Data Mining in Forensic Image Databases

Zeno Geradts*, Jurrien Bijhold
Netherlands Forensic Institute

ABSTRACT

Forensic Image Databases appear in a wide variety. The oldest computer database is with fingerprints. Other examples of databases are shoeprints, handwriting, cartridge cases, toolmarks drugs tablets and faces.
In these databases searches are conducted on shape, color and other forensic features. There exist a wide variety of methods for searching in images in these databases. The result will be a list of candidates that should be compared manually.
The challenge in forensic science is to combine the information acquired. The combination of the shape of a partial shoe print with information on a cartridge case can result in stronger evidence. It is expected that searching in the combination of these databases with other databases (e.g. network traffic information) more crimes will be solved.
Searching in image databases is still difficult, as we can see in databases of faces. Due to lighting conditions and altering of the face by aging, it is nearly impossible to find a right face from a database of one million faces in top position by a image searching method, without using other information. The methods for data mining in images in databases (e.g. MPEG-7 framework) are discussed, and the expectations of future developments are presented in this study.

Keywords: image databases, data mining, forensic science, biometrics

1. INTRODUCTION

The importance of image databases in forensic science has long been recognized. For example, the utility of databases of fingerprints[1] is well known. Over the past four years, DNA databases have received particular attention and have been featured on the front page of newspapers. These databases have proven extremely useful in verifying or falsifying the involvement of a suspect in a crime, and have led to the resolution of old cases. DNA databases are also playing an increasingly prominent role in the forensic literature[2,3] However, a variety of other databases are also crucial to forensic casework, such as databases of fingerprints, faces, and bullets and cartridge cases of firearms[4]. Research into forensic image databases is a rapidly expanding field of scientific endeavor that has a direct impact on the number of criminal cases solved.

Throughout the 20th century, many databases were available in the form of paper files or photographs (e.g., cartridge cases, fingerprints, shoeprints). Fingerprint databases were computerized in the 1980s, and became the first databases to be widely used in networks. In addition, the Bundes Kriminal Amt attempted to store images of handwriting in a database. These databases were all in binary image format. At the beginning of the 1990s, computer databases of shoeprints, tool marks and striation marks on cartridge cases and bullets became available. Improvements in image acquisition and storage made it economically feasible to compile these databases in gray-value or color format using off-the-shelf computers

Some forensic databases contain several millions of images, as is the case with fingerprints. If databases are large, the forensic examiner needs a method for selecting a small number of relevant items from a database, because if this cannot be achieved the investigation becomes time consuming and therefore either expensive or impossible. The retrieval of similar images from a database based on the contents of each image requires an automatic comparison algorithm that is

* zeno@forensic.to;phone +31704135681;fax +31704135454;Netherlands Forensic Institute;Volmerlaan 17;2288 GD Rijswijk;Netherlands;

fast, accurate, and reliable. To formulate such an algorithm, one must first identify which parts or features of the images are both crucial and suitable for finding correspondences. The development of the retrieval system then requires a multidisciplinary approach with knowledge of multimedia database organization, pattern recognition, image analysis and user interfaces.

Forensic image databases often contain one or more sub-databases:

- Images that are collected from the scene of crime (e.g., shoeprints recovered from the crime scene) à with this database it is possible to link cases to each other

- Images that are collected from the suspect (e.g., shoeprints that are collected from a suspect) à with this database in combination with the database of images that are collected from the scene of crime, it is possible to link suspects with cases

- Reference images (e.g., shoeprints from shoes that are commercially available, that can be used to determine which brand and make of shoe a certain shoe print is from)

The compiling of large-scale forensic image databases has made available statistical information regarding the uniqueness of certain features. For example, at the beginning of the 20th century there have been arguments about the number of matching points that are needed for concluding that a fingerprint matches. (depending on the country5 this could be 8-16 points). Nowadays, however, statistical ranking in fingerprint databases provides more information regarding the uniqueness of fingerprints and the number of points versus the statistical relevancy. Up until very recently, most forensic conclusions were drawn more on experience of the investigator than on real statistics. The statistical information now available from databases should result in forensic investigation that is more objective. This is necessary since courts and lawyers are asking questions that are more critical about forensic investigation and conclusions that are based on experience of the investigator instead of real statistics[*].

2. Visual Content

Previous studies on information retrieval from image databases based on visual contents have used the following features[6,7] such as color, texture, shape, structure, and motion, either alone or in combination.

Color
Color reflects the chromatic attributes of the image as it is captured with a sensor. A range of geometric color models[8] (e.g., HSV, RGB, Luv) for discriminating between colors are available. Color histograms are the most traditional technique for describing the low-level properties of an image.

Texture
Texture has proved to be an important characteristic for the classification and recognition of objects and scenes Haralick and Shapiro[9] defined texture as the uniformity, density, coarseness, roughness, regularity, intensity, and directionality of discrete tonal features and their spatial relationships. Haralick[10] reviewed the two main approaches to characterizing and measuring texture: statistical approaches and structural approaches. Tuceryan and Jain[11] carried out a survey of textures in which texture models were classified into statistical models, geometrical models, model-based methods, and signal processing methods.

---

[*] *On January 9th 2002 U.S. District Court Judge Louis H. Pollak in Philadelphia, ruled that finger print evidence does not meet standards of scientific scrutiny established by the U.S. Supreme Court, and said fingerprint examiners cannot testify at trial that a suspect's fingerprints "match" those found at a crime scene. It is expected that other forensic fields (e.g., handwriting, toolmarks, hairs) will also be challenged for court..*

Shape

Shape features are expressed in text, for example squares, rectangles, and circles. However, complex forms are more difficult to express in text.

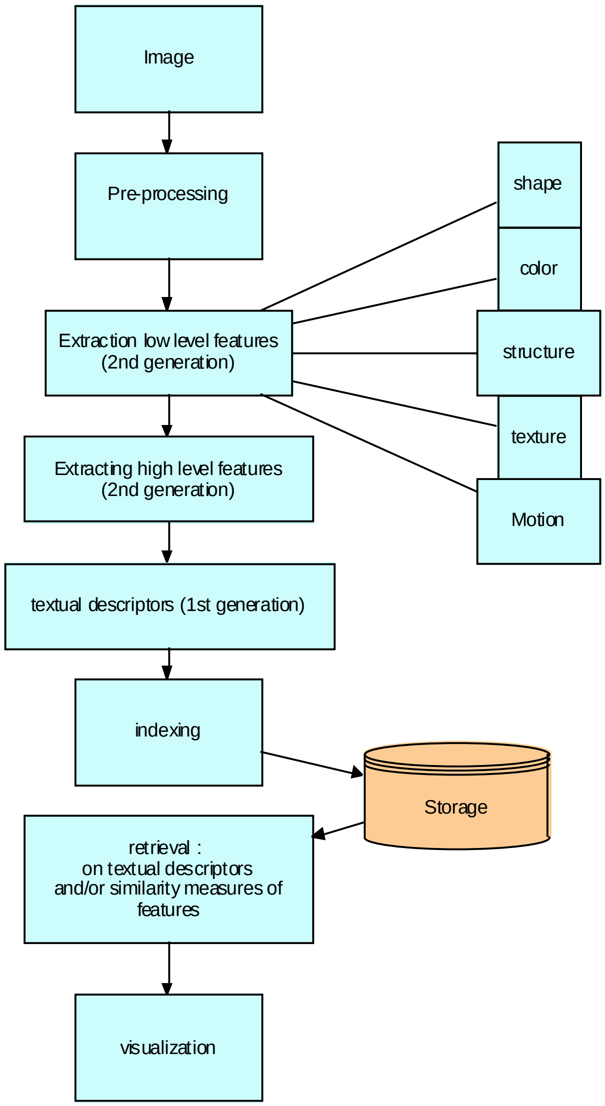

Figure 1: Flowchart for a visual information retrieval system

Traditionally, shapes are expressed through a set of features that are extracted using image-processing tools. Features can characterize the global form of the shape such as the area, local elements of its boundary or corners, characteristic points, etc. In this approach, shapes are viewed as points in the shape feature space. Standard mathematical distances are used to calculate the degree of similarity of two shapes.

Preprocessing is often needed to find the shapes in an image. Multiple scale techniques[12] are often used as filters to elucidate the shapes in an image. In many cases, shapes must be extracted by human interaction, because it is not

always known beforehand which are the important shapes in an image.
The property of invariance (e.g.; the invariance of a shape representation in a database to geometric transformations such as scaling, rotating, and translation) is important in the comparison of shapes

Structure

The structure of an image is defined as the set of features that provide the “gestalt” impression of an image. The distribution of visual features can be used to classify and retrieve an image. Image structure can be important for achieving fast pre-selection of a database (i.e., the selection of a part of the database) based on the image contents. A simple example is distinguishing line drawings from pictures gray values and location in the image.

Motion

Motion is used in video databases and is analyzed in a sequence of frames. There are several models for calculating the motion vectors from a sequence of images. These methods range from simply determining the motion from the difference between two images to more complex approaches using optical flows for different objects.

3. Similarity Measures

Similarity measures are used to identify the similarities in images. These measures are usually based on histograms of a feature in the image; however, more sophisticated implementations taking into account other measurements from the images are also possible. A metric model is frequently used to compare features that can be expressed within a metric system. The implementation of this method is particularly efficient in terms of the computing power required. Several distance functions are commonly used:[13] the Euclidean distance, the city-block-distance, and the Minkowsky distance for histograms. However, many other metric models can be implemented. Another approach to measuring similarity is the use of transformational distances.14 This approach requires the definition of a deformation process that transforms one shape into another, where the amount of deformation is a measure of the similarity. One such group of models are the elastic models, which use a discrete set of parameters to evaluate the similarity of shapes. Evolutionary models, on the other hand, consider the shapes that result from a process in which at every step forces are applied at specific points of the contours.

Several studies in psychology have pointed out that the human visual system has a number of inadequacies compared to algorithm-based systems. The Earthmover’s distance[15] is an example of a new kind of implementation that, according to the developers, seems to be equivalent to the human visual system. The developers of the Earthermover’s distance claim that it is easy to implement and that efficient indexing methods are possible.

4. Biometric databases

Biometric systems are an important area of research into visual information retrieval systems. In this section, examples of practical implementations are discussed in relation to the methods described earlier.

Biometric systems serve two potential goals: identification or recognition.[16] Recognition is often used to distinguish a particular individual from a limited number of people whose biometric data is known. For instance, a company that has 40 employees uses face recognition access control, in which the 40 employees have to be distinguished from each other. If an error (false positive or false negative) is made with this access control system, this is acceptable. Identification is much more difficult to achieve than recognition, because false positives are unacceptable.

For biometric systems, a group of measures are used to identify or recognize a person. For a measure to be suitable for a biometric system, it must satisfy several requirements[17]:

- Universality: the biometric measure must be a characteristic possessed by all people.
- Uniqueness: no two individuals should have the same biometric measure
- Permanence: the biometric measure should not change with time.

Examples of biometric measures are fingerprints, palm prints, DNA, iris, face, handwriting, gait, and speech. In the next section, some implementations of databases of fingerprints, faces, and handwriting are described, which meet most of the requirements mentioned above. A study to gait comparison is also included in this chapter, which does not meet all requirements.

Fingerprints

The most commonly used system for classifying fingerprints is the Henry Classification scheme[18]. Although this classification system dates back to the early 1900s, it is still used in commercial systems where it has been used since the era in which fingerprints were classified manually. Commercial Automatic Fingerprint Identification Systems (AFIS) usually employ this classification scheme; however, other approaches have been developed.

Faces

Images of faces are easy to obtain with a camera, and are important for the surveillance industry. The problem with recognizing faces in images from a camera system is that the faces are not acquired in a standardized fashion. The face can be in any position and the lighting and magnification can vary. Furthermore, hairstyle, facial hair, makeup, jewels and glasses all influence the appearance of a face. Other longer-term effects on faces are aging, weight change, and facial changes such as scars or a face-lift. Images of faces are generally taken under more standardized conditions for police investigations[19].

Before a face can be analyzed, it must be located in the image[20]. For face recognition systems to perform effectively, it is important to isolate and extract the main features in the input data in the most efficient way. One of the main problems that must be overcome in face recognition systems is removing redundant sampling so as to reduce the dimensionality[21] Sophisticated pre-processing techniques are required to attain the best results.
Some face recognition systems represent the primary facial features (e.g., eyes, nose, and mouth) in a space based on their relative positions and sizes. However, important facial data may be lost using this approach[22], especially if variations in shape and texture are considered to be an important aspect of the facial features.

Template matching involves the use of pixel intensity information, corresponding to the largest eigenvalues. This can be in the original gray-level dataset or in pre-processed datasets. The templates can be the entire face or regions corresponding to general feature locations such as the mouth and eyes. Cross correlation of test images with all training images is used to identify the best matches.

Statistical approaches can also be used in the analysis of faces. For example, Principal Components Analysis (PCA[23]) is a simple statistical dimensionality reducing technique that is frequently employed for face recognition. PCAs extract the most statistically significant information for a set of images as a set of eigenvectors (often referred to as 'eigenfaces'). Once the faces have been normalized for their eye position they can be treated as a one-dimensional array of pixel values, which are called eigenvectors.

For forensic investigation based on facial comparison that is focused on identification, there exist different approaches for positioning. It is necessary to get a reference image in which the suspect is positioned exactly as in the questioned image.. Approaches for positioning a suspect are described by van den Heuvel[24] and Maat[25]. Positioning takes much time, and this is often not available in commercial systems.

In practice it appears that it is not possible to identify persons with the regular CCTV-systems. The expectations of face recognition software is sometimes too high, as can be seen in a trial in Palm Beach for searching terrorists[26]. It appeared that the system falsely identified several people as suspects and has at times been unable to distinguish between men and women

Handwriting

Various handwriting comparison systems[27] exist on the market. The oldest system is the Fish system, which was developed by the Bundes Kriminal Amt in Germany. Another well-known system is Script, developed by TNO in the Netherlands.

In both systems, handwriting is digitized using a flatbed scanner and the strokes of certain letters are analyzed with user interaction. The features of the handwriting are represented as content semantics.

Gait

Gait is a new biometric[28] aimed to recognize a subject by the manner in which they walk. Gait has several advantages over other biometrics, most notably that it is non-invasive and perceivable at a distance when other biometrics are obscured.

At the present time, many crimes, including bank robberies, are captured using CCTV surveillance systems sited at stores, banks and other public places. These recordings are often passed to our institute for the purpose of identification. If a criminal has covered his face, the recognition is much more difficult. The question then asked is if it is possible to compare the gait of the perpetrator with the gait of the suspect. For this purpose it is necessary that some of the gait parameters have subject characteristic features.

Since there is not much known on the characteristics of gait in literature and use in forensic analysis, we have started a research project on gait analysis. Human gait contains numerous parameters. These parameters can be categorized into spatial-temporal and kinematic parameters. Because it was impossible to investigate all gait parameters in our study[29], a selection has been made on the criteria that the gait parameters could probably also be obtained in non-experimental settings, and could be characteristic of a person. In our experiments markers were used as is shown in. The subjects wore only their underwear and shoes.

From this research it appeared that with this measurement the gait was not unique. Some non-characteristic differences were measured between different persons, so it could be used for distinguishing people from each other. For forensic analysis of CCTV-images, gait analysis is even more difficult, since there are most often a few images available and the subjects wear clothes.

5. Databases in the forensic laboratory

An overview is shown on methods for automatic comparison of images. For databases of tool marks, cartridge cases, shoe prints and drugs tablets a wide variety of implementations have been accomplished at our laboratory in the last ten years.

Toolmarks[30]

From the methods available for image matching, texture is an option for making a smaller set. Spatial relationships in the tool mark are in the end necessary for the comparison. The similarity model that has been used is based on registration of two line patterns, in view of the fact that there might be a shift and a zoom. The main contribution of this research was to develop a method for extracting relevant forensic information from the striation marks (Figure 2). With a 3D-approach that is implemented in 1999, this resulted in higher correlation factors for matching striation marks.

For the toolmarks, the practical use of these databases is still limited. In the Netherlands, there exist some databases of tool marks, however automatic comparison has not been implemented. Several efforts have been made to implement automatic comparison of striation marks. The results of these algorithms are promising depending on the amount of time that is available, as also can be seen in the bullet systems (IBIS).

Cartridge Cases[31]

This research is focused on pre-selection of features. For improving the features that are selected, a manual intervention might improve the results. Registration methods are used for the comparison. Shape for the firing pin, could be a good option for making a faster pre-selection. A faster pre-selection is also possible with texture and or structure of the impression marks. Spatial relationships have been implemented in our research.
The cartridge case systems are widely used compared to toolmarks databases. These systems have correlation engines, and modification to a 3D-system will result in better correlation ranks. In the Netherlands we use the system Drugfire,

as is used by many agencies in the United States. The company that produces this product is phasing out the software, since there appeared to be patent infringement problems with the other company IBIS. In the United States, the IBIS system will be the standard. This system has a more reproducible way of imaging the cartridge cases, and better results are possible with this system. In practice, the system results in "cold" hits

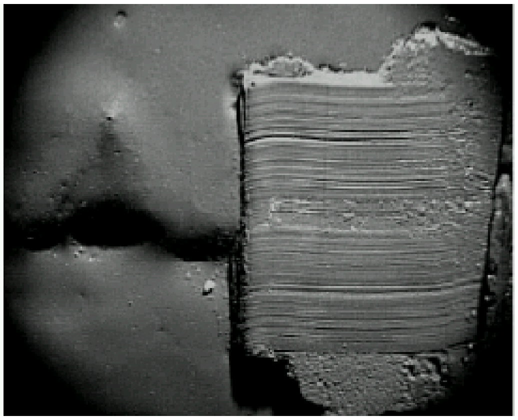

Figure 2: Example of striation mark from screw driver

.

Shoe Prints[32]

This research has been focused on shape recognition. At the time of the study (1995), computing power and memory was limited which caused that more sophisticated implementations have not been tested. In the mean time other methods for shape recognition have been implemented (as implemented with the drugs tablets study), that could result in a better matching. The spatial relationships between the different shapes of the shoe profile have not been implemented. This will improve the results.

Digital shoe print systems are often used in the United Kingdom and Japan. Many shoe print systems could not survive in the market, since government forensic laboratories that develop these systems, develop them for themselves, and this means that they do not really want to market them. In the Netherlands, we have seen some results with these systems, when used manually. Automatic classification and comparison is possible for clear shoeprints. In practice, the problem with shoeprints is that they are often vague, and for that reason a human being should classify. Shoe prints are valuable in forensic science. They are time-consuming for comparison and collection, and it depends on the police region if they are used. In regions with much violent crimes, we see that this kind of evidence is used less. Shoe prints in blood are an important part of evidence for a homicide that is sometimes skipped due to limitation in time per case. For this reason, the use of shoe print databases should be promoted, since more crimes can be solved, and more forensic information is available.

Drugs Tablets[33]

This research is focused on shape recognition. The different methods for shape recognition should be refined for the approach of the 3D shape. The other approach is by making better images, as this will improve the results importantly. Registration methods have been implemented for the comparison. The combination with text (description of logo) of the drugs tablet will greatly reduce the effort, and for larger databases (> 1000) this research should be implemented.
In future, if more images are in the database, it could result in more need for such a logo comparison. The problem with shape comparison of these logos is to filter out the images of the logo. It appeared that the acquisition with a regular camera system, the quality of those images is limited, and this results in huge problems of splitting the logos from the drugs tablets with an image processing method. With 3D-techniques, the real shape of the stamp is available, which is under current research.

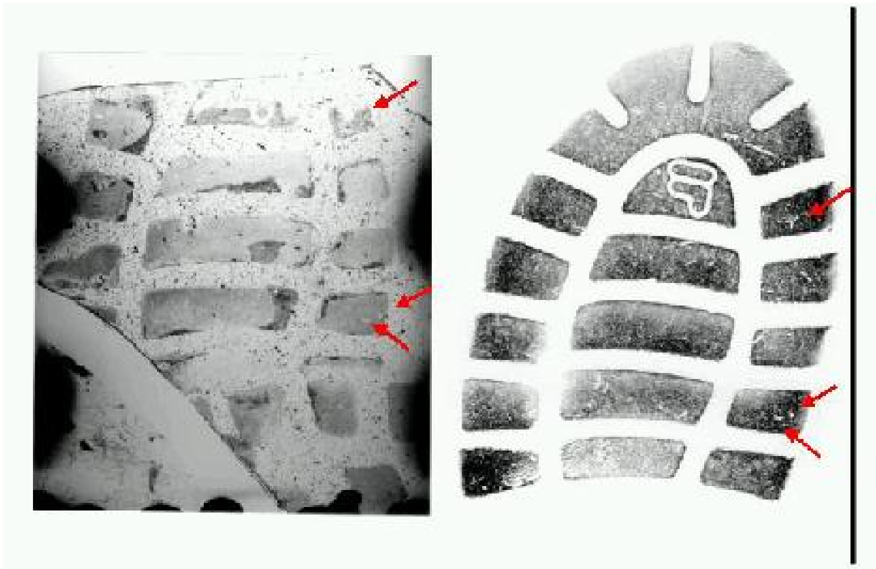

Fig. 3: Shoe print comparison of shoeprint found at the crime scene (left) with a test shoeprint (right). The characteristic marks are pointed with arrows.

6. Conclusions

In this study it appeared that general approaches of searching in forensic image databases are applicable to forensic science databases. The human being is in general still much better in interpreting the separate images, compared to an algorithm. For larger databases, it is not feasible to ask an investigator to compare all forensic evidence, and for this reason the use of image databases in forensic science is very useful, even for crimes in "cold" cases.
Pre-processing of the images is important in all forensic applications we have evaluated. The data acquisition should be handled with a standardized approach for optimal results.

7. Discussion and future research

For getting more results out of the separate databases, it is necessary to combine the information of the forensic database. This will result in relations between crimes, which were not considered before. Relations can be found that afterwards appear to be a coincidence and misleading, and sometimes they will really result that the case is solved.
It is also possible to use data mining techniques in combination with forensic image databases. Biometric databases for access control (which are expected to be more widely implemented) can also be used. Also information of (cellular) phones and other network systems can be used to track and trace criminals, which will result in an efficient way of reducing data. Furthermore, credit card and banking information combined with the loyalty programs of shops are a way of reducing the number of possibilities between a crime scene and suspects. These methods can not be used for all cases, since privacy issues are involved. For this reason the law in many countries will protect us against un-proportional use of data for criminal investigation.

All these methods will result that a criminal is more aware that traces can be found. In court, they are confronted with the evidence. This evidence is public, and the networks of criminals will be informed about new methods for solving crimes. Another issue is that they will try to alter the evidence in a way that another person is charged for the crime, as can be seen with DNA evidence.

The developments in image acquisition (3D) and image matching algorithms (e.g. MPEG-7) combined with the computing power of future systems, will result in systems that can be used effectively for solving more crimes.
It remains important to analyze the results of the search and to have user feed back about the position that a certain image is found on the hit list. The conclusion on matching of two marks or prints should be done by a forensically qualified human being. The forensic examiner should also learn from the image databases and more statistical information of these databases should be used in their investigation.

ACKNOWLEDGEMENTS

For this work we would like to acknowledge the Netherlands Forensic Institute, especially Hage Postema for his contribution to this work.

REFERENCES

1. J.H. Wegstein, J.F. Rafferty, "The LX39 Latent Fingerprint Matcher", NBS Special Publication, **500**(36), pp.15-30, 1978.
2. T.D. Kupferschmid,, T. Calicchio, B. Budowle**; "**Maine Caucasian population DNA database using twelve short tandem repeat**.",** Journal of Forensic Sciences, **44**(2), pp. 392-395, 1999.
3. D.J. Werrett, "The National DNA Database" Forensic Sciences International, **88**(1) , pp. 33 42. 1997.
4. R.W. Sibert, "DRUGFIRE Revolutionizing Forensic Firearms Identification and Providing the Foundation for a National Firearms Identification Network", USA Crime Lab Digest, **21**(4), pp. 63-68, 1994.
5. I.W. Evett; R.L. Williams; "A Review of the Sixteen Points Fingerprint Standard in England and Wales", UK Fingerprint World, **21**(82) P125-143, 1995.
6. S.K. Chang, A. Hsu, "Image Information systems: Where do we go from here?", IEEE Transactions on Knowledge and Information Engineering, **4**(5), pp.644-658, 1994.
7. P. Agrain; H. Zhang; D. Petkovic; "Content-based representation and retrieval of visual media: a state of the art review", Multimedia Tools and Applications, **3**(3), pp. 179-202, 1997.
8. M.J. Swain; D.H. Ballard; "Colour Indexing",International Journal of Computer Vision, **7**(1), pp. 11-32, 1995.
9. R. Haralick, L. Shapiro; "Glossary of computer vision terms", Pattern Recognition, **24**(1), pp. 69-93, 1991.
10. R. Haralick, "Statistical and structural approaches to texture", Proceedings of IEEE, **67**(5), pp. 786-804, 1977.
11. N. Tuceryan, A.K. Jain; *Handbook of Pattern Recognition and Computer Vision*, World Scientific Publishing Company, pp. 235-276, 1993.
12. A. Witkin; "Scale Space filtering", Proceedings of 7th International Conference on Artificial Intelligence, pp. 1010-1021. 1983.
13. S. Aksoy; R.M. Haralick; "Probabilistic vs. Geometric Similarity Measures for Image Retrieval", Proceedings of Computer Vision and Pattern Recognition (CPRV'00), pp. 112-128, 2001.
14. A.L. Youille, K. Honda, C. Peterson; "Particle tracking by deformable templates", Proceedings International Joint Conference On Neural Networks, pp. 7-11, 1991.
15. R. Yossi R; C. Tomasi; *Perceptual metrics for Image Database Navigation*., Kluwer Academic Publishers, 2001.
16. P.J. Phillips PJ, "An Introduction to Evaluating Biometric Systems", Computer, **33**(2), pp. 56-63, 2000.
17. B. Miller; "Vital Signs of Identity", IEEE Spectrum, **31**(2), pp. 22-30, 1994.
18. E.R. Henry, *Classification and uses of Finger Prints*, Routledge, London, 1900.
19. J. Atick, P.M. Griffin, A.N. Redlich, "FaceIt: face recognition from static and live video for law enforcement" Publication: Proc. SPIE **2932**, p. 176-187, Human Detection and Positive Identification: Methods and Technologies, 1997.
20. H.A. Rowley, S. Baluja, T. Kanade, "Human Face detection in visual scenes', in D.S. Toretzky, M.C. Mozer, M.E. Hasselmo (Eds), Advances in Neural Information Processing, Volume 8, pp. 875-881, Cambridge, MIT Press, 1996
21. A.J. Howell, "Introduction to Face Recognition", CRC Press International Series on Computational Intelligence, pp. 219-283, 1999.

22. A. Samal, P.A. Iyengar, "Automatic recognition and analysis of human faces and facial expressions : A survey". Pattern Recognition, **15**, pp. 65-77, 1992.
23. P. Penev; L. Sirovich, "The Global Dimensionality of Face Space", Proceedings of Fourth IEEE International Conference on Automatic Face and Gesture Recognition, 2000.
24. H. van den Heuvel; "Positioning of persons or Skulls for Photo comparisons using Three-Point Analyses and One Shot 3D Photographs", Proceedings of SPIE, **3576**, pp. 203-215, 1998.
25. G.J.R. Maat, "The positioning and magnification of faces and skulls for photographic superimposition", Forensic Science International, **41**, pp. 225-235, 1989.
26. L. Bierman; "FaceScan Ineffective ACLA says", January 4th 2002, Palm Beach Post, page 1.
27 L. Schomaker, L. Vuurpijl; "Forensic writer identification: A benchmark data set and a comparison of two systems", Report from the University of Nijmegen, 2000.
28 J. Foster; "New Area Based Measures for Gait Recognition", Audio-And Video-Based Biometric Person Authentication : Proceeding of the Third International Conference, Avpa 2001 Halmstad, Sweden, June 6-8, 2001, pp. 312-317.
29 Geradts; Merlijn M; Bijhold J; "Use of gait parameters of persons in video surveillance systems.", submitted to Proceedings SPIE Investigative and Trial Image Processing, 2002.
30. Z.J. Geradts, D. Zaal, H. Hardy, J. Lelieveld, I. Keereweer, J. Bijhold; "Pilot investigation of automatic comparison of striation marks with structured light", Proc. SPIE Vol. **4232**, pp. 49-56, 2001.
31. Z.J. Geradts, J. Bijhold, R. Hermsen, F. Murtagh, Image matching algorithms for breech face marks and firing pins in a database of spent cartridge cases of firearms, Proc. SPIE Vol. **4232**, pp. 545-552, 2001.
32. Z. Geradts, J. Keijzer, The Image database REBEZO for Shoeprints with developments on automatic classification of shoe outsole designs, Forensic Science International, **82**, No. 1, pp. 21-31, 1996.
33. Z.J. Geradts, H. Hardy, A. Poortman, J. Bijhold, "Evaluation of contents-based image retrieval methods for a database of logos on drug tablets", Proceedings SPIE Vol. **4232**, p. 553-562, 2001.