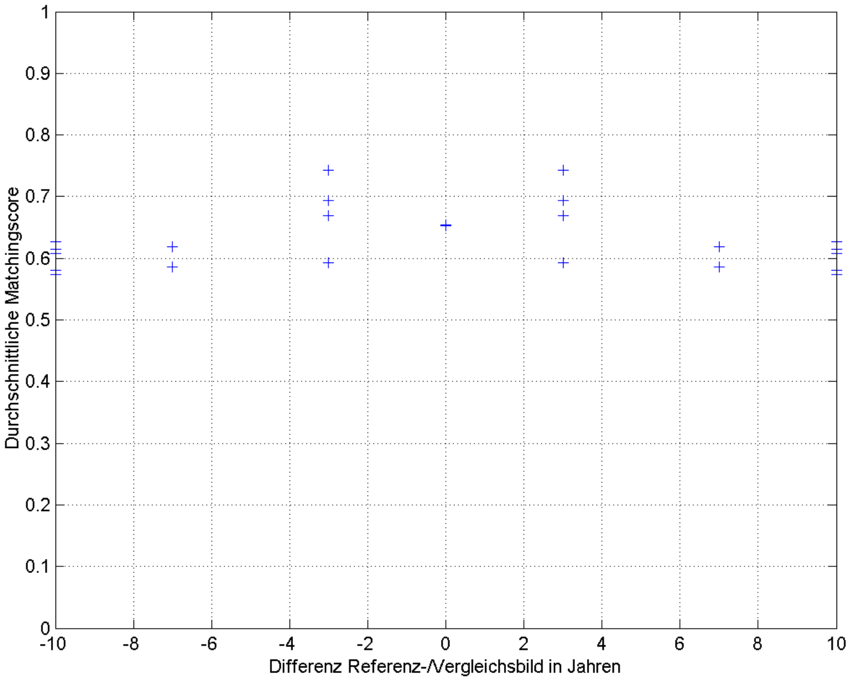


Figure 153: 5-29 Algorithm 1

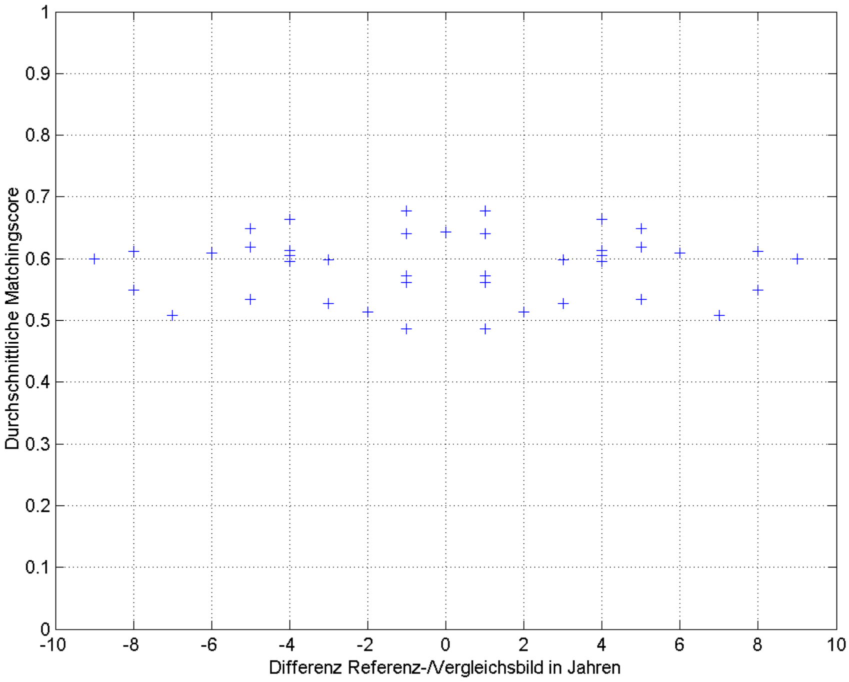


Figure 154: 5-30 Algorithm 1

Durchschnittliche Matchingscore	Average matching score
Differenz Referenz-/ Vergleichsbild in Jahren	Difference between reference and comparison images in years

B.3 Algorithm 2

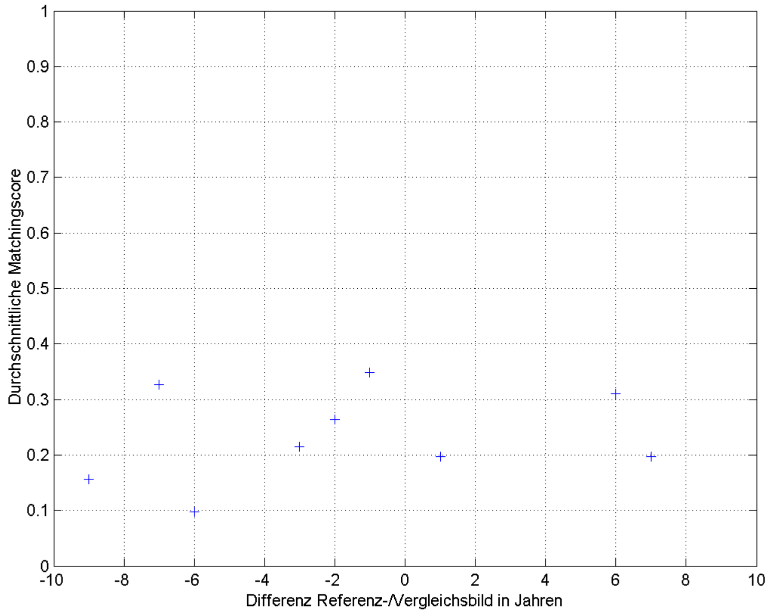


Figure 155: 5-6 Algorithm 2

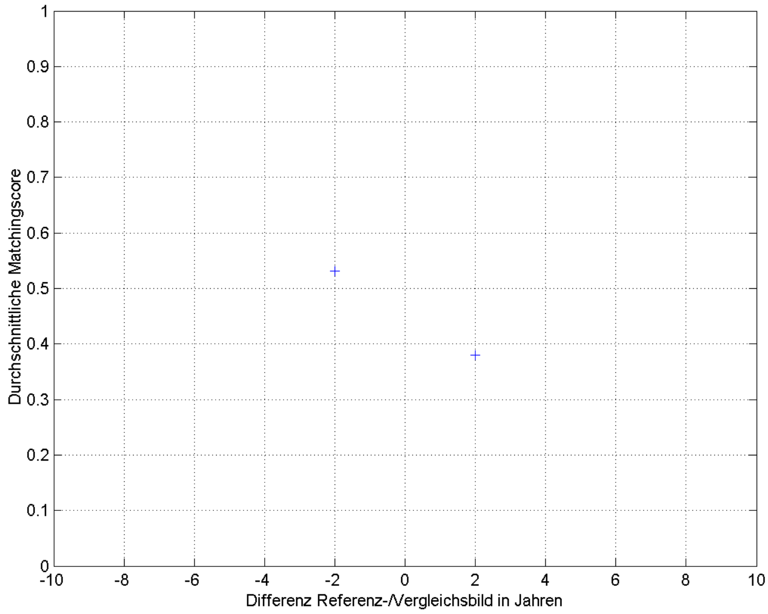


Figure 156: 5-7 Algorithm 2

Durchschnittliche Matchingscore	Average matching score
Differenz Referenz-/ Vergleichsbild in Jahren	Difference between reference and comparison images in years

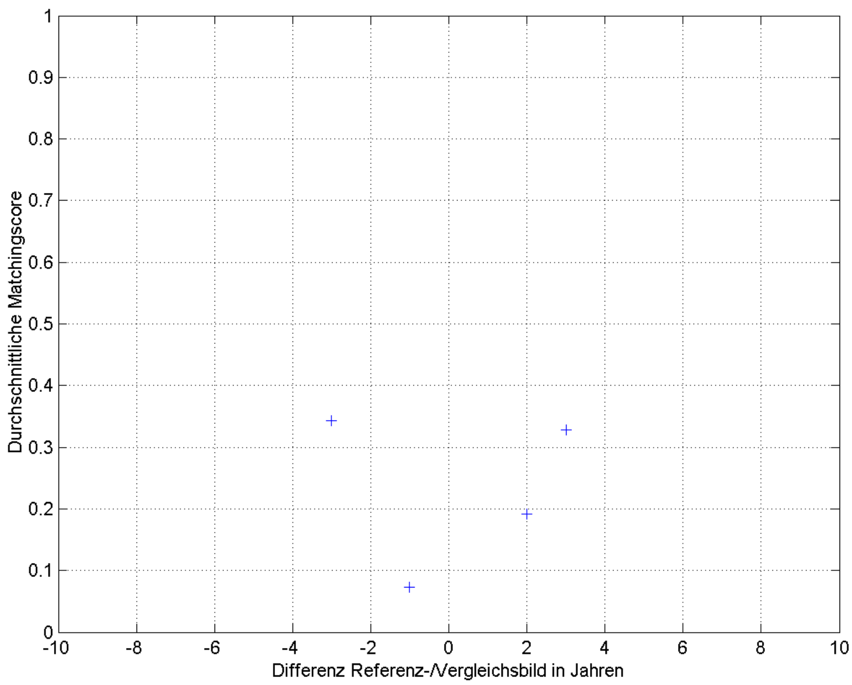


Figure 157: 5-8 Algorithm 2

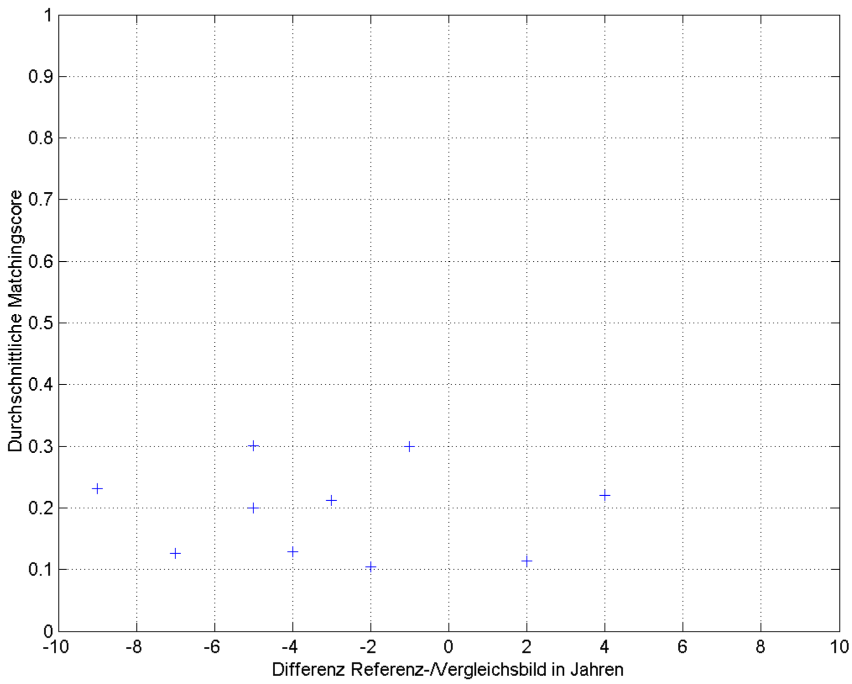


Figure 158: 5-9 Algorithm 2

Durchschnittliche Matchingscore	Average matching score
Differenz Referenz-/ Vergleichsbild in Jahren	Difference between reference and comparison images in years

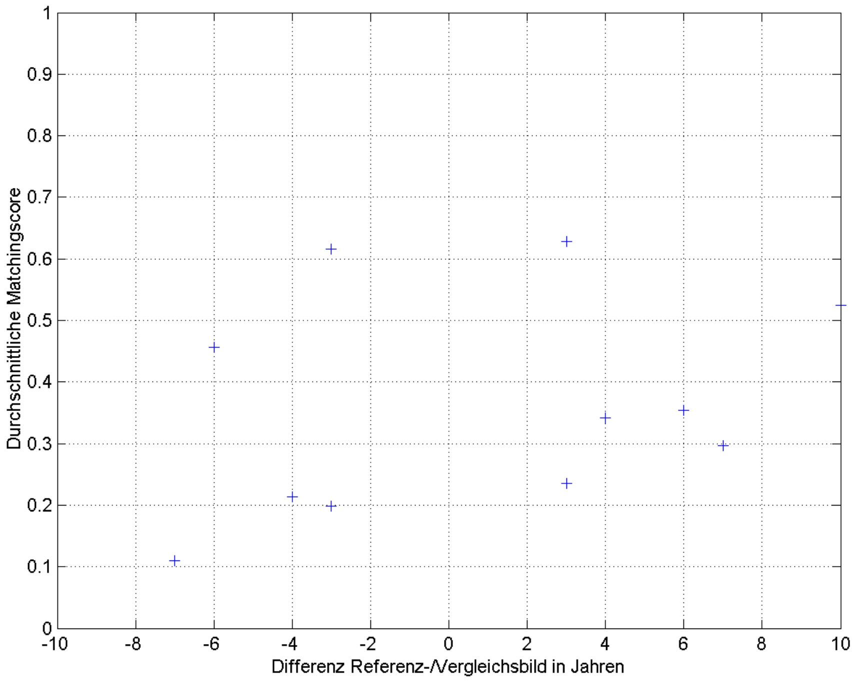


Figure 159: 5-10 Algorithm 2

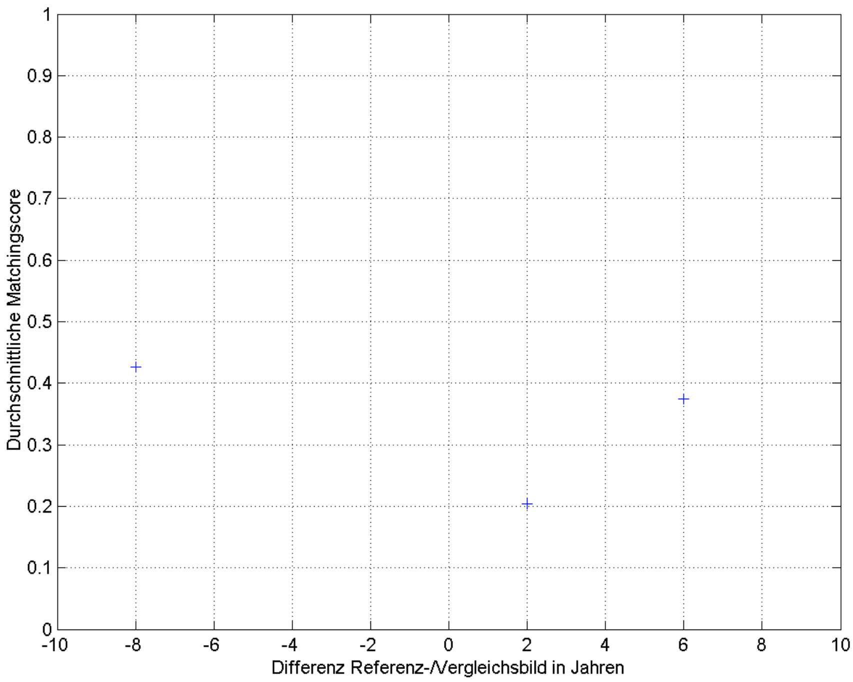


Figure 160: 5-11 Algorithm 2

Durchschnittliche Matchingscore	Average matching score
Differenz Referenz-/ Vergleichsbild in Jahren	Difference between reference and comparison images in years

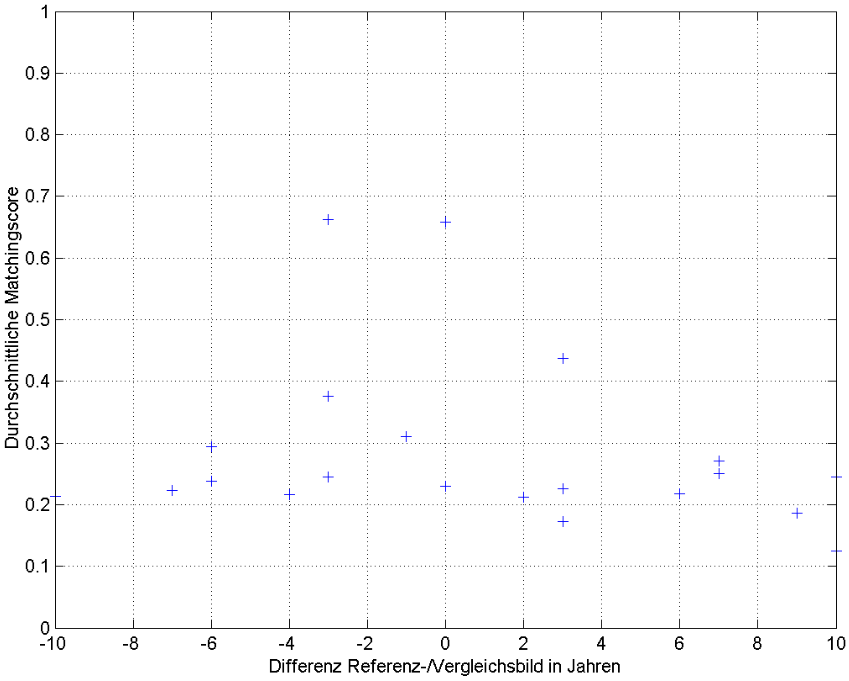


Figure 161: 5-12 Algorithm 2

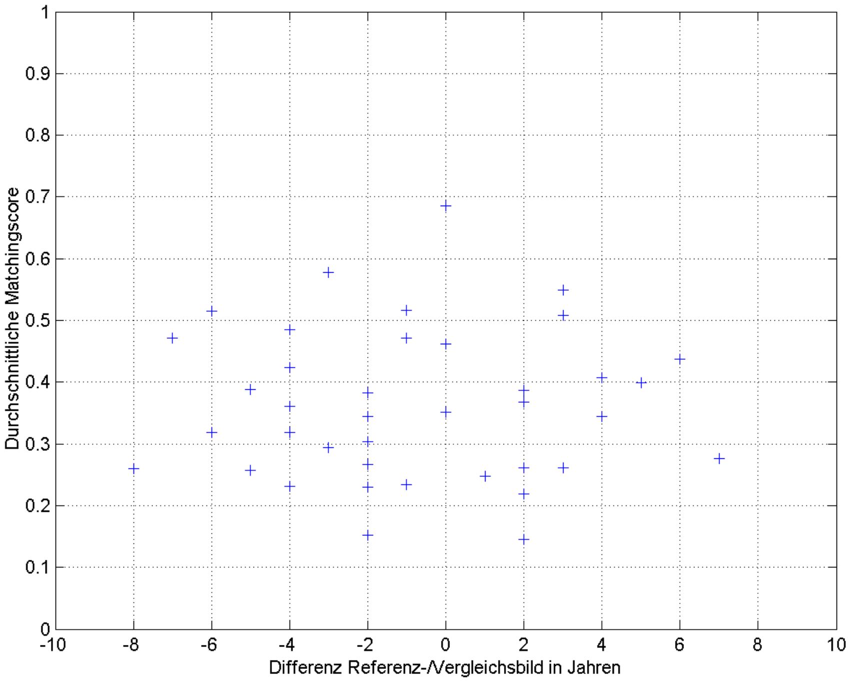


Figure 162: 5-13 Algorithm 2

Durchschnittliche Matchingscore	Average matching score
Differenz Referenz-/ Vergleichsbild in Jahren	Difference between reference and comparison images in years

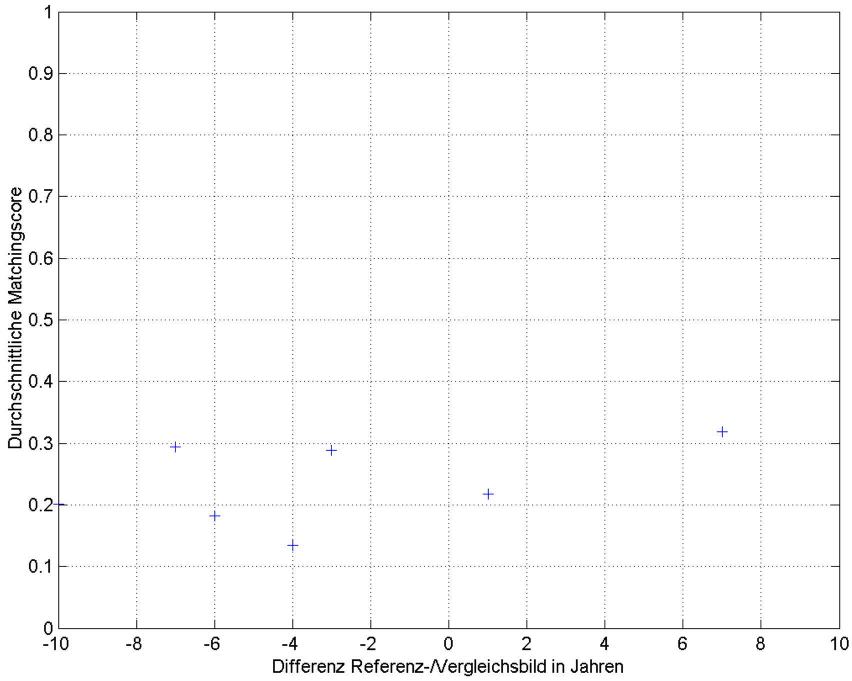


Figure 163: 5-14 Algorithm 2

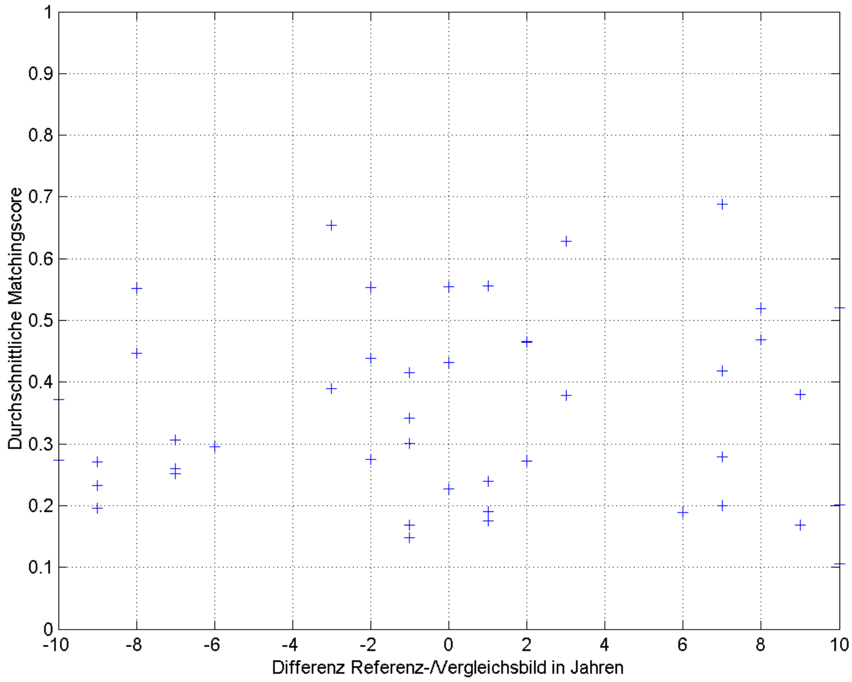


Figure 164: 5-15 Algorithm 2

Durchschnittliche Matchingscore	Average matching score
Differenz Referenz-/ Vergleichsbild in Jahren	Difference between reference and comparison images in years

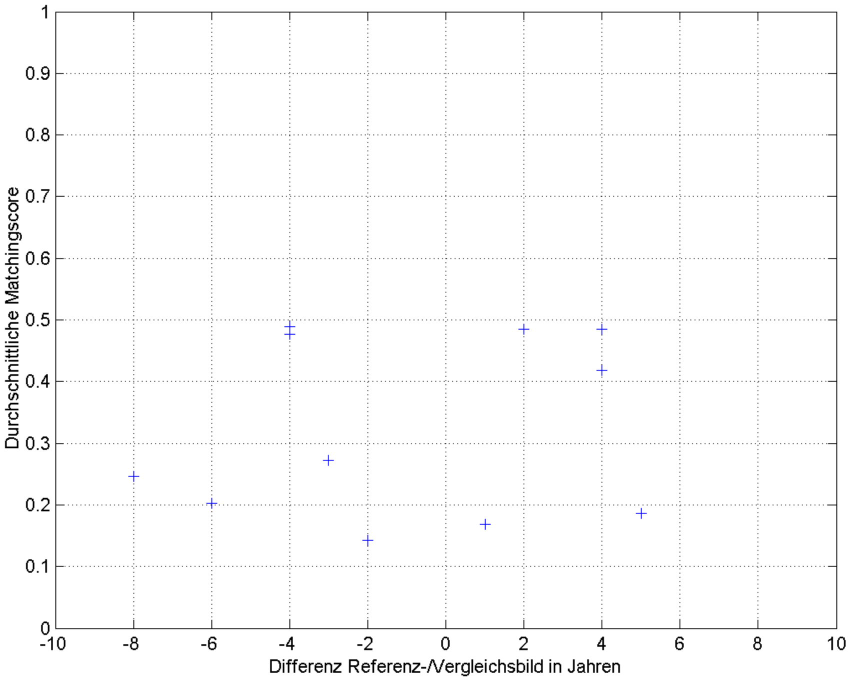


Figure 165: 5-16 Algorithm 2

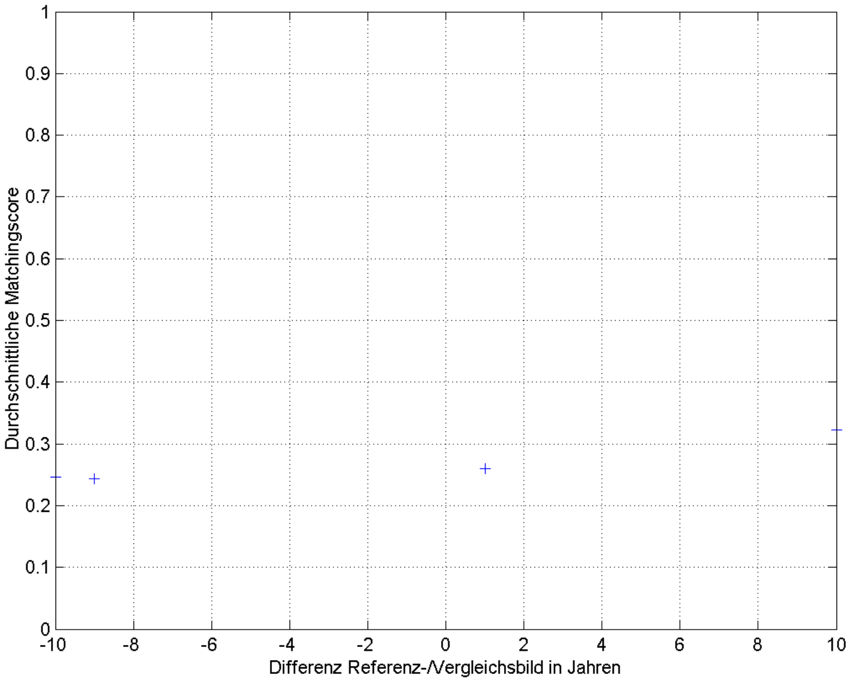


Figure 166: 5-17 Algorithm 2

Durchschnittliche Matchingscore	Average matching score
Differenz Referenz-/ Vergleichsbild in Jahren	Difference between reference and comparison images in years

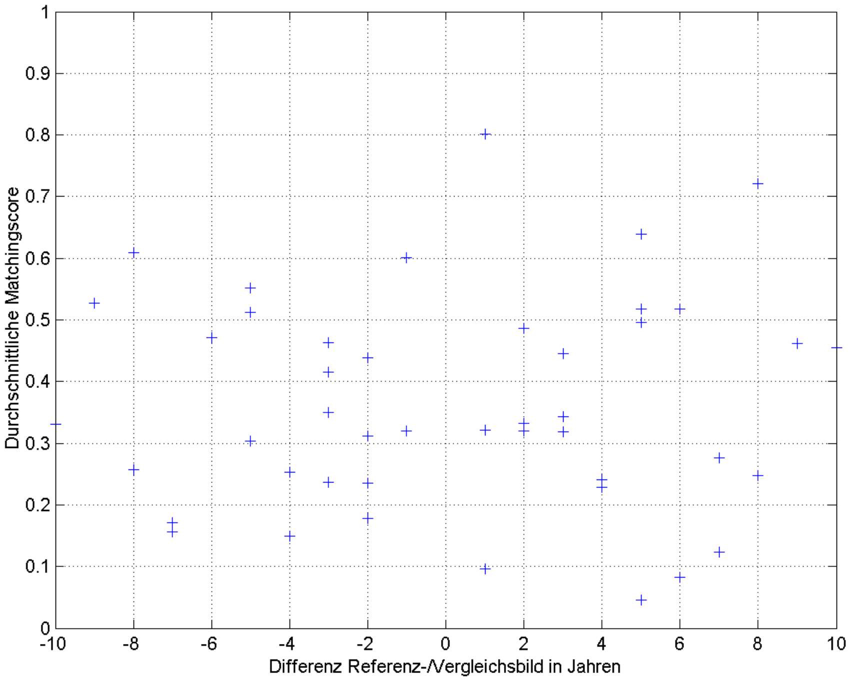


Figure 167: 5-18 Algorithm 2

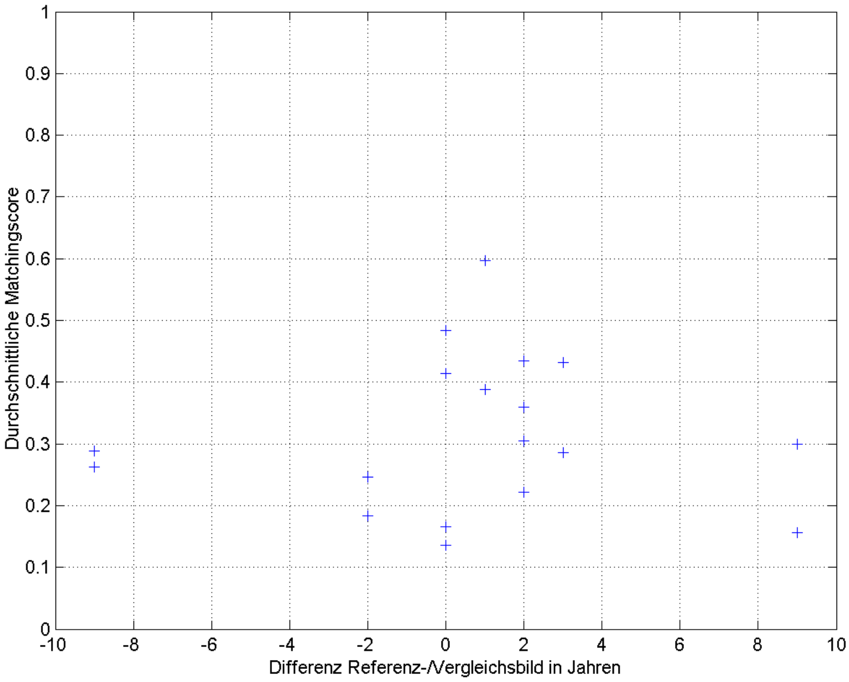


Figure 168: 5-19 Algorithm 2

Durchschnittliche Matchingscore	Average matching score
Differenz Referenz-/ Vergleichsbild in Jahren	Difference between reference and comparison images in years

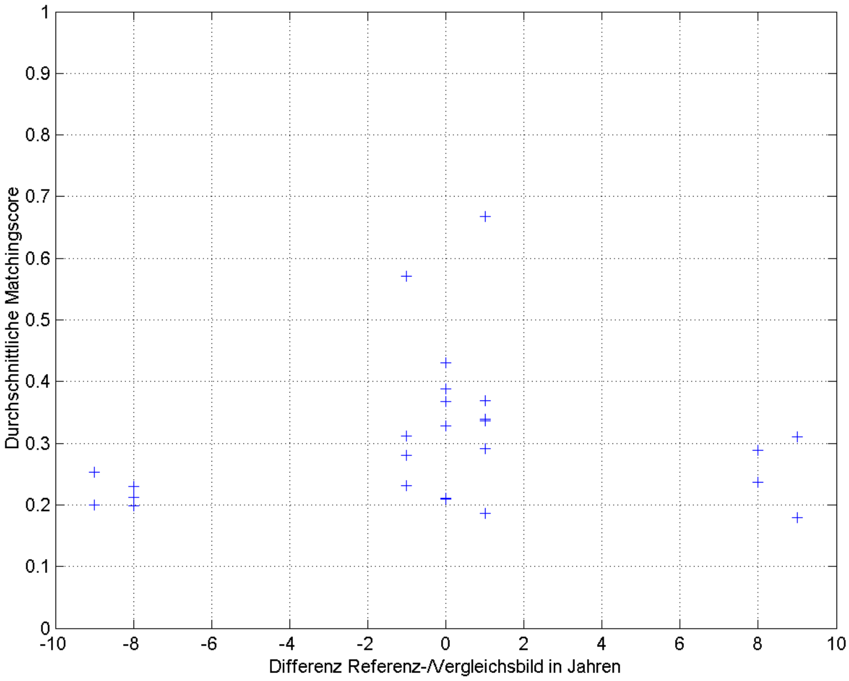


Figure 169: 5-21 Algorithm 2

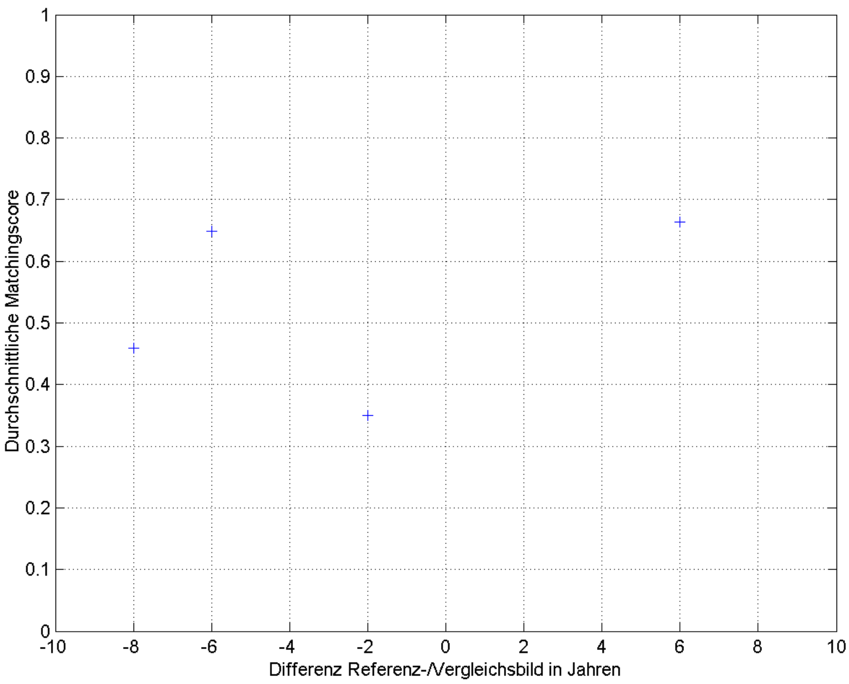


Figure 170: 5-23 Algorithm 2

Durchschnittliche Matchingscore	Average matching score
Differenz Referenz-/ Vergleichsbild in Jahren	Difference between reference and comparison images in years

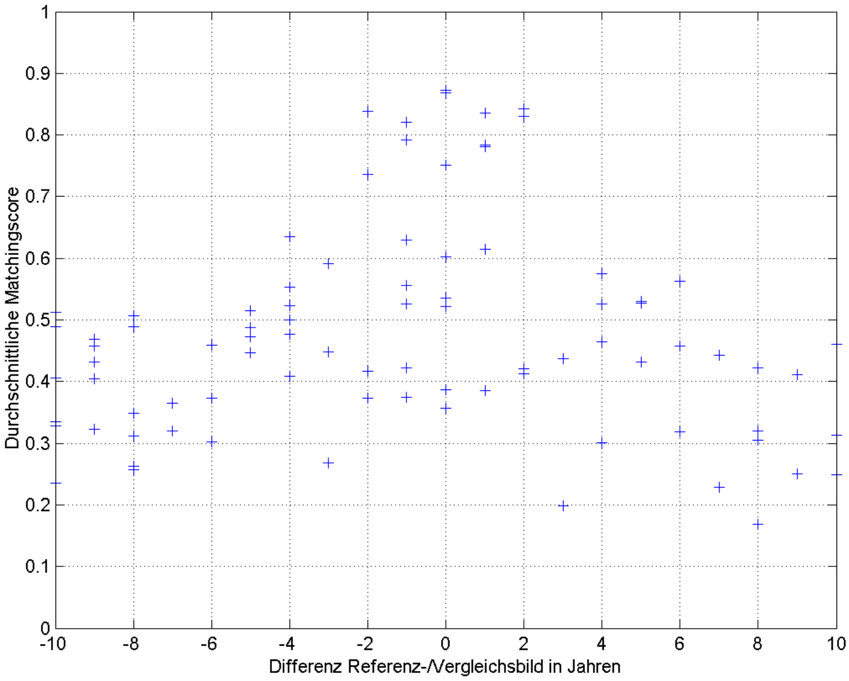


Figure 171: 5-24 Algorithm 2

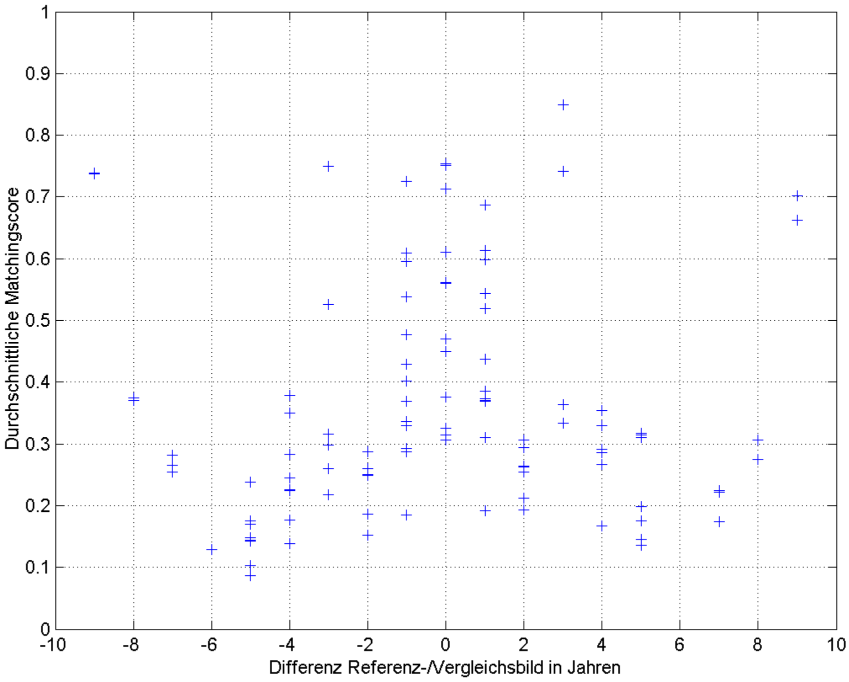


Figure 172: 5-25 Algorithm 2

Durchschnittliche Matchingscore	Average matching score
Differenz Referenz-/ Vergleichsbild in Jahren	Difference between reference and comparison images in years

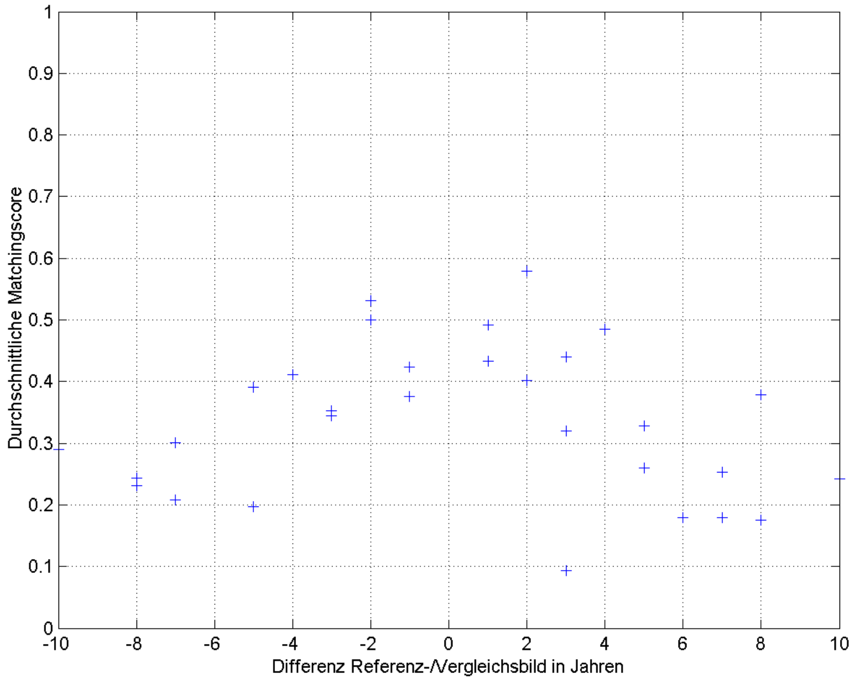


Figure 173: 5-26 Algorithm 2

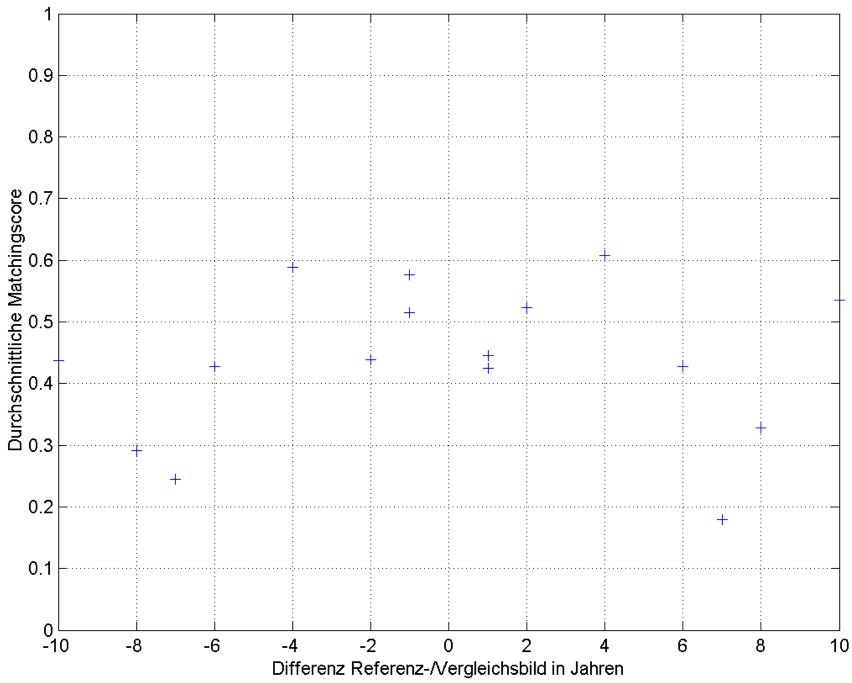


Figure 174: 5-27 Algorithm 2

Durchschnittliche Matchingscore	Average matching score
Differenz Referenz-/ Vergleichsbild in Jahren	Difference between reference and comparison images in years

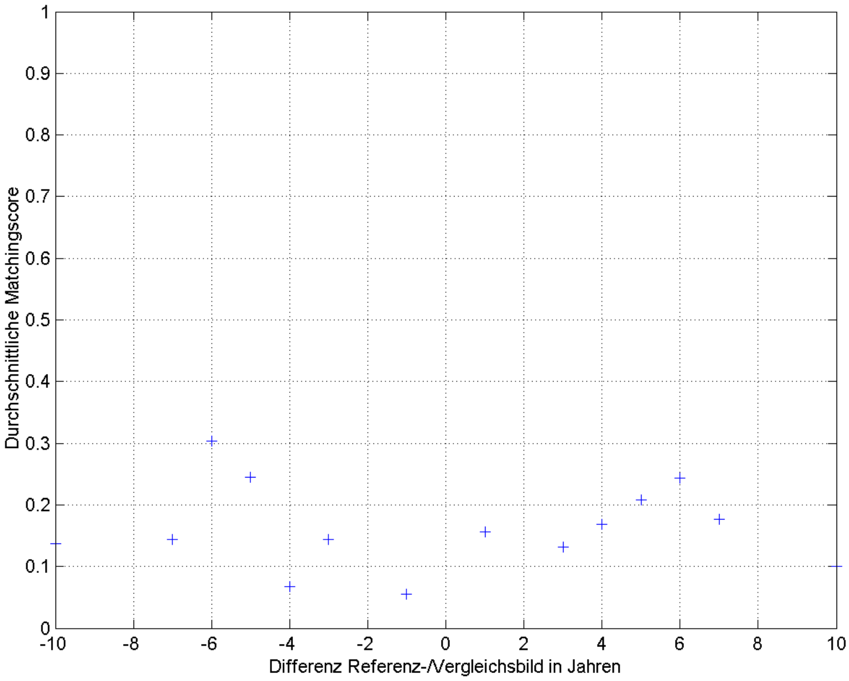


Figure 175: 5-28 Algorithm 2

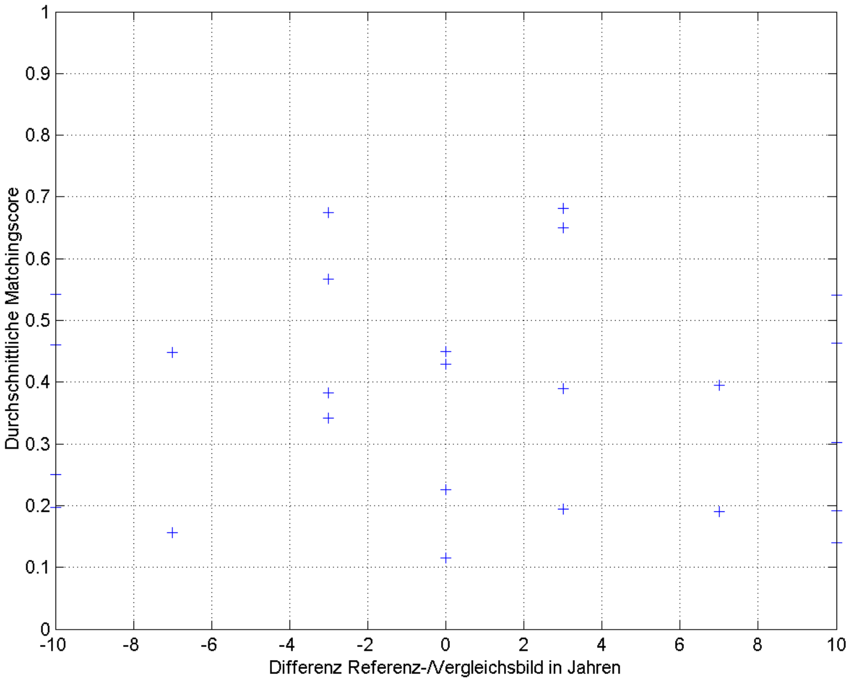


Figure 176: 5-29 Algorithm 2

Durchschnittliche Matchingscore	Average matching score
Differenz Referenz-/ Vergleichsbild in Jahren	Difference between reference and comparison images in years

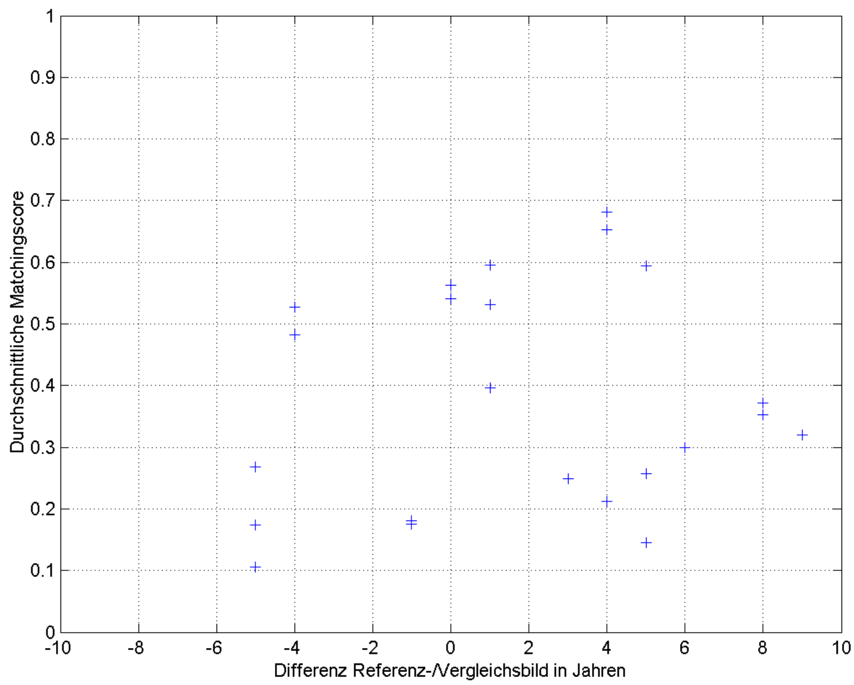


Figure 177: 5-30 Algorithm 2

Durchschnittliche Matchingscore	Average matching score
Differenz Referenz-/ Vergleichsbild in Jahren	Difference between reference and comparison images in years

B.4 Algorithm 3

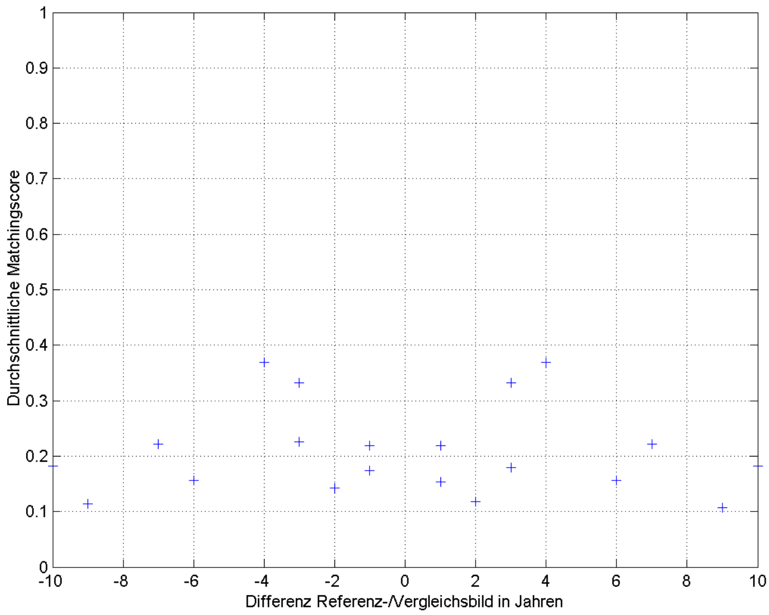


Figure 178: 5-6 Algorithm 3

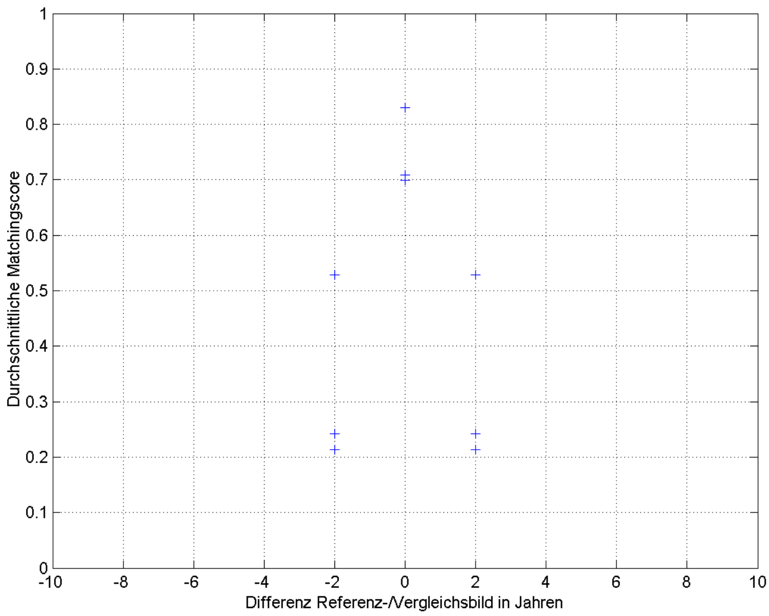


Figure 179: 5-7 Algorithm 3

Durchschnittliche Matchingscore	Average matching score
Differenz Referenz-/ Vergleichsbild in Jahren	Difference between reference and comparison images in years

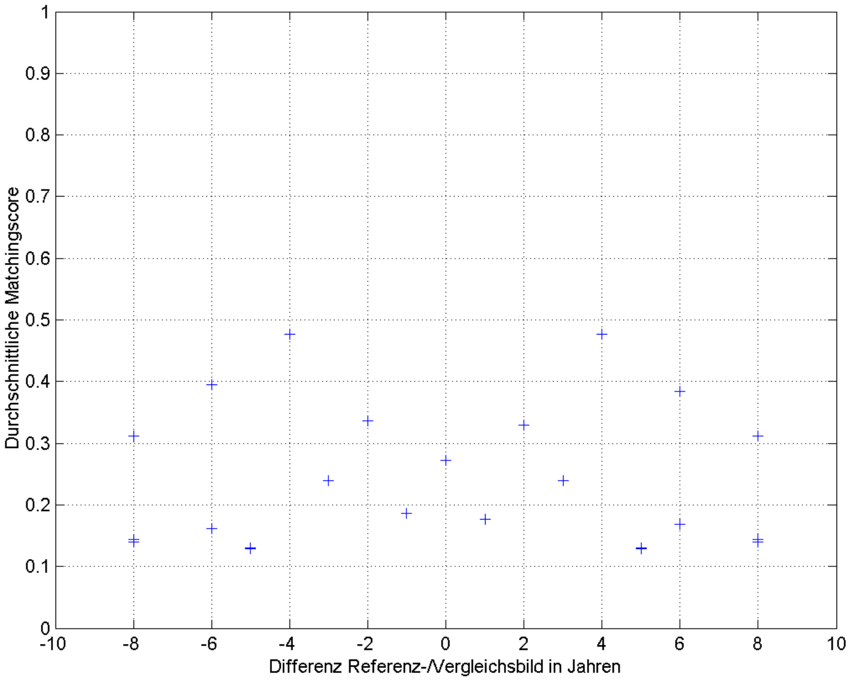


Figure 180: 5-8 Algorithm 3

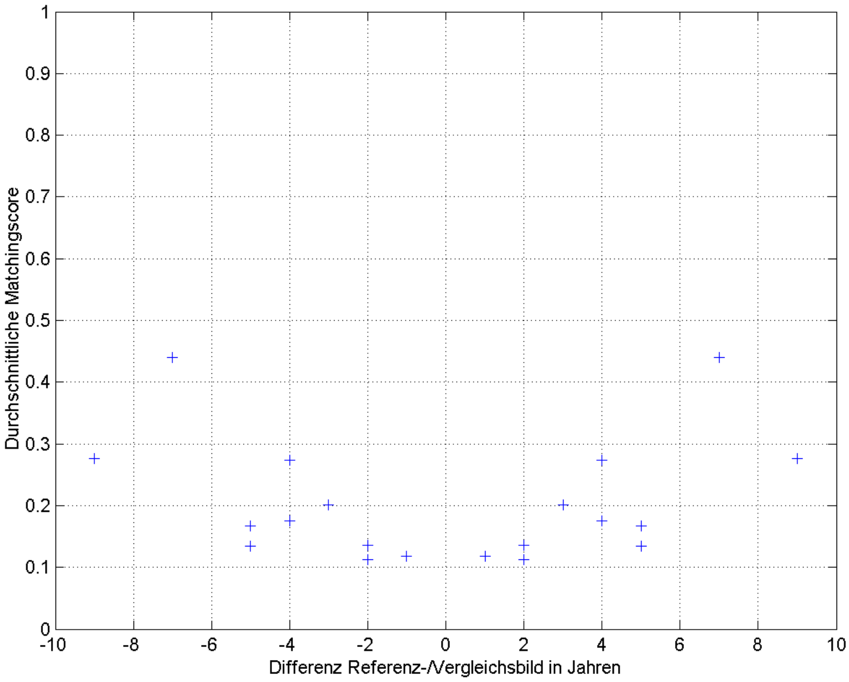


Figure 181: 5-9 Algorithm 3

Durchschnittliche Matchingscore	Average matching score
Differenz Referenz-/ Vergleichsbild in Jahren	Difference between reference and comparison images in years

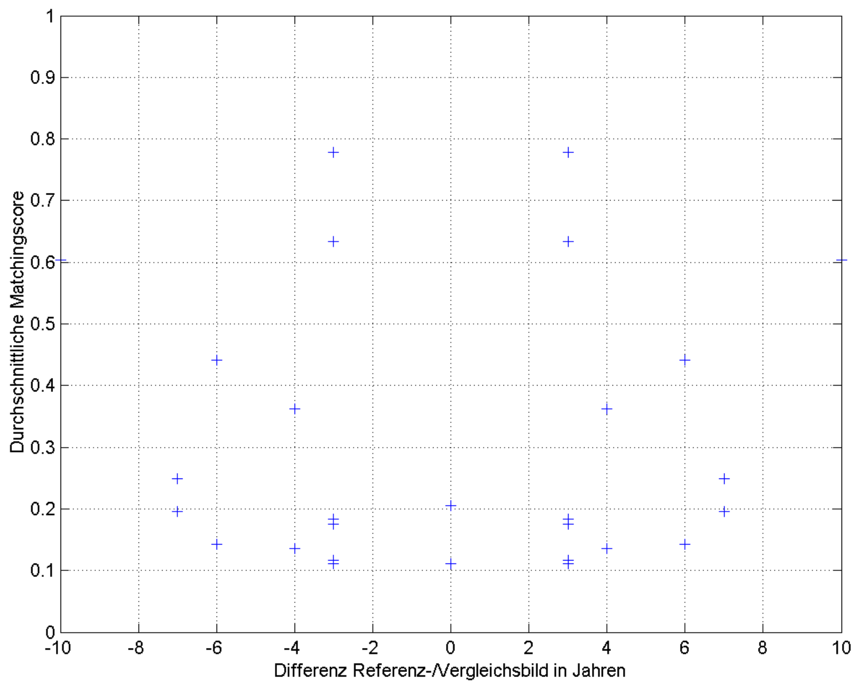


Figure 182: 5-10 Algorithm 3

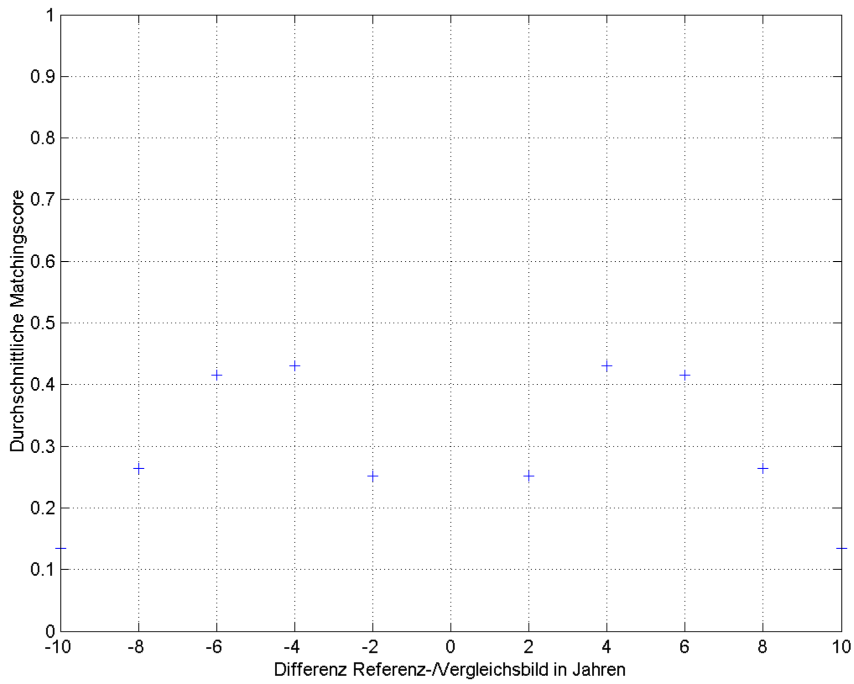


Figure 183: 5-11 Algorithm 3

Durchschnittliche Matchingscore	Average matching score
Differenz Referenz-/ Vergleichsbild in Jahren	Difference between reference and comparison images in years

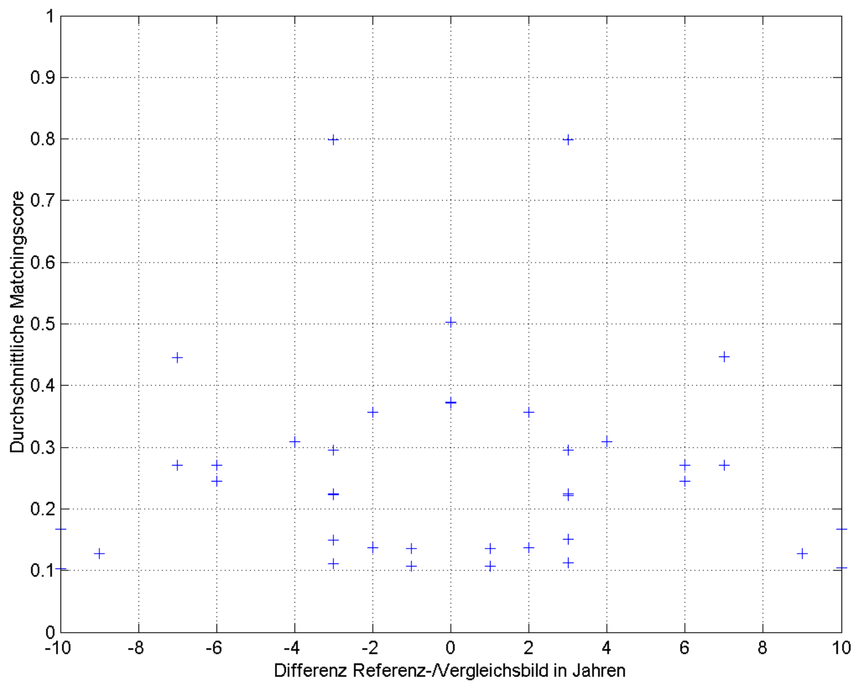


Figure 184: 5-12 Algorithm 3

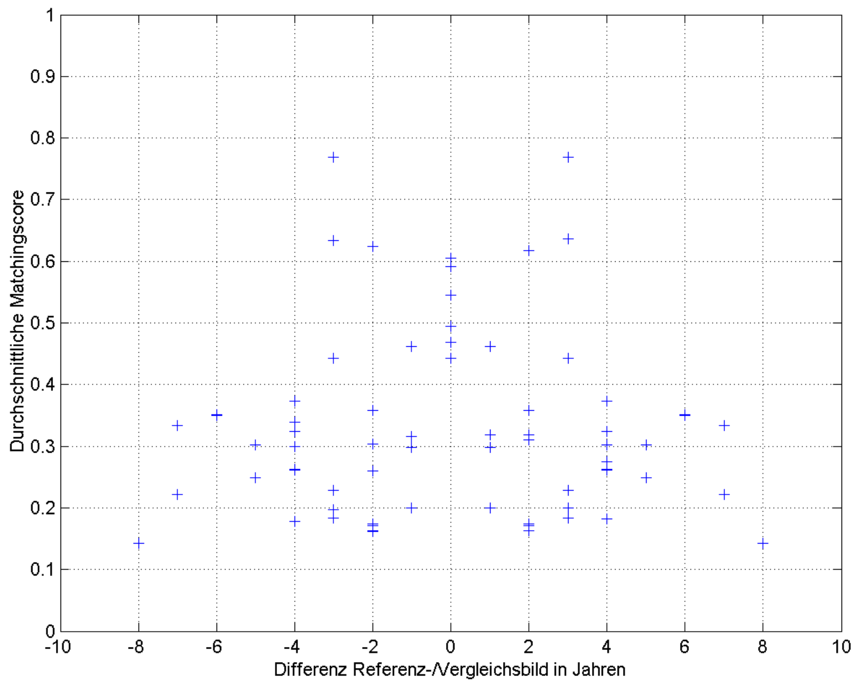


Figure 185: 5-13 Algorithm 3

Durchschnittliche Matchingscore	Average matching score
Differenz Referenz-/ Vergleichsbild in Jahren	Difference between reference and comparison images in years

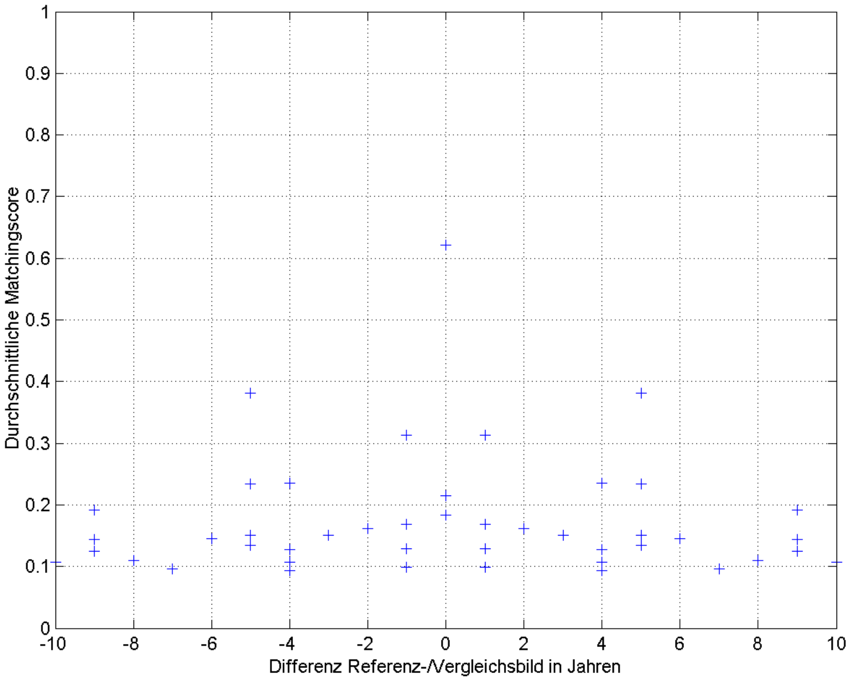


Figure 186: 5-14 Algorithm 3

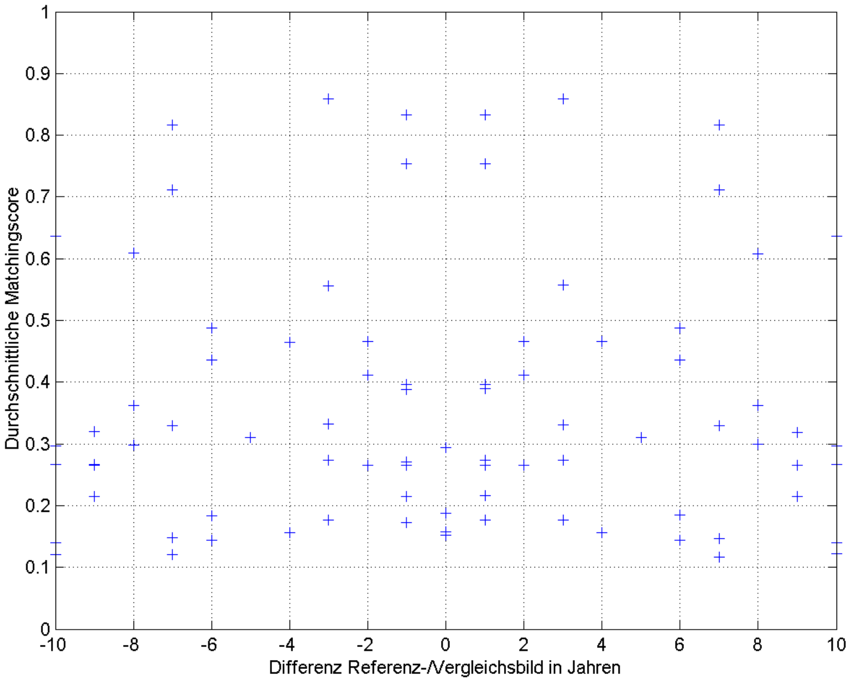


Figure 187: 5-15 Algorithm 3

Durchschnittliche Matchingscore	Average matching score
Differenz Referenz-/ Vergleichsbild in Jahren	Difference between reference and comparison images in years

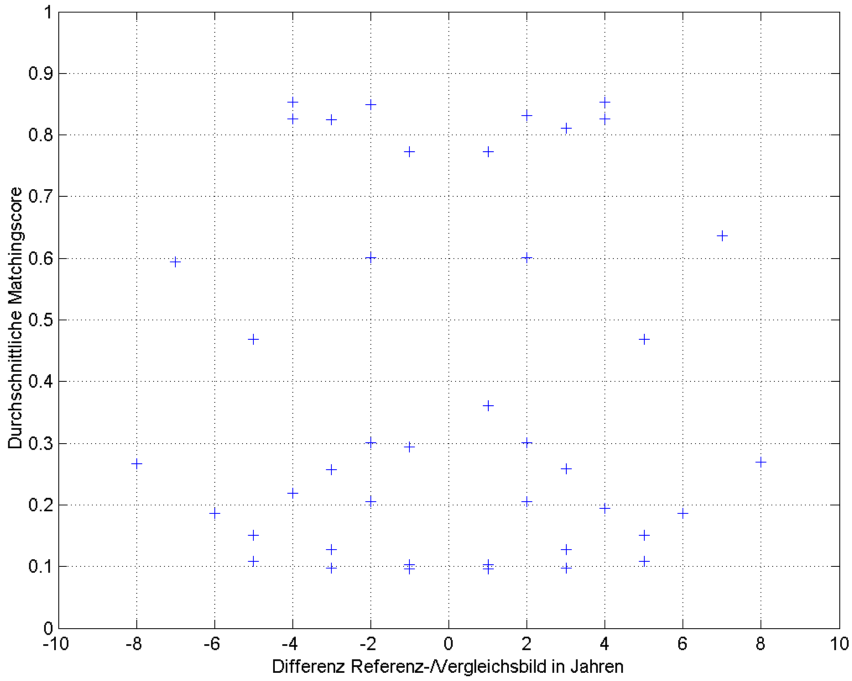


Figure 188: 5-16 Algorithm 3

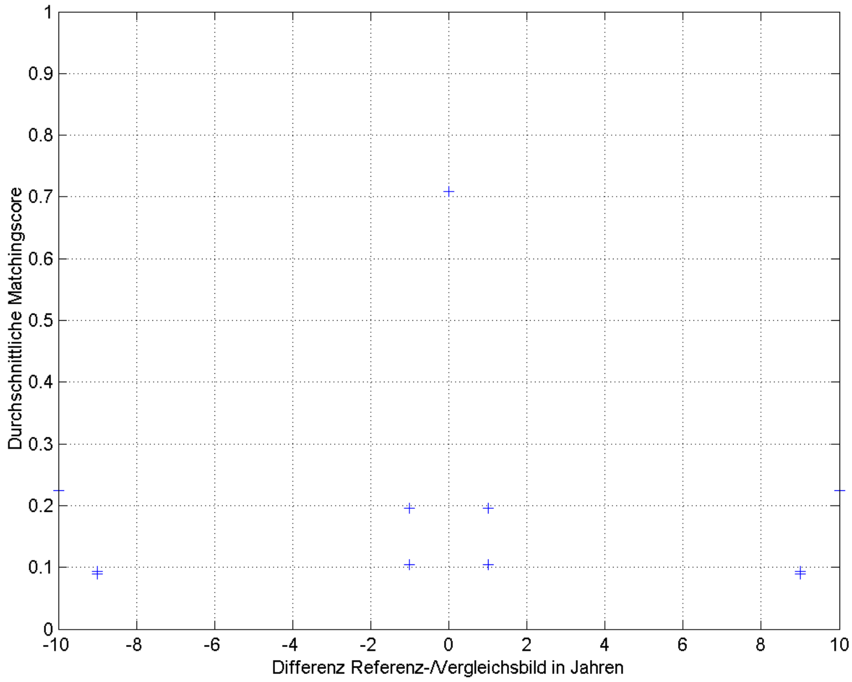


Figure 189: 5-17 Algorithm 3

Durchschnittliche Matchingscore	Average matching score
Differenz Referenz-/ Vergleichsbild in Jahren	Difference between reference and comparison images in years

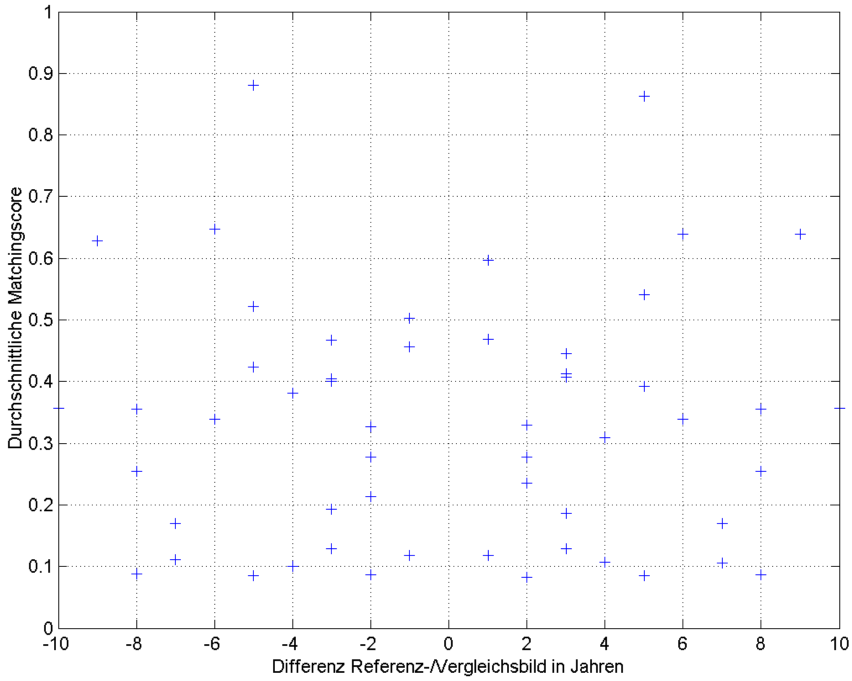


Figure 190: 5-18 Algorithm 3

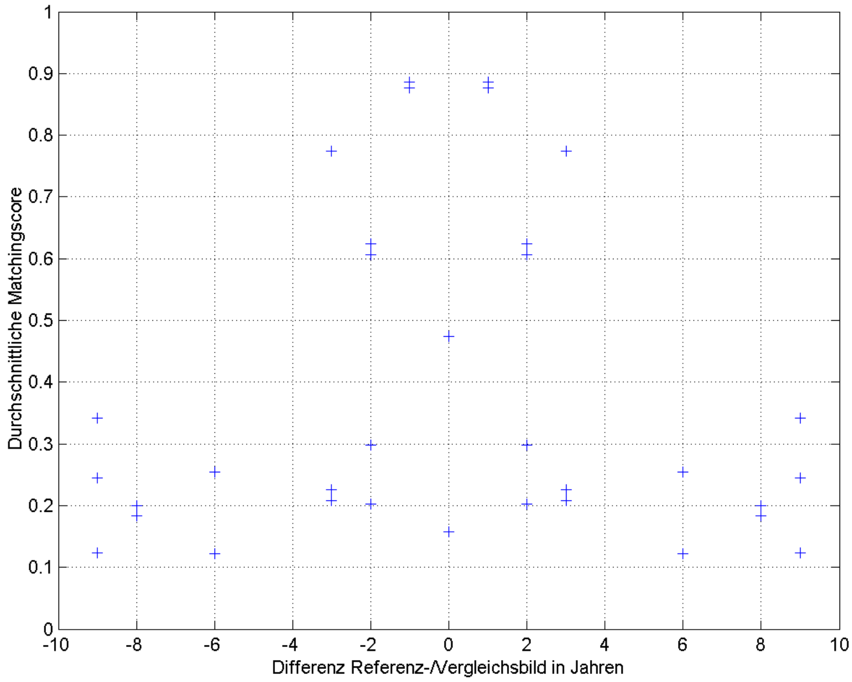


Figure 191: 5-19 Algorithm 3

Durchschnittliche Matchingscore	Average matching score
Differenz Referenz-/ Vergleichsbild in Jahren	Difference between reference and comparison images in years

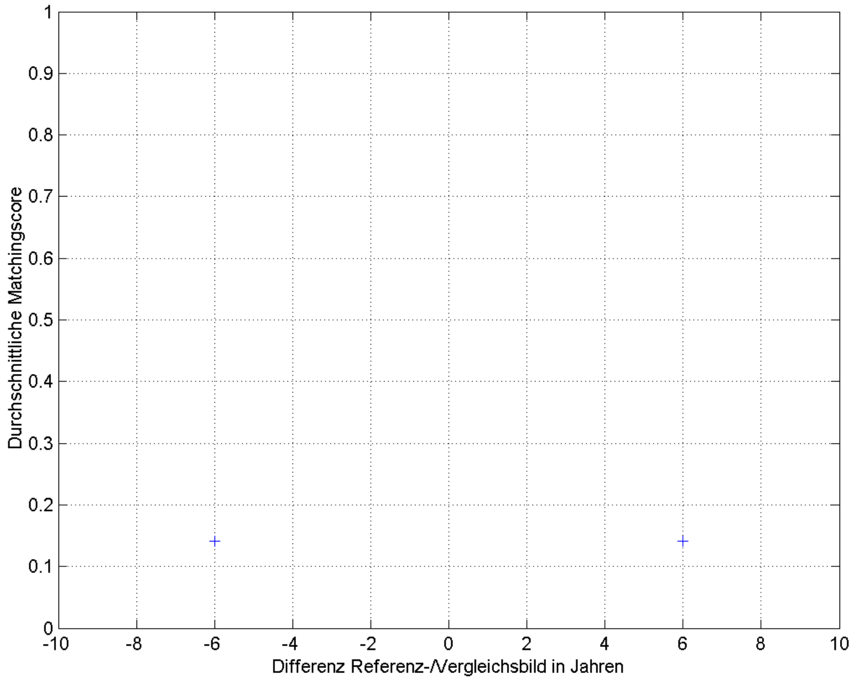


Figure 192: 5-20 Algorithm 3

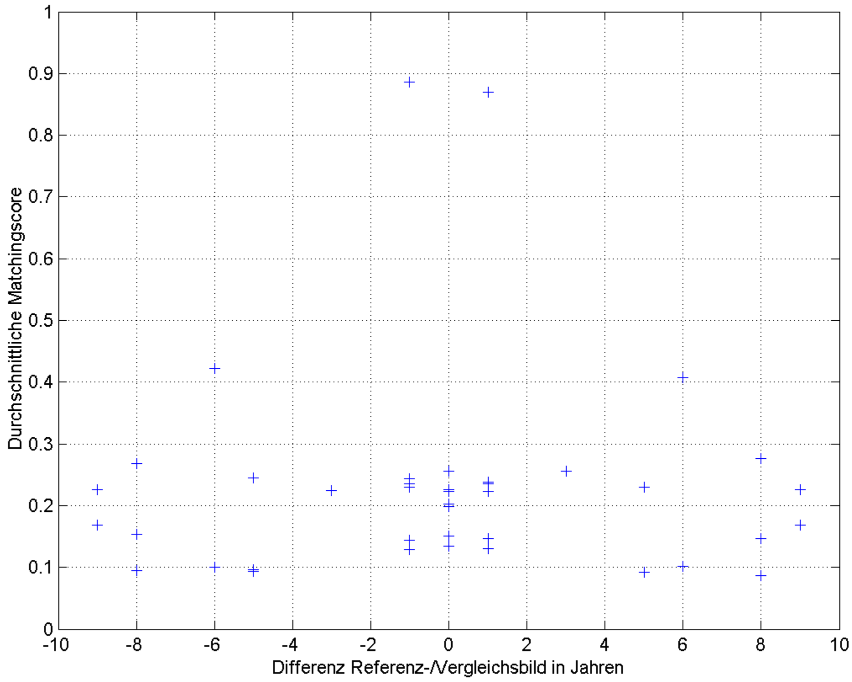


Figure 193: 5-21 Algorithm 3

Durchschnittliche Matchingscore	Average matching score
Differenz Referenz-/ Vergleichsbild in Jahren	Difference between reference and comparison images in years

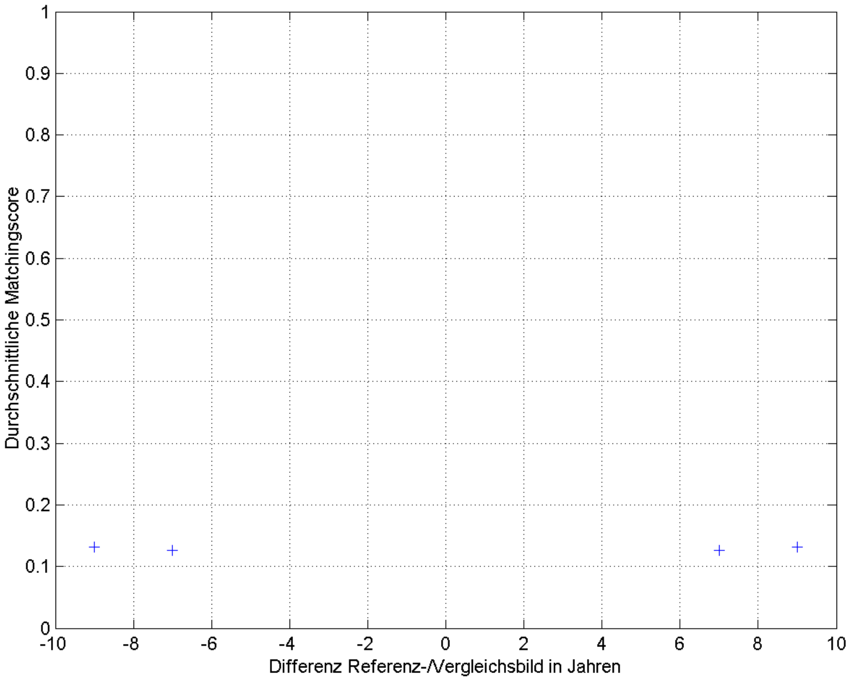


Figure 194: 5-22 Algorithm 3

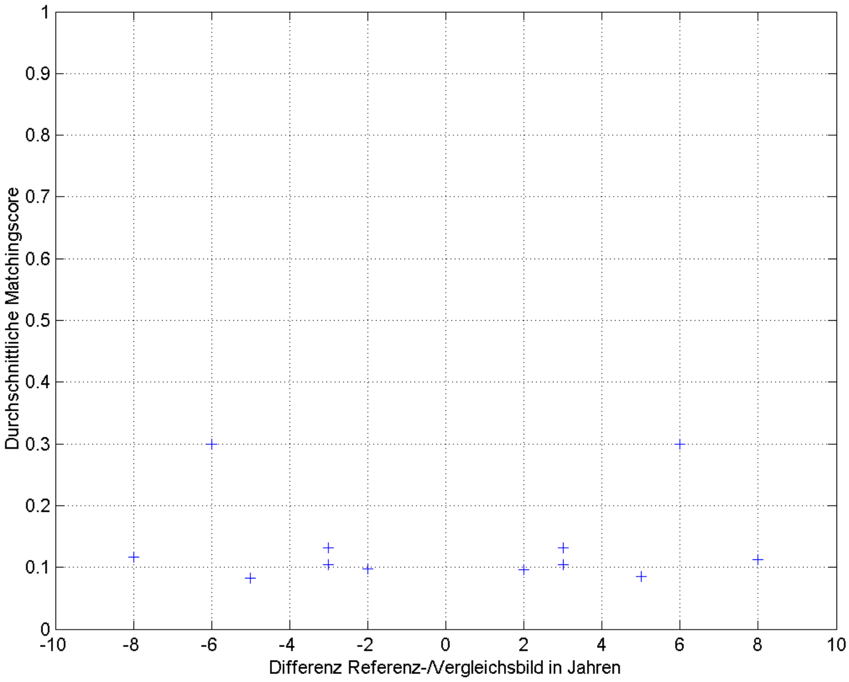


Figure 195: 5-23 Algorithm 3

Durchschnittliche Matchingscore	Average matching score
Differenz Referenz-/ Vergleichsbild in Jahren	Difference between reference and comparison images in years

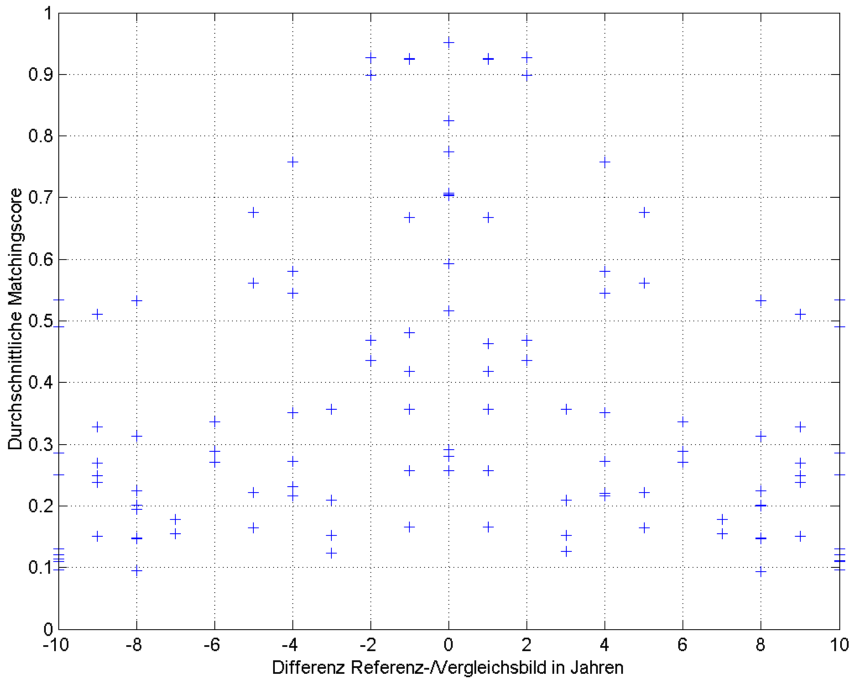


Figure 196: 5-24 Algorithm 3

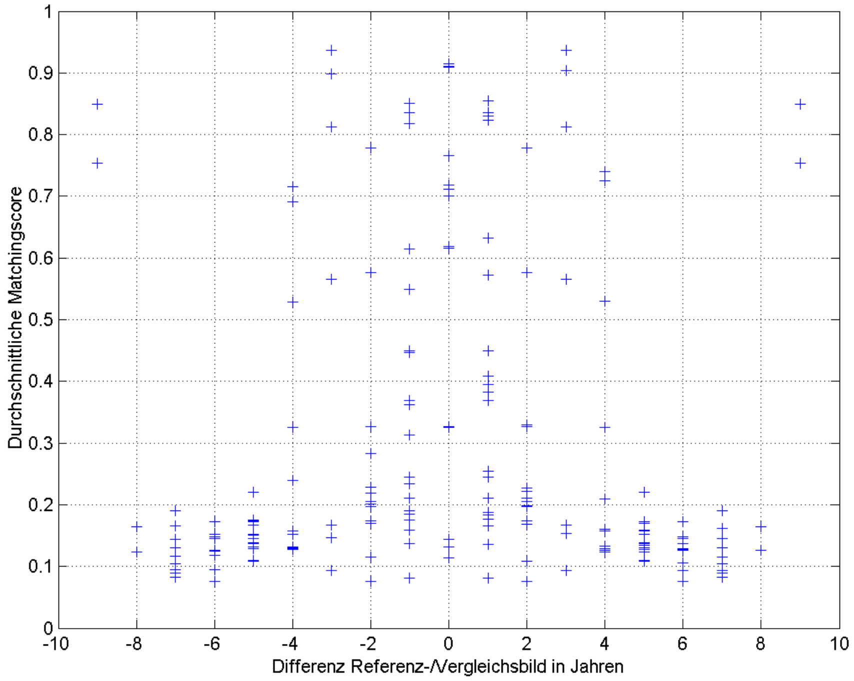


Figure 197: 5-25 Algorithm 3

Durchschnittliche Matchingscore	Average matching score
Differenz Referenz-/ Vergleichsbild in Jahren	Difference between reference and comparison images in years

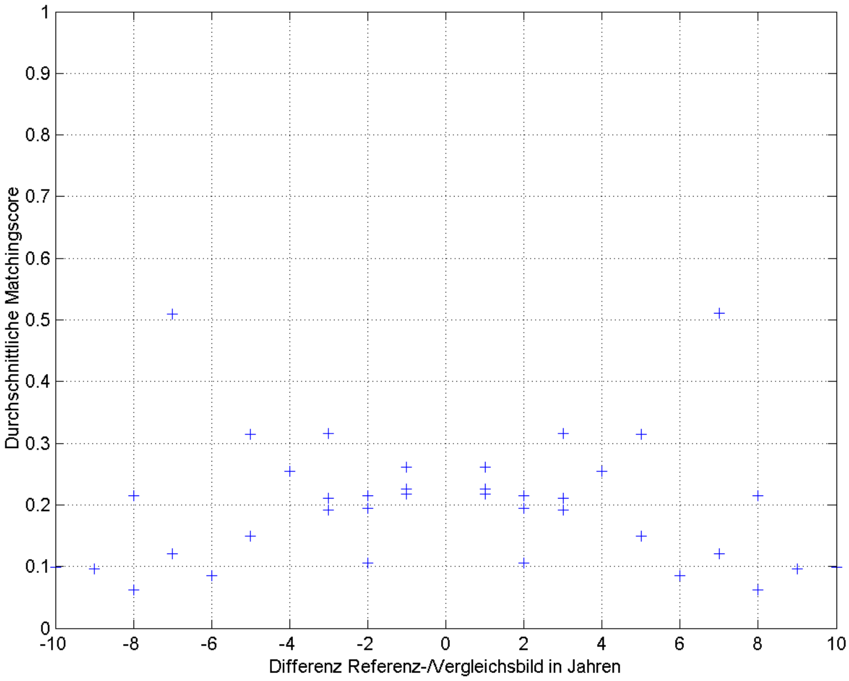


Figure 198: 5-26 Algorithm 3

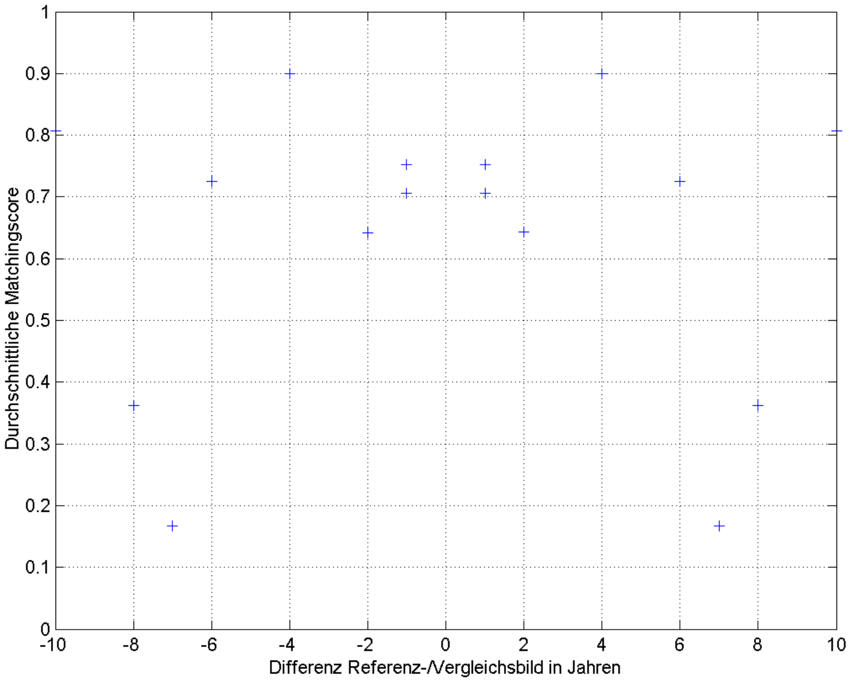


Figure 199: 5-27 Algorithm 3

Durchschnittliche Matchingscore	Average matching score
Differenz Referenz-/ Vergleichsbild in Jahren	Difference between reference and comparison images in years

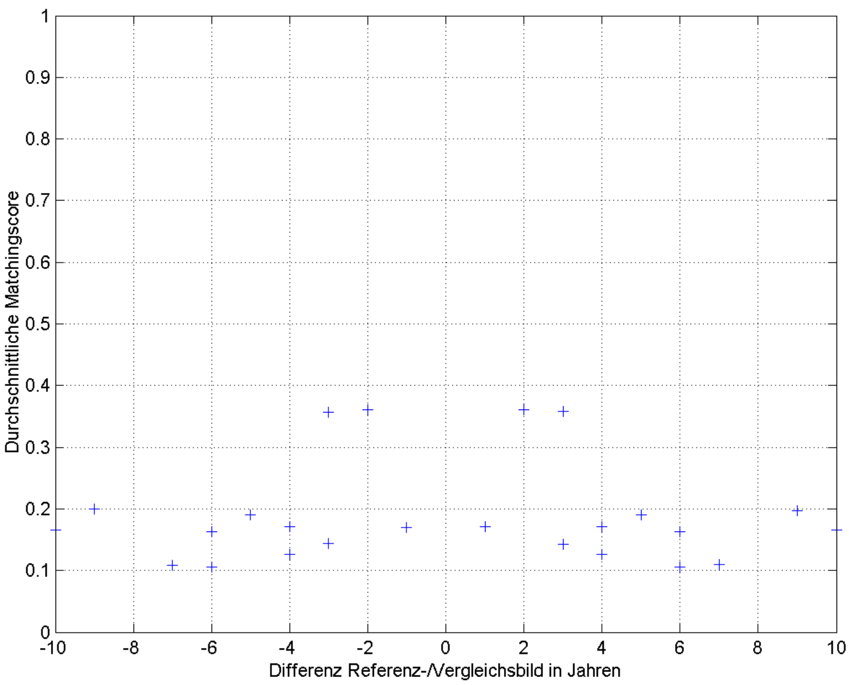


Figure 200: 5-28 Algorithm 3

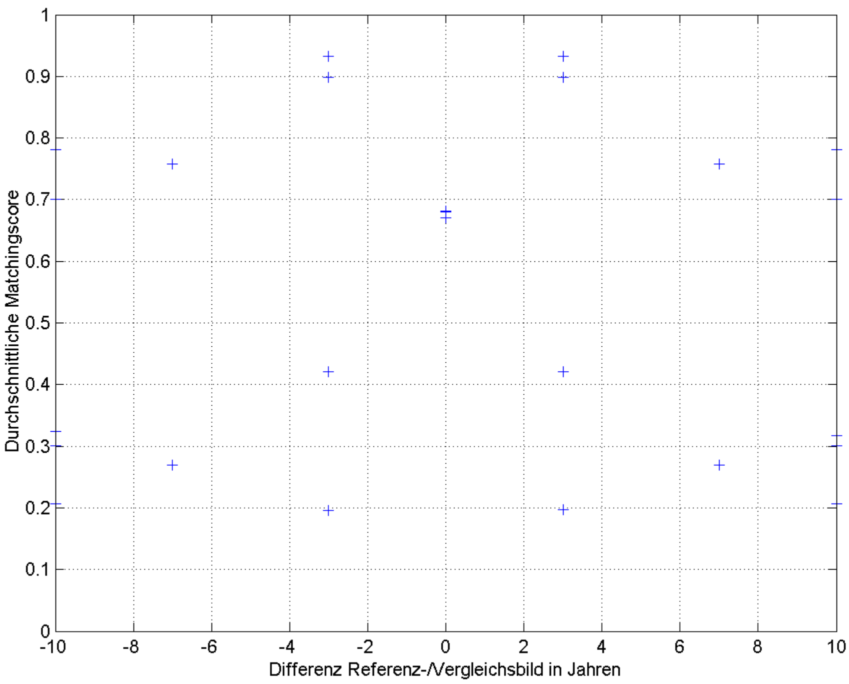


Figure 201: 5-29 Algorithm 3

Durchschnittliche Matchingscore	Average matching score
Differenz Referenz-/ Vergleichsbild in Jahren	Difference between reference and comparison images in years

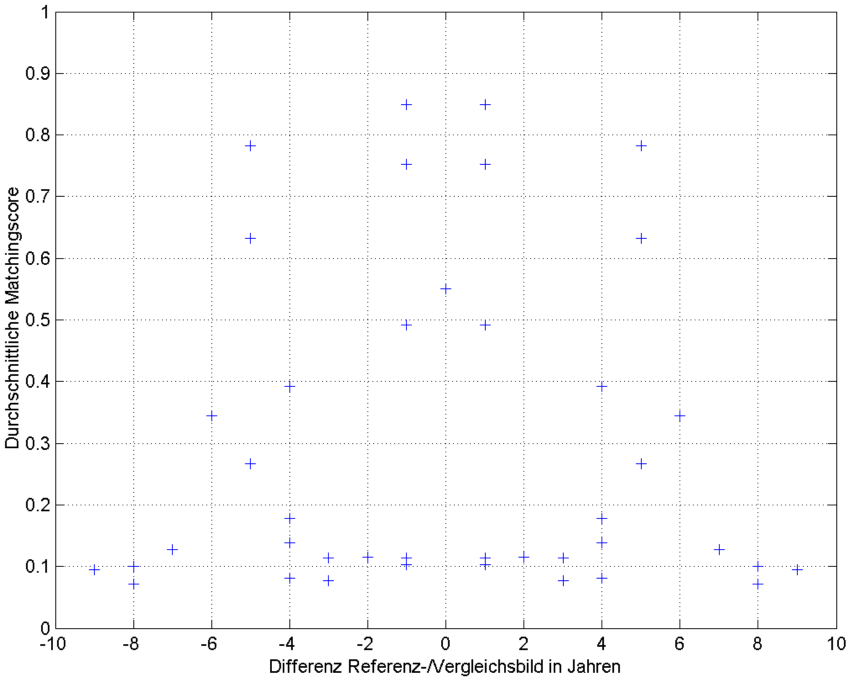


Figure 202: 5-30 Algorithm 3

Durchschnittliche Matchingscore	Average matching score
Differenz Referenz-/ Vergleichsbild in Jahren	Difference between reference and comparison images in years

Annex C Company profile s

The following company profiles and product descriptions were provided by the companies themselves.

C.1 Astro Datensysteme AG

Astro Datensysteme AG has been making its name as an innovative systems house in the areas of computer and security technology for over 14 years. Founded in 1989, it commenced production of UNIX-based SPARC computers under licence to SUN Microsystems Technology in 1991. Quality and closeness to the customer's requirements have always been its top priority, so it is not surprising that Astro Datensysteme was appointed Master Distributor by SunPro© in 1993. In 1995 Astro Datensysteme progressed its strategy of offering the maximum customer friendliness by developing the one-stop shopping idea. Customers should be able to receive from a single supplier everything that they needed, from servers and workstations, peripherals and accessories through to expert advice and technical know-how, instead of being referred from one vendor to the next when they had a question, as was previously frequently the case. True to this motto, in 1996 Astro Datensysteme 1996 began to distribute Hewlett-Packard© and IBM© workstations and concluded teaming agreements with ORACLE© and Informix©.

In 1997, at the same time as the foundation stone was being laid for the Astro Technology Centre in Moosinning, Astro Datensysteme founded its Internet Solution Centre. In 1998 it relocated its headquarters to the now finished complex. That year and the following year it expanded its product range through authorised distribution partnerships with SUN© Microsystems, Microsoft© Solution Provider and SGI© (Silicon Graphics).

In 1998 Astro Datensysteme began research and development in the area of biometric security systems, initially with the focus on fingerprint technology. In 1998, Astro Datensysteme went on to develop "Astro Security Solutions E-Guard" and "CryAstro High-Speed Encryption".

In 2000, its IDS, ProNet and Bio-Laptop biometric solutions were launched on the market. As one of the innovative pioneers in the area of biometric security, Astro Datensysteme was one of the founding partners of the Silicon Trust and entered into a strategic partnership with Infineon Technologies AG. The year 2002 saw the market launch of Astro BioSmartCards, as well as enhancements to IDS and ProNet. Atmel Inc. joined the team of Astro partners, and Astro Datensysteme also concluded a development partnership for biometric solutions with Sun.

In September 2002, Astro Datensysteme extended its research and development activities to the area of facial recognition. Based on the algorithms of two companies whose technology it purchased, AFIS-like software for photographs and passport photos, video surveillance and checkpoint surveillance and various security systems based on facial recognition or combined with other security systems (fingerprints, PIN, smart card) have been developed.

C.2 Cognitec Systems GmbH

Based in Dresden, Cognitec Systems GmbH was founded in May 2002 and acquired all rights to the FaceVACS facial recognition technology from the former Cognitec Vision Software AG. The core team was taken over by the new company in order to continue the development and marketing of the already widely used and world-leading FaceVACS software.

Cognitec's software experts have been developing facial recognition technology and innovative applications based on this since 1995. This software technology is able to recognise persons in video images or on photographs by their characteristic facial features. FaceVACS has developed into one of the leading biometric products on the market. Industry and government customers have been using FaceVACS for physical access control since 1996.

Software for the identification of persons in video imagery is enhancing security in public areas. Border checkpoints based on automatic facial recognition can check the identity of passport holders. Cognitec's software can also be integrated into other software packages or devices, using the FaceVACS software development kit. Software houses all over the world have started developing applications that use automatic official recognition.

C.3 Controlware GmbH

As a leading systems integrator and competent service provider of over 20 years' standing, Controlware stands for expertise in the area of telecommunications.

Controlware provides an all-round service extending from communications solutions, information security and IT management solutions through to storage networking, from consulting to installation, maintenance, operating and training. On top of this, through the Network Control Centre (NCC) Controlware offers network support 24 hours a day. Its range of services includes partial or complete network operations and customer-specific programming of network management solutions.

In this way all of the services offered by Controlware are optimally tailored to individual customer requirements. The company's extensive and modular service programme covers both individual solutions and also extensive major projects.

Controlware GmbH is a global services company that is in private ownership. In the financial year 2002 the company's German workforce numbered over 500. The company headquarters are located in Dietzenbach near Frankfurt am Main. As well as 10 sites in Germany, Controlware also has branches in Belgium, France, the United Kingdom, the Netherlands, Austria, Switzerland, Singapore and the USA.

C.4 ZN Vision Technologies AG

ZN Vision Technologies AG is one of the technologically leading and most innovative companies in the field of security technology. For 10 years, ZN has been developing and marketing biometric security solutions using face recogition. ZN offers a broad range of products including standard systems, customer-specific solutions and consulting services for all areas of biometric applications, notably border management, identification documents, airport and building security, forensics, national ID management systems, intelligent video surveillance and privacy enhancing technologies (PET). Both with regard to technology and economy, ZN's solutions have proven themselves among over 170 customers world-wide.

ZN was originally founded in 1992 as a spin-off of the Ruhr University Bochum and the University of Southern California, Los Angeles, by leading researchers in the area of artificial vision. In April 2003 the company announced that it was merging with Viisage Technology (NASDAQ: VISG), to form the biggest facial recognition company and one of the biggest corporate players in biometrics in the world.

ZN Vision Technologies AG supplies standard systems and customer-specific solutions for the automatic identification and verification of persons using state-of-the-art facial recognition techniques. The ZN-Face®, ZN-Phantomas® and ZN-SmartEye® systems developed by ZN constitute the basic modules in applications such as physical access control, border control, police and government investigations and intelligent video surveillance. ZN AG's impressively sized clientele includes high-security areas such as nuclear power stations and computer centres, the Berlin airports, large companies like Microsoft and Siemens, leisure amenities such as Hanover Zoo and fitness studios, plus a number of police authorities both in Germany and abroad.

ZN's systems are based on ZN's specially developed and patented Hierarchical Graph Matching method, an approach to facial recognition that is derived from human vision. Approx. 1,700 features are used to characterise the face. Thus the ZN systems produce superior recognition results even under difficult environmental conditions and changes in appearance, for example, due to ageing, beards, wigs or spectacles. The ZN facial recognition technology has won world-wide acclaim, as recognised in numerous prizes such as the Innovationspreis der Deutschen Wirtschaft (Deutsche Wirtschaft innovation prize) (1996), the Karl-Heinz-Beckurts Preis (Karl-Heinz-Beckurts prize) (1998) and the Körber Preis für die Europäische Wissenschaft (Körber Prize for European Science) (2000). Recently it won the "Product of the Year Award 2002" organised by the American magazine BiometriTech for the ZN-Face system.

Annex D Product profiles

D.1 Astro Datensysteme AG

D.1.1 Algorithm description

METHOD OF THE COMPARISON AND PERSONAL IDENTIFICATION BY IMAGE OF THE FACE

Theoretical motivation of the method

Described technology is founded on using the method of principle components or procedure of the data compression Karunen-Loev. Is it theoretically proved, that Karunen-Loev transformation is the most optimum by degree of the compression in space of squarely-summarized functions among all orthogonal transformations: Furie, Uolsh-Adamar and others. This allows effectively and portably to convert the two-dimensional image to vector, which components present the most important, "main" components. Its length much less then original size of the image, that enables effectively to reveal the fine particularities of analyzed data, as well as vastly accelerates and simplifies the further procedure of the comparison of images.

As measures of the comparison of vectors is used selective correlation, which is statistical motivated and significant measure to dependencies of the random quantities of most often applicable for decision of the many practical problems.

The Method consists of the following stage:

1. Coding and rationing of the images

At this stage procedures are intended for executing rationing: searching for the person and transformation of the current image to image with given sizes and determined by position of the face on him. For this is produced automatic consequent searching at first for head on image, then eyes and next pupils of eyes. The coordinates of pupils of eyes are used for performing operation of turning, shift and scaling at time of transformation of the image. On got image is present only directly image of the face without background forming original image: clothes, hairstyle, surroundings background and etc. The used algorithms are founded on using the methods of linear filtering, digital processing of signal, categorizations of images and filters of Kalman.

Possible processing of images, comes from movie camera, digital camera, and from computer database. Received images can be used hereinafter for identification of the person or shaping the base of persons.

2. Making the base of the compared persons

First procedure of using the method main of general components is shaping of ortonormal basis. For example, is it let given certain extensive choose out of F normalized images of the persons after performing the stage1. It is started procedure of iterated selection of most variable faces from this choose out. The image from ensemble of F in 2D digital type presents itself matrix A numbers, expressing brightness in determined point (the pixel) of the image, size N X M, where N - height of the image, M - its width.

$$A = \begin{pmatrix} a11 & a12 & a1M \\ a21 & a22 & a2M \\ aN1 & aN2 & aNM \end{pmatrix} \qquad (1)$$

The Matrix A presenting in the manner of vector f lengths L = N * M by way of the consequent sensing on columns.

Then choose out of faces F length K possible to show in the manner of matrix

$$F = \begin{pmatrix} f11 & f12 & f1K \\ f21 & f22 & f2K \\ fL1 & fL2 & FLK \end{pmatrix} \qquad (2)$$

in which each column presents the separate face from multitude of F.

Usually, matrix F is an array of numbers, which have a high degree of correlation between itself. So for the reason of reduction of redundancy of information, it is used procedure of Karunen-Loev or method of principal components. At the beginning of this procedure, it is computed covariance matrix

$$D = F * F' \qquad (3),$$

where F' - matrix, transponed to F.

Matrix of covariation D introduces in the manner of

$$D = \begin{pmatrix} 1 & d12 & d1K \\ d21 & 1 & d2K \\ dK1 & dK2 & 1 \end{pmatrix} \qquad (4)$$

for which are computed own numbers and vector. For each of they are executed following equality

$$D * zi = li * zi \qquad (5),$$

where li eigin number, zi eigin vector.

The Matrix D is symmetrical and positively determined that provides stability of the calculation process and that own numbers will be material. We shall Produce sorting own vector on decrease the module own number and shall hereinafter use only first NS of them.

It is known that Eigen vectors {z i} forms orthonormalized basis, on which it is possible to distribute any vector of the length L. In our case this means that any normalized image of the face is possible to present in the manner of linear combination of the eigen vectors. Herewith first eigen vector is average arithmetical of all faces from ensemble F, but next vectors present some variation from it in order of decreasing module of their coefficients of decomposition.

Thereby each image of the person with certain inaccuracy representable in the manner of vector, definied from equality

$$b=Z' * x \qquad (6),$$

where b - a vector of the decomposition in orthogonal basis, Z - a matrix, columns of it are an eigen vectors, x - a vector of the image. While size of the vector b greatly less then size x that means the significant degree of the compression of information at presentation of the image. As a result of this we got vectors, on which is produced the most further comparison. After showing all vectors from F, it is produced their comparison with "average" face from F by means of selective correlation and determination of most different persons from it. Hereinafter this procedure iteratively is repeated, as a result of that happens multitude of F1 of most variable faces is got from first choose out F. From multitude F1 by way of the repeated using of the decomposition Karunen-Loev is got new orthonormal basis for comparison of the faces. Now let G be certain multitude of the faces, on which will be produced searching for of the similar faces.

Each image from G shall be presented in the manner of vector on formula (6).Thereby, there are received multitude standard faces of multitude G and orthonormal basis, intended for search of the similar faces.

3. Searching for of the similar persons.

Let b - image of face, given in the manner of decomposition coefficients vector, got from (6). The known multitude of standard vectors C = {ci}. At searching can be put 2 different tasks: verification and identification. Verification is search "head-to-head", which means determination of degree comparing face with face from the database. The Identification search process for "one to many", which means determination of

multitude of "similar" persons from the base.

As measures of "similarity" is vector in linear space, formed by basis {zi} determined as selective correlation. Normalized correlation of two vectors a and b is defined by formula

$$r = \frac{a * b}{||a||_{L2} * ||b||_{L2}} \qquad (7),$$

$||a||_{L2}$ norm in space L2.

Significance of the correlation are found in interval from -1 to 1 and if more is significance of the correlation, then more is statistically depends are the compared vectors. The Threshold, by which are filtered "similar" vectors are floating depending on type of the solving task and used algorithm.

D.1.2 Product description

Astro GateKeeper

Astro GateKeeper is a physical access control system based on facial recognition. A camera (webcam, video camera) installed on a door records the face of the person seeking access and passes it to the connected computer (workstation, desktop or laptop), which uses the installed software to compare the face with a predefined database of persons cleared for access. If the face recorded matches along the predefined parameters, the connected door lock is released and access is granted.

The advantages of physical access control based on facial recognition are manifold.

The biometric key, in this case the face, and the person cleared for access are inseparably linked together. Non-biometric keys can get lost, be forgotten or even stolen. By contrast, physical access control governed by facial recognition is more convenient, more secure and less expensive to the operator.

Compared with most other physical access control systems, both biometric and ones operated with PIN or password entry, magnetic card or smart card (apart from radio smart card), the person whose identity is being checked does not have to actively prove his identity. Ideally, he is not even aware of the physical access control system. His face is recorded, compared and verified while he is still approaching the door. The other access control systems have to be operated actively (by entering a code, placing a finger over a scanner, looking into a special camera etc.), which takes longer and introduces sources of error.

The technology that we use to verify or identify a person on the basis of an image is based on an algorithm involving comparisons and re-recognition of images of the face. The vector-based algorithm analyses the primary features of face, consisting of those elements in the description of a human face that differ most frequently.

For this purpose, the system records every movement digitally by camera. Using special algorithms for recognising the map of a human face, it marks and finds the exact position of the pupils, and measures the face from that starting point.

The map of the face is then converted to the individual main features. The data thus generated is around 300 bytes, allowing simple integration of identification systems even in large data quantities.

The system is capable of learning and is trained by checking a searched for set of images using statistical yardsticks.

Test results

In January 2003, the Astro facial recognition system (FRS) tests were completed.

A database containing 454,764 faces served as the basis of the test.

1,000 randomly selected images had to be recognised by the system. The test images were a series of photographs taken over a period of seven years and under quite different lighting conditions and images taken from driving licences.

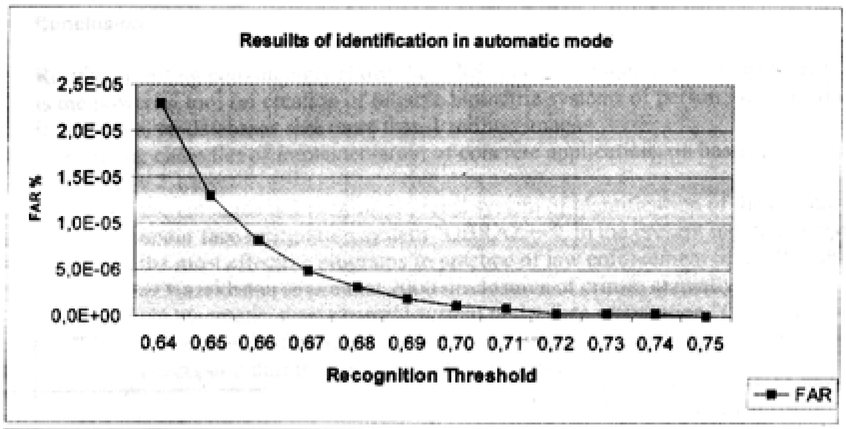
The test computer was an industry standard Pentium IV – 2.8 GHz, 512MB RAM, 80GB HDD with Microsoft Windows 2000 Professional.

Results

Technical specifications:

	Parameters		Value
1	Maximum possible volume of the database		> 1 000 000 images
	One to One (Verification)		
2	Time of search		<1 sec.
3	Accuracy of search (Recognition Threshold = 0.65)	False Accept Rate	1.3E-05%
4		False Reject Rate	> 34.5%
	Identification in automatic mode		
5	Identification Rate		800 000 images/ sec.
6	Accuracy of search (Recognition Threshold = 0.65)	False Reject Rate	34.5%
7		False Accept Rate	1.3E-05%
	Identification in expert mode		
8	Identification Rate		800 000 images/ sec
9	Accuracy of search	Probability of eigen detection on the 1st place	> 84.8%
10		Probability of eigen detection in first 10 images of the recommendatory list	> 92.1%
11		Probability of eigen detection in first 20 images of the recommendatory list	> 93.7%
12		Probability of eigen detection in first 50 images of the recommendatory list	> 95.2%
13		False Reject Rate	< 4.8 %

Specifications are reached by using Pentium IV / 2.8 GHz / RAM 512 Mb; Windows 2000. Test database 454 764 images.



Recognition Threshold	0,64	0,65	0,66	0,67	0,68	0,69	0,70	0,71	0,72	0,73	0,74	0,75
False Accept Rate (FAR) %	2,3E-05	1,3E-05	8,1E-06	4,8E-06	3,2E-06	1,9E-06	1,1E-06	8,1E-07	2,7E-07	2,7E-07	2,7E-07	0,0
False Reject Rate (FRR) %	31,7	34,5	37,3	39,1	41,7	45,4	48,7	52,1	55,8	59,1	62,9	66,0

D.2 Cognitec Systems GmbH

The technology of FaceVACS

FaceVACS processes each image (8 bit greyscale) as follows:

Face localization: The image is analyzed to determine the position and size of one or more faces. (In all of the following steps, we assume that only a single face is found.)

Eye localization: The positions of the eye centers in the face are determined.

Image Quality Check: The quality of the face image is checked to see whether it is sufficient for the steps that follow.

Normalization: The face is extracted from the image and is scaled and rotated such that an image of fixed size results, with the eye centers at fixed positions in that image.

Preprocessing: The normalized image is preprocessed with standard techniques such as histogram equalization, intensity normalization, and others.

Feature extraction: In the preprocessed image, features are extracted that are relevant for distinguishing one person from another.

Construction of the reference set: During enrollment the facial features of (usually) several images of a person are extracted and combined into a reference set, also called the "biometric template".

Comparison: For verification, the set of extracted features is compared with the reference set of the person who the person in the image just processed is claimed to be; for identification, the feature set is compared to all stored reference sets, and the person with the largest comparison value is selected; in both cases the recognition is considered successful if the (largest) comparison value - which is interpreted as a similarity value - exceeds a certain threshold value.

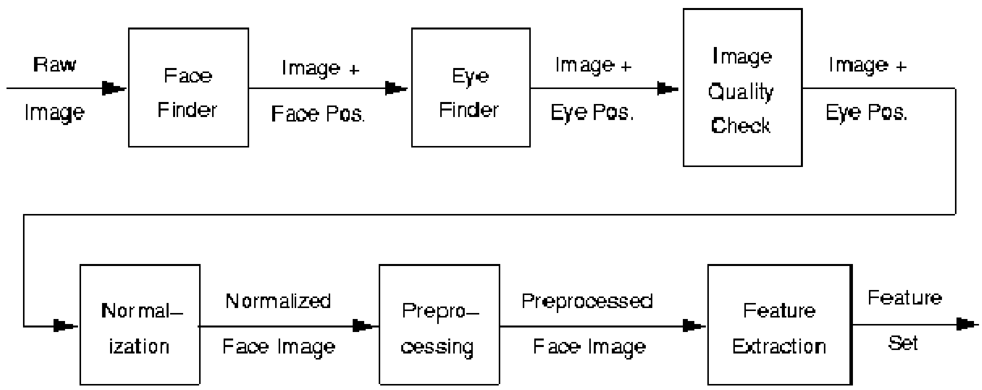


Figure 1. FaceVACS architecture: Feature set creation

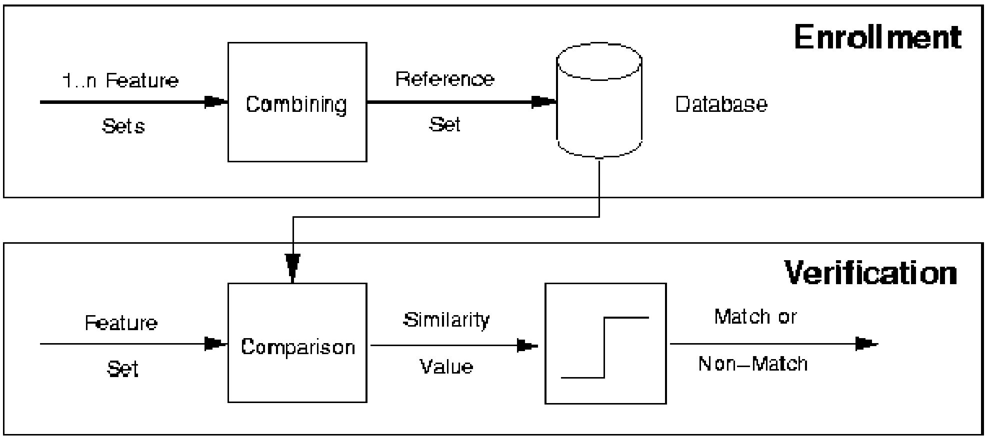


Figure 2. FaceVACS architecture: enrollment and verification

In addition, FaceVACS has a "live check" to ensure that the face in front of the camera is a real one and not just a photograph. To this end, the changes in appearance occurring during movement of the face (rotations around the vertical axis in particular) are exploited. Due to the special 3D structure of a real face, those changes are quite different for a real face than for a photo. So when the user wa nts to pass the live check, he or she should briefly rotate his or her head back and forth.

In the following subsections, more details of the individual steps are given. An example image is used to illustrate the effect of each processing stage.

Figure 3. Example image

1. Face and eye localization

To locate the face, a so-called image pyramid is formed from the original image. An image pyramid is a set of copies of the original image at different scales, thus representing a set of different resolutions. A mask is moved pixelwise over each image in the pyramid, and at each position the image section under the mask is passed to a function that assesses the similarity of the image section to a face. If the similarity value is high enough, the presence of a face at that position and resolution is assumed. From that position and resolution, the position and size of the face in the original image can be calculated. From the position of the face, a first estimate of the eye positions can be derived. In a neighborhood around these estimated positions, a search for the exact eye positions is started. This search is very similar to the search for the face position, the main difference being that the resolution of the images in the pyramid is higher than the resolution at which the face was found before. The positions yielding the highest similarity values are taken as final estimates of the eye positions.

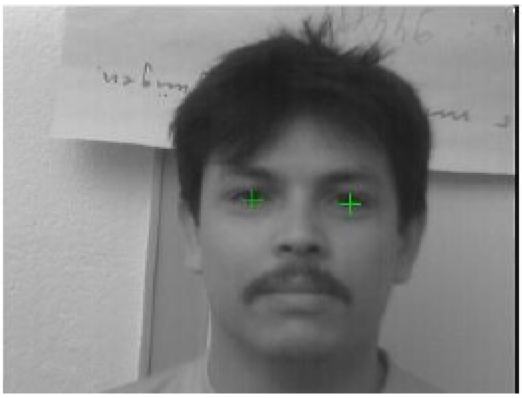


Figure 4. Eye locations found by the algorithm

2. Image Quality Check

To be usable for the subsequent steps, the part of the image occupied by the face has to meet certain quality requirements; e.g., it should not be too noisy or blurred. The quality is measured by means of a set of functions that are applied to the image. If the quality is considered too low, the image is rejected.

3. Normalization and preprocessing

In the normalization step, the face is extracted, rotated and scaled such that the centers of the eyes lie at predefined positions. More precisely, they are positioned to lie on the same horizontal pixel row such that the midpoint of this row is aligned with the midpoint between the centers of the eyes.

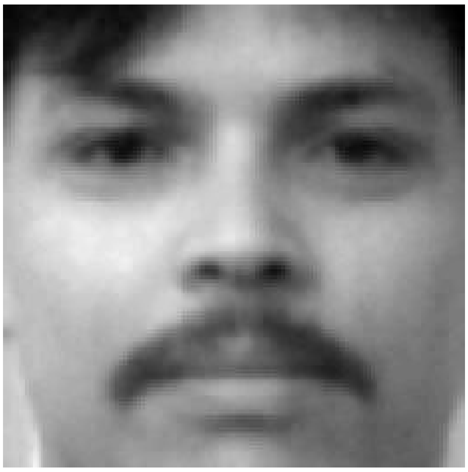


Figure 5. After normalization

The preprocessing step comprises, among other transformations, the elimination of very high and very low spatial frequencies and the normalization of contrast.

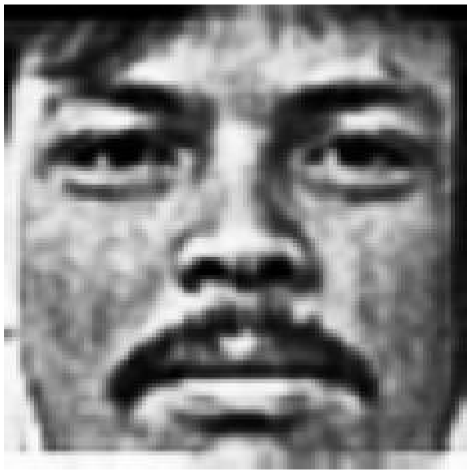


Figure 6. After preprocessing

4. Feature extraction and reference set creation and comparison

Feature extraction starts with local image transforms that are applied at fixed image locations. These transforms capture local information relevant for distinguishing people, e.g. the amplitudes at certain spatial frequencies in a local area. The results are collected in a vector.

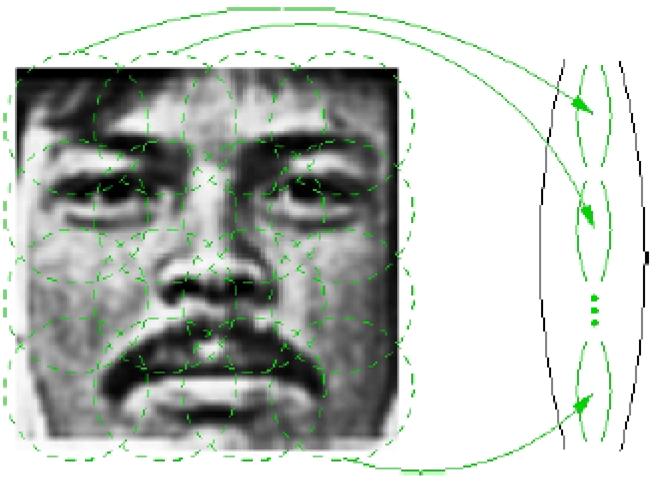


Figure 7. Extracting local features

A global transform is then applied to this vector. Using a large face-image database, the parameters of this transform were chosen to maximize the ratio of the inter-person variance to the intra-person variance in the space of the transformed vectors; i.e., the distances between vectors corresponding to images of different persons should be large compared to distances between vectors corresponding to images of the same person. The result of this transform is another vector that represents the feature set of the processed face image.

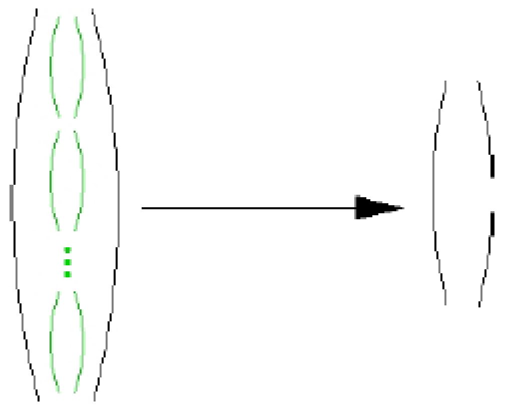


Figure 8. Global transform, yielding the feature set of the face image

For the creation of the reference set, several images are usually taken of each person during enrollment in order to better cover the range of possible appearances of that person's face. The reference set generated for a person consists of up to five feature sets, which are the centers of clusters obtained through a clustering process on the feature sets created from those images.

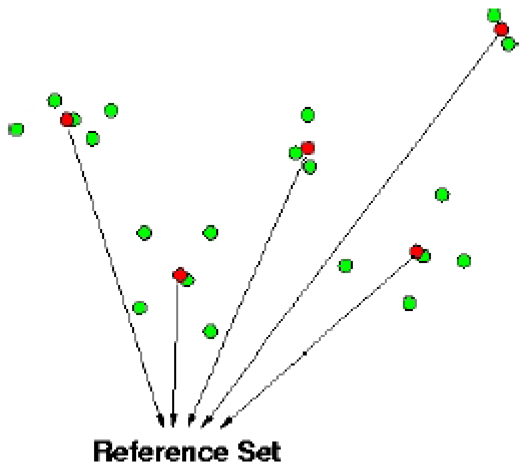


Figure 9. Combining cluster centers (red) into a reference set. (Green dots are feature sets created from images.)

The function that is used to compare a feature set with a reference set is simple and can be computed very fast. It makes identification a matter of seconds, even if a million reference sets have to be compared.

D.3 Controlware GmbH

Face Algorithm Highlights

1. General
   - 1.1. Proprietary Algorithm
   - 1.2. Originally designed for multi-view face representations obtained by video tracking / from multiple face samples
   - 1.3. Recently modified to cope with single images
2. Detection = finding face-like regions in video frame / images)
   - 2.1. Based on morphological (shape-based), structural and statistical criteria
   - 2.2. Insensitive to background variations (does not rely on head contour)
   - 2.3. Does not rely on color or motion
   - 2.4. Uses tracking to capture multi-view face representations
3. Coding = creating Face Characteristic Data (FCD) from (multi-view) face regions
   - 3.1. Coding is preceded by geometric normalization based on accurate estimation of face landmarks location
   - 3.2. Coding uses 5 main face regions for redundancy, robustness
   - 3.3. Each region: represented by multiple vectors
   - 3.4. Each vector represents the information content = face shape and texture in a small neighborhood of the face
   - 3.5. Vectors sets are arranged in hierarchical / multiple level of detail (LOD) manner: small set for fast database searching to larger set for accurate matching score computation
4. Matching = one-to-one comparison process between enrolled and live FCD (vector sets). Matching results are presented by the probability of error for the specific comparison. Thus, 1.0 stands for 1.0% error or 99% confidence.
5. Searching a live FCD against a database of enrolled FCD, using vectors with small level of detail (LOD) for rapid rejection of unlikely subjects and then vector sets with high LOD for computing the similarity score.
6. Implementation highlights:
   - 6.1. Distributed, scalable architecture
   - 6.2. Efficient software implementation for Pentium 4 processors

D.4 ZN Vision Technologies AG

The facial recognition algorithms developed by ZN Vision Technologies AG constitute a systematic enhancement of elastic graph matching (Hierarchical Graph Matching, HGM), first used by Christoph von der Malsburg and colleagues at the Ruhr University Bochum and at the University of Southern California. Under elastic graph matching, the facial geometry is represented by a flexible grid which is adapted to facial pose and expression through adjustment of size, position and internal distortion. A set of special filter structures (Gabor wavelets), which analyse the local facial characteristics (e.g. skin texture), is assigned to every grid point in the elastic graph. The actual facial comparison entails two stages. First of all the grid is superimposed on a face in such a way that the nodal points correspond to specific landmarks, and then the features are compared node-by-node. The similarity between two facial images is then based on a comparison of all the nodal points.

The basic procedure of elastic graph matching has been proven in independent algorithm tests run by the U.S. Army Research Laboratory (FERET tests) and in other comparisons as the most powerful basic approach to facial recognition. The Hierarchical Graph Matching enhancement is designed for use in large databases and for the rapid comparison of faces (over 150,000 comparisons per second on a standard PC) and geared up for processing video sequences and three-dimensional facial data.

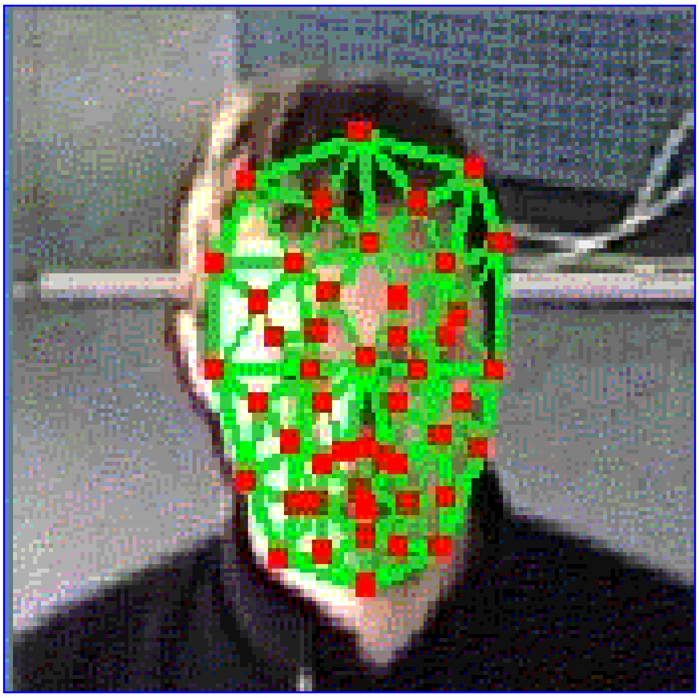


Figure: under Hierarchical Graph Matching, the face is modelled by an elastic graph, with local facial features fixed at the nodal points of the graph.

References

[1] J. Buhmann, J. Lange and Chr. von der Malsburg. Distortion invariant Object Recognition by Matching Hierarchically Labelled Graphs. IJCNN International Conference on Neural Networks. Washington. Vol. I, 155-159 (1989)

[2] L. Wiskott, J.-M. Fellous, N. Krüger, Chr. von der Malsburg. Face Recognition by Elastic Bunch Graph Matching. IEEE Transaction on Pattern Analysis and Machine Intelligence, 19, 775-779 (1997).

Annex E Definition of nois e factors

One of the aims of the BioFace project was to find out what influence noise factors have on the recognition performance of the systems. It was therefore possible in the database to assign any number of image attributes to each image. Every image attribute indicated the presence of one of the noise factors.

In order to distinguish between the case where an image is in fact free of noise factors and the case where it is simply not known whether an image is afflicted with noise factors, there are also image attributes which state that an image is free of noise factors.

The image attributes are represented by numerical values. For BioFace, the following image attributes were defined:

Value	Description
[100-199] Normal photographs	
100	This image does not have any noise factors. This attribute is exclusive, i.e. if this attribute is set for an image, then none of the other attributes can be set for the image.
101	This image does not have any noise factors and has been marked by the BKA explicitly as a search image. This attribute is exclusive, i.e. if this attribute is set for an image, then none of the other attributes can be set for the image.
[200-299] Person-related noise factors	
210	Facial expression (e.g. eyes closed)
215	Position of head (e.g. head inclined)
230	Wounds, illness-induced changes
240	Wearing of items that cover part of the head, e.g. spectacles, beards
[300-399] Photographic noise factors	
[310-319] Problems due to object mapping	
311	Too much/too little light
313	Inadequate contrast
314	Low resolution
315	Fuzziness
316	Silhouette
[320-329] Problems due to light source (photograph platform)	
321	Photographs printed on normal paper
322	Film stuck over (e.g. foreign passports
323	Scanned photograph on plastic background (German identity card)
324	Stamp, rivets attached
329	JPEG error
[330-339] Problems due to aspects of the photographic process	
331	No head

332	Head too small
333	Head not entirely visible
334	Profile photograph
335	Other aspects of the photographic process

Annex F Authors

Christoph Busch	Fraunhofer Institute for Computer Graphics Research, Darmstadt (IGD)
Henning Daum	Fraunhofer Institute for Computer Graphics Research, Darmstadt (IGD)
Frank Graf	Fraunhofer Institute for Computer Graphics Research, Darmstadt (IGD)
Manfred Martin	Federal Criminal Police Office, Wiesbaden, (BKA)
Axel Munde	Federal Office for Information Security (FOIS)
Andrew Pretzel	Federal Criminal Police Office, Wiesbaden, (BKA)
Tom Reinefeld	Federal Criminal Police Office, Wiesbaden, (BKA)
René Salamon	Federal Office for Information Security (FOIS)
Christoph Vogel	Federal Criminal Police Office, Wiesbaden, (BKA)
Mike Wiegand	Federal Criminal Police Office, Wiesbaden, (BKA)