Ear Recognition

Biometric Identification using 2- and 3-Dimensional Images of Human Ears

Anika Pflug

Thesis submitted to Gjøvik University College
for the degree of
Doctor of Philosophy in Information Security

2015

Ear Recognition

Faculty of Computer Science and Media Technology
Gjøvik University College

Ear Recognition - Biometric Identification using 2- and 3-Dimensional Images of Human Ears / Anika Pflug
Doctoral Dissertations at Gjvik University College 2-2015
ISBN: 978-82-8340-007-6
ISSN: 1893-1227

We choose to go to the moon in this decade and do the other things, not because they are easy, but because they are hard.

(J. F. Kennedy)

Declaration of Authorship

I, Anika Pflug, hereby declare that this thesis and the work presented in it is entirely my own. Where I have consulted the work of others, this is always clearly stated.

Signed:

(Anika Pflug)

Date:

Summary

The outer ear is an emerging biometric trait that has drawn the attention of the research community for more than a decade. The unique structure of the auricle is long known among forensic scientists and has been used for the identification of suspects in many cases. The next logical step towards a broader application of ear biometrics is to create automatic ear recognition systems.

This work focuses on the usage of texture (2D) and depth (3D) data for improving the performance of ear recognition. We compare ear recognition systems using either texture or depth data with respect to segmentation and recognition accuracy, but also in the context of robustness to pose variations, signal degradation and throughput. We propose a novel segmentation method for ears where texture and surface information are fused in the feature space. We also provide a reliable method for geometric normalization of ear images and present a comparative study of different texture description method and the impact of their parametrization and the capture settings of a dataset. In this context, we propose a fusion scheme, were fixed length spectral histograms are created from texture and surface information.

The proposed ear recognition system is integrated into a demonstrator system as a part of a novel identification system for forensics. The system is benchmarked against a challenging dataset that comprises of 3D head models, mugshots and CCTV videos from four different perspectives. As a result of this work, we outline limitations of current ear recognition systems and provide possible direction for future applied research.

Having a complete ear recognition system with optimized parameters, we measure impact of image quality on the accuracy during ear segmentation and ear recognition. These experiments focus on noise, blur and compression artefacts and are hence only conducted on 2D data. We show that blur has a smaller impact on the system performance than noise. In scenarios where we work with compressed images, we show that the performance can be improved by optimizing the size of local image patches for feature extraction and the size of the compression artefacts.

This thesis is concluded by work on automatic classification of ears for the purpose of narrowing the search space in large datasets. We show that classification of ears using texture descriptors is possible. Furthermore, we show that the class label is influenced by the skin tone, but also by the capture settings of the dataset. In further work, we propose a method for the extraction of binary feature vectors of texture descriptors and their application in a 2-stage search system. We show that our 2-stage system improves the recognition performance, because it removes images from the search space that would otherwise have caused recognition errors in the second stage.

Acknowledgements

I would like to express my gratitude towards CASED and Hochschule Darmstadt, who hosted me for four years and provided me with everything I needed for conducting my research. I am grateful for the opportunities I had at Gjøvik University College, where my research in biometrics started back in 2010 with my master thesis on vein recognition. This would not have been possible without the grateful and unprejudiced support of Christqoph Busch and Daniel Hartung. During the entire time of my studies, I received invaluable feedback and guidance from Christoph.

Further, I would like to say thank you to all the project partners in GES-3D and my co-workers and lab mates Chris Stein and Xuebing Zhou. It was a pleasure working with you and I really do not know, how the project would ever have been completed without your commitment.

During my work at Hochschule Darmstadt, I was happy to supervise Adrian, Johannes, Philip and Ulrich, who were excellent students. All of them have done great work and provided valuable support for my research. Along with these great students, I am grateful to be part of the da/sec research group, where I was working together with great lab mates, such as Andreas, Björn, Christan D., Christan R., Jessica and Martin.

I would also like to thank Prof. Arun Ross for giving me the opportunity to visit his lab at Michigan State University. Thanks to my lab mates Ajita, Antitza, Asem and Thomas, I was having a wonderful time in Michigan.

Finally, special thanks go to my family for their support during my entire life. Finally, I would like to thank Kim, who accompanied me though all the years and gave a meaning to so many things.

Contents

List of Figures

List of Tables

Part I

Introduction

Chapter 1

Biometric Ear Recognition

1.1 Introduction

Being originally used for forensic identification, biometric systems evolved from a tool for criminal investigation into a number of commercial applications. Traditional means of automatic recognition such as passwords or ID cards can be stolen, faked or forgotten. Contrary to this, a biometric characteristic should be universal, unique, permanent, measurable, high-performing, acceptable for users and it should be as hard as possible to circumvent the biometric authentication system. Today we find fingerprint recognition system in cellphones, laptops and front doors. With the increased dissemination of biometric systems, the recognition performance in the field of forensics has also improved during the last years.

Biometric systems have helped to identify the two suspects of the Boston Bombings in 2013 [103] and also for identifying a suspect for a series of robberies in gas stations in the Netherlands [79]. In the latter case, ear recognition played an important role for the identification of the suspect.

The outer ear as a biometric characteristic has long been recognized as a valuable means for personal identification by criminal investigators. The French criminologist Alphonse Bertillon was the first to investigate the potential for human identification of ears more than one century ago in 1890 [26]. In 1893 Conan Doyle has published an article in which he describes particularities of ears from selected famous people and argues that the outer ear, just like the face, reflects properties of a person's character [53]. In his studies about personal recognition using the outer ear in 1906, Richard Imhofer only needed four different characteristics to distinguish 500 different ears [83]. Later on in 1964 the American police officer Alfred Iannarelli collected more than 10 000 ear images and determined 12 characteristics to identify a subject [81]. Iannarelli also conducted studies on twins and triplets, where he discovered that ears are even unique among genetically identical subjects. Scientific studies confirm Iannareli's assumption, that the shape of the outer ear is unique [124], and it is also assumed to be stable throughout a human life time [125].

In numerous research papers from the last decade, it could be shown that the biometric recognition performance achievable with automated ear recognition systems is competitive to face recognition systems [6]. Lei *et al.* have also shown that the outer ear can be used for gender classification [110].

In many criminal cases there is no further evidence than a CCTV video where we see the perpetrator committing a crime. Such footage can only be used as evidence in court, if the perpetrator's identity can be determined from the CCTV footage without a doubt. Criminal investigators usually try to combine information coming from witnesses or the victim with cues from the CCTV footage. Biometric identification is a helpful tool for this task, however the uncontrolled conditions in CCTV videos remain a difficult setting for every automated identification system. For this purpose, national police offices maintain a database of criminals with mugshots that serve as the biometric reference during an automatic search. Typical difficulties for automatic identification in this particular use case are for example pose variations, sparsely lit scenes, image compression, blur and noise.

If there is no further information available but the CCTV footage, the German criminal police (Bundeskriminalamt or short BKA) tries to identify suspects with the help of an au-

Figure 1.1: CCTV footage of a man who stole money from several ATMs in the Rhein-Main area. The right image is a close up of the video frame on the left. This image was taken in a bank in Frankfurt. Note that face identification is impossible, because the subject in wearing a baseball cap. (image source: BKA. The case was closed in August 2014, but the video is still available at [34]

tomatic face recognition system, called GES (An abbreviation of "Gesichtserkennungssystem", which is German for face recognition system). The reference images for GES are collected by criminal police officers during crime investigation processes and are stored in a central database that is maintained in the premises of Bundeskriminalamt (BKA) in Wiesbaden. A set of reference images consists of a series of mugshots that contains at least one frontal portrait image, a half profile and a full profile image (from the right side). Newly collected datasets follow a new standard, where left and right half profile, as well as left and right full profile images are acquired [175].

In order to protect CCTV equipment from vandalism, cameras are often installed in corners or underneath the ceiling. Figure 1.1 shows an example of such video footage from a real case, where we see a man leaving a bank after he robbed an ATM. CCTV cameras are usually arranged to deliver face images. Perpetrators, who are aware of the presence of a CCTV camera will avoid to directly look into the camera and some times wear hats in order to cover their faces. This means that investigators frequently have to work with half profile or profile views where the face can be partly or fully occluded. In such scenarios, ear recognition can be a valuable amendment to existing face recognition systems for identifying the subject.

As long as the video contains a single frame, where the subject's face is clearly visible from one of the angles that match with one of the reference mugshots, automatic identification has a chance to succeed. In practice, however, clear images from a defined pose are rarely the case. Typical countermeasures against being identified are baseball caps, hats or scarves. Additionally, the resolution of surveillance cameras in relation to the monitored region can be small and we may encounter additional signal degradations, such as interlacing and blur.

In a small study, which was conducted at Darmstadt central station between September 2013 and June 2013, we tried to estimate the average probability that the outer ear is visible in public. The observation took place in the entrance hall of the train station where people walk through a door. At different times during the day, we counted the number of visible and occluded ears. We also made notes about the type of occlusion and the weather outside. We observed that the ear was fully visible in 46% of the cases (5431 observations in total). At the same time, the probability of occlusion is highly dependent on the gender. Whereas the ears of women were only fully visible in 26.03% of the cases, the ears of men were fully visible in 69.68% of the cases. More details about this study can be found in Appendix A. With forensic identification of suspects in mind, this result is still encouraging because 74.3% of the suspects in Germany in 2013 were male [136].

Table 1.1: Level of support for imagery as evidences at court as agreed upon by the FIAG. The table is taken from [140]

Level	Description
0	Lends no support
1	Lends limited support
2	Lends moderate support
3	Lends support
4	Lends strong support
5	Lends powerful support

1.2 Forensic Identification Using Ear Images

In order to be valid in court, any imagery should provide an objective evidence that is independently measurable and verifiable [140]. Many courts require an estimation of the strength of evidence within an image from an independent expert. Without such an expertise, CCTV footage is not accepted as a valid evidence. Table 1.1 shows an example of different levels of certainty, which comply with the standard of the Forensic Imagery Analysis GROUP (FIAG). In an expertise, a certainty level is assigned to each trait in the image. The conclusion of a forensic expertise is a likelihood estimation of the probability that the suspect and the subject in the video are one and the same person. The expert testimony can be supported by an automated identification system, but an automated decision may never be the only source of evidence. Automatic identification is usually used for selecting the most likely candidates (usually between 50 and 100 subjects) that are then further examined by a forensic expert. According to the Daubert Standard, the way of how an expert came to a given conclusion must be clear and comprehensive for the judge and it should be based upon sufficient facts. Secondly, any sources of error must be transparent and the principle that was used for the conclusion must be an established scientific standard. Finally the principle must be applicable for the particular case [25].

The probe presentation of the outer ear can either be analyzed directly from any imagery where it is clearly visible or ear prints can be taken from surfaces. Ear prints can be left on windows or doors, when burglars try to make sure that the house is empty [125]. Meijerman has shown that Ear prints are individual, but are dependent on different factors such as pressure and temperature, which can result in a high intra-class variation [123]. In Germany, there are several example cases, where ear prints have helped the investigators to identify the suspects, such as in Osnabrueck [1] and Hamburg [2].

The evidential values of ear prints is heavily discussed for several reasons. Firstly, ear prints are usually left *before* a crime is committed and are hence not necessarily a proof that the subject who left the ear print is the same subject who committed the crime. Secondly, there is also no indication of the time when the ear print was left [125]. Finally it is argued that comparing actual ears and an image of the ear is different from comparing two ear prints. The tissue of the outer ear is soft and additional factors such as pressure, moist, the lifting process and surface structure have an influence on the appearance of the ear print. In research on ear recognition, images of known subjects are compared with each other, showing that the appearance of the ear is unique enough to distinguish between the subjects in the database. There are controversial discussions whether ear prints are equally unique and permanent [101].

For the purpose of identifying subjects from CCTV images, a systematic way to describe

[1] http://www.ksta.de/html/artikel/1218660630642.shtml

[2] http://www.n-tv.de/panorama/Ohrabdruck-ueberfuehrt-Einbrecher-article6144406.html

the outer ear is essential. Such standard descriptions exist for all kinds of characteristics that can be observed in imagery, including the outer ear. Such description methods consist of a list of relevant landmarks (*e.g.* concha, outer helix, tragus). During the analysis, the forensic expert describes the appearance of each part and then gives an estimate of how similar the representations in the probe CCTV image and the reference image are (usually the mugshot from the database but sometime other images are used as well). For the ear lobe such descriptions could be for instance: hanging, attached, partly attached. The sum of all of these descriptions together with an estimated similarity between the suspect and the subjects in the reference image are summarized in an expertise that can be used at court. For good reasons, a expertise must be prepared from a human expert and not by an automated system. Consequently, current ear analysis concepts are highly optimized towards manual analysis.

In forensic identification, biometric identification systems are used for retrieving the most likely candidates from a large database. Current systems mainly rely of facial images for reducing the number of possible suspects. The reliability of these systems is reduced by pose variations and occlusions, but also by low image contrast and compression artefacts. Automated identification systems with full 3D images as references can provide a pose-independent view of the subject, which can potentially result into more accurate estimates of the similarity by offering the possibility of adjusting the pose of the reference image to the pose of the probe image. Such an estimation could also include the shape of the outer ear, especially in cases where only half profile views of the suspect are available. In such a scenario, ear recognition is a valuable amendment to facial descriptors that enables forensic experts to use all the evidences available in the probe image.

As soon as a list of the most likely suspects is available, evidence may be manually collected with photogrammetry or superimposition of tracings. In photogrammetry, we measure the precise distances between landmarks in a pair of images. The analysis must ensure that only identical points are compared and, in case of pose variation, two images with the same pose are needed. If a 3D model of the subject is available, we could also use re-rendered view of the model. In some cases it may also be possible to compensate slight pose variations by applying affine transformations to the reference image. For superimposition of tracings, the outlines of the ear are extracted and then fitted onto another image (presumably from the database or another crime scene). Subsequently, the analyst checks how well the outlines of the two images match. When analyzing face images with this method, the analyst can also investigate the symmetry two half faces from two different images [140]. The techniques described above currently are mostly applied on facial images, but may - in principle - be used for any type of imagery, including ear images.

1.3 Goals of This Work

This work aims at exploring new techniques for 2D and 3D ear recognition. We focus on, but are not limited to forensic identification from CCTV footage. Instead of 2D mugshots, we assume that police station have full 3D head models stored in their forensic databases.

With this background we investigate possibilities to combine 2D and 3D information with the goal of increasing the performance of ear recognition systems with respect to the segmentation accuracy, normalization accuracy and recognition rates. We combine 2D and 3D information (rendered depth images) by exploiting the fact that depth and texture information are co-registered in rendered views of the 3D model and propose different ways of how these information channels can be combined. In order to measure the virtues of combining depth and texture information, we compare the performance rates of our algorithm with the performance accomplished with 2D data or 3D data only. We further analyze the statistical properties of fixed length histogram features and propose a generic method for creating binary representations for a more efficient search technique. We apply these binary feature vectors in a sequential search approach, where the binary feature vectors are used

for creating a short list of the most likely candidates and the real-valued features are used for refining the search within the short list. An additional focus is set on the impact of image quality (*i.e.* blur and noise) on segmentation and recognition performance. Finally, we explore the suitability unsupervised clustering for classification of fixed length histogram features.

The goals of the thesis are can be summarized with the following research questions:

Q1: How can the outer ear be automatically detected from 2D and 3D images?

Q2: How can cropped ear images be normalized with respect to rotation and scale?

Q3: Is it possible to combine 2D and 3D data in order to obtain a better descriptor that yields a better performance than 2D or 3D alone?

Q4: How can ear templates be represented in order to enable fast search operations in large datasets?

Q5: Which impact does signal degradation have on the performance of ear recognition systems?

Q6: Is it possible to automatically find categories of ear images?

As an extension to our research results, we develop a demonstrator ear recognition module that is part of a multi-modal face and ear recognition system. This system is evaluated and tested using a challenging dataset that is collected and provided by forensic experts from the German criminal police. This dataset is comprised of 3D models as reference data. Mugshots and CCTV videos are used as probe data. The dataset represents a typical scenario in forensic identification, where an unknown subject is to be identified from a video sequence. We explore the virtues and limitations of ear recognition in this scenario and point out future directions for forensic ear recognition systems.

1.4 Structure

This thesis is divided into three parts. In the remainder of this first part of the document, we will give an overview of the publications and contributions in the context of this work. Subsequently, we give an overview of the GES-3D project, which was conducted in the context of this work, including an explanation of system requirements, the image capture system, the workflow of our biometric service provider and some concluding remarks on the overall performance of the system.

The structure of the second part of the document roughly follows the general work flow in a biometric pipelines as proposed in the ISO/IEC SC37 SD11 standard document [89] (a brief summary can be found in the Appendix B). The structure of this thesis is also summarized in Figure 1.2. The figure will show up in each chapter in part II and is intended to guide the reader through the structure of this thesis and maintain a link between the single publications and the research questions (see previous Section 1.5).

We start with an elaborate overview of the state of the art. A brief update of this survey is given later in this chapter in chapter C. We start with the initial segmentation step. For segmentation, we propose a novel ear detection method, where depth and texture information is combined as expressed as a number of shape descriptors. We select the shape that is in the largest cluster of the best-rates shapes in the image. We also propose a sliding window technique using a circular detection window and evaluate it with respect to its robustness against rotations.

The segmentation step is concluded with a geometric normalization approach that does not rely on any symmetry constraints. We show that the outer ear can be normalized with this approach by measuring the recognition rates of a simple texture descriptor. We then move forward to the feature extraction step and present an evaluation of different texture

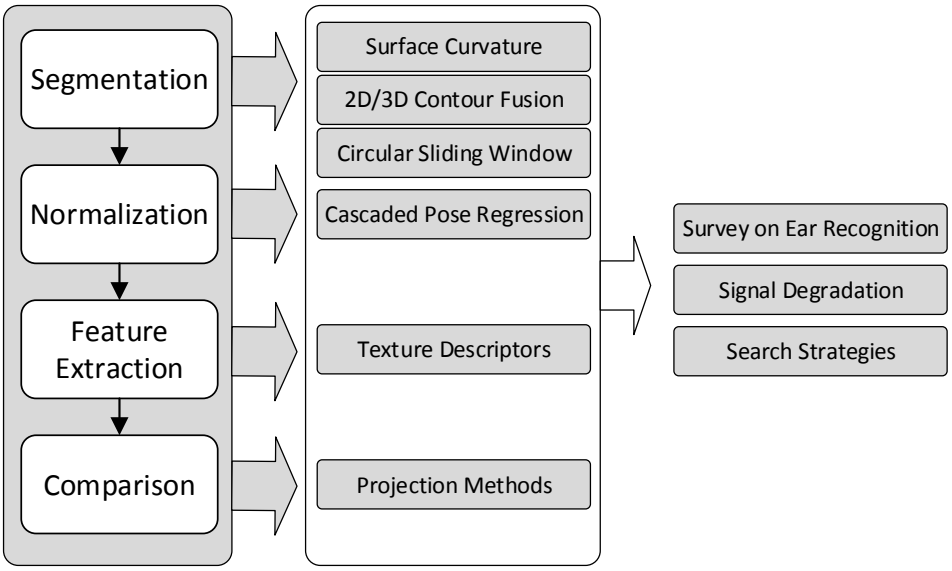


Figure 1.2: Illustration of the structure of this thesis. At the beginning of each chapter, we highlight one or several processing steps and the topics that are discussed.

descriptors in combination with selected subspace projection techniques. We benchmark the parameter sets for selected texture descriptors with three different datasets. Moreover, we propose a new descriptor that creates a fixed length histogram descriptor from surface and texture data.

The Chapters 9, 10, 11 and 12 of the thesis concentrate on applications and further investigations on the basis of the aforementioned ear recognition system. We first propose a binarization method for histogram features and then focus on the impact of signal degradation on the performance of segmentation and recognition with respect to noise and blur. Finally, we examine different texture feature spaces for clustering tendencies with the goal of providing an unsupervised classification scheme for ear biometrics.

The thesis is concluded with part III. This part summarizes the findings in part II and gives an outlook to future work and remaining challenges for 2D and 3D ear recognition.

1.4.1 List of Publications

1.4.1.1 Attached Research Articles

- [147]ANIKA PFLUG, CHRISTOPH BUSCH, Ear Biometrics - A Survey of Detection, Feature Extraction and Recognition Methods, IET Biometrics, Volume 1, Number 2, pp. 114-129

- [154]ANIKA PFLUG, ADRIAN WINTERSTEIN, CHRISTOPH BUSCH, Ear Detection in 3D Profile Images Based on Surface Curvature, International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP), 2012

- [155]ANIKA PFLUG, ADRIAN WINTERSTEIN, CHRISTOPH BUSCH, Robust Localization of Ears by Feature Level Fusion in Image Domain and Context Information, 6th IAPR International Conference on Biometrics (ICB), 2013

- [146]ANIKA PFLUG, PHILIP MICHAEL BACK, CHRISTOPH BUSCH, Towards making HCS Ear detection robust against rotation, International Carnahan Conference in Security Technology (ICCST), 2012
- [148]ANIKA PFLUG, CHRISTOPH BUSCH, Segmentation and Normalization of Human Ears using Cascaded Pose Regression, Nordic Conference on Secure IT Systems (NordSec), 2014
- [150]ANIKA PFLUG, PASCAL N. PAUL AND CHRISTOPH BUSCH, A comparative Study on Texture and Surface Descriptors for Ear Biometrics, International Carnahan Conference in Security Technology (ICCST), 2014
- [186]JOHANNES WAGNER, ANIKA PFLUG, CHRISTIAN RATHGEB, CHRISTOPH BUSCH, Effects of Severe Signal Degradation on Ear Detection, 2nd International Workshop on Biometrics and Forensics (IWBF), 2014
- [153]ANIKA PFLUG, JOHANNES WAGNER, CHRISTIAN RATHGEB AND CHRISTOPH BUSCH, Impact of Severe Signal Degradation on Ear Recognition Performance, Biometrics, Forensics, De-identification and Privacy Protection (BiForD), 2014
- [152]ANIKA PFLUG, ARUN ROSS, CHRISTOPH BUSCH, 2D Ear Classification Based on Unsupervised Clustering, In Proceedings of International Joint Conference on Biometrics (IJCB), 2014
- [151]ANIKA PFLUG, CHRISTIAN RATHGEB, ULRICH SCHERHAG, CHRISTOPH BUSCH, Binarization of Histogram Models: An Application to Efficient Biometric Identification, Conference on Cybernetics (CYBCONF), 2015

1.4.1.2 Additional Research Articles

- [38] CHRISTOPH BUSCH, ANIKA PFLUG, XUEBING ZHOU, MICHAEL DOSE, MICHAEL BRAUCKMANN, JÖRG HELBIG, ALEXANDER OPEL, PETER NEUGEBAUER, KATJA LEOWSKI, HARALD SIEBER, OLIVER LOTZ, Multi-Biometrische Gesichtserkennung, 13. Deutscher IT-Sicherheitskongress, 14-16 May 2013
- [162]R. RAGHAVENDRA, KIRAN B. RAJA, ANIKA PFLUG, BIAN YANG, CHRISTOPH BUSCH, 3D Face Reconstruction and Multimodal Person Identification from Video Captured Using a Smartphone Camera, 13th IEEE Conference on Technologies for Homeland Security (HST), 2013
- [149]ANIKA PFLUG, DANIEL HARTUNG, CHRISTOPH BUSCH, Feature Extraction from Vein Images using Spatial Information and Chain Codes, Information Security Technical Report, Volume 17, Issues 12, February 2012, pp. 26-35
- [77]DANIEL HARTUNG, ANIKA PFLUG, CHRISTOPH BUSCH, Vein Pattern Recognition Using Chain Codes, Spacial Information and Skeleton Fusing, GI-Sicherheit, 2012

1.5 Contribution

The following questions were derived from the requirements of the GES-3D project on the one hand and from questions that occurred during the ongoing research on the other hand. These research questions mark the red line though this thesis.

Q0: What is the current state of the art in ear recognition?
(Addressed in Chapters 3. Also see[147])
In preparation of defining a number of research questions, an elaborate survey on the state of the art in ear recognition has been compiled [147]. The survey consists of three main

parts. In the first part, we give an overview over the publicly available datasets that can be used for evaluating ear recognition systems. Then we describe different approaches for ear detection in 2D and 3D images and compare their detection accuracy. We move on to ear recognition systems and give a complete overview of different approaches and their recognition performance. This survey serves as the basis for the further work in this thesis. A summary of more recent work on ear recognition that was published after the publication of the literature survey is provided in appendix C.

Q1: How can the outer ear be automatically detected from 2D and 3D images?
(Addressed in Chapters 4, 5 and 6. Also see [154], [155], [146])
A reliable segmentation is important for any recognition system. In this work, we explore the geometrical features of the outer ear and how they can be used for segmentation. Special focus is set on the fusion of texture and depth information and the robustness to pose variations.

For the detection of 2D ears, state of the art techniques from face recognition, such as the Haar-like features [183] yield satisfactory segmentation accuracy as long as the capture setting is controlled carefully and the image quality is sufficient. High detection accuracies can also be achieved with LBP (original implementation as suggested by Ahonen *et al.* [9]). It is also possible to detect the outer ear with Cascaded Pose Regression [148] from coarsely segmented images (*i.e.* face profile images). In order to make sure that ear detection in 2D image performs well enough, the ear should not be smaller than 50×80 pixels.

Ears in depth images (3D) can be segmented by searching for the unique surface structure in the ear region. In [146] the ear detection approach by Zhou *et al.* [218] is extended to detect the outer ear in 3D profile images under different in-plane rotations. Due to the projection of the ROI from cartesian coordinates to a polar coordinates, the detection accuracy drops.

In [154] we have introduced a novel ear detection method, in which we reconstruct the ear outline by combining regions with high surface curvature to an ear outline. This work was extended in [155], where we added edge information from the co-registered texture image to the reconstructed 3D shapes. The high detection performance confirms that the surface structure and texture information in the ear region is clearly distinguishable for the surrounding areas.

Q2: How can cropped ear images be normalized with respect to rotation and scale?
(Addressed in Chapter 7. Also see [148])
In order to apply ear recognition in more challenging environments (as in GES-3D), the ear region needs to be normalized with respect to rotation and scale.

Cascaded Pose Regression (CPR) is an approach for face normalization, originally propose by Dollar et al [62]. CPR optimizes a loss function that tries to minimize the difference between local grey level-based features within the ellipse. Instead of localizing a number of visually defined landmarks, CPR uses weak features for estimating the orientation of the ear. Given that we have a sufficient number of training images, CPR can also be optimized towards being robust to partial occlusions and for normalizing ear images in different poses.

Using Cascaded Pose Regression (CPR), we fit an ellipse around the ear, where the major axis of the ellipse represents the largest distance between the lobule and the upper helix [148]. We then compensate scale and rotations by adjusting the center, the length of the major axis and the tilt of the ellipse such that the major axis is vertical and has a fixed length.

We show that the recognition performance of a pipeline using CPR prior to extracting the feature vector, is significantly higher than the same pipeline without normalization. We also show that CPR crops the ear region accurately for different ROIs representing different capture settings. Obviously, the benefit of using CPR increases with a larger variation

rotation and scale in the dataset.

Q3: Is it possible to combine 2D and 3D data in order to obtain a better descriptor that yields a better performance than 2D or 3D alone?
(Addressed in Chapter 8, Also see [146])
Motivated by the performance increase with fused texture and depth data in ear segmentation, we propose a combined descriptor for co-registered pairs of texture and depth images. We consider the texture image and the depth image as separate channels of information that are be merged into a fixed-length histogram descriptor. The optimal settings for the merged feature vector are determined in a series of experiments using three different datasets.

The combined 2D/3D descriptor uses surface curvature information for determining the histogram bin and texture information for determining the bin magnitude. The method can be applied within a sliding window, which results in a spatial histogram. Our experiments show that the method has some potential. However, it is vulnerable against noise, especially in the depth channel.

Along with this, we conduct a study on different techniques for texture description and empirically determine the optimal algorithm and parameter settings for the capture settings represented by three publicly available datasets. We conclude that the optimal parameter set for each of the texture and surface descriptors is highly dependent on the resolution and the quality of the input images.

Q4: How can ear templates be represented in order to enable fast search operations?
(Addressed in Chapter 9, also see [151])
Given that we have a fixed length histogram descriptor that yields satisfactory performance in a given scenario, we would like to optimize search operations towards being as fast and reliable as possible. Based on the observation that many histogram descriptors are sparsely filled, we propose a sequential identification system (1:N search) that uses binary descriptors in the first stage and real-valued descriptors in the second stage (*i.e.* with double precision numbers).

In our test system (implemented in C++), the comparison of binary feature vectors is up to ten times faster than the comparison with real-valued feature vectors of the same length.

Obviously, there is a loss of information during the binarization process, which results in a lower true positive identification rate (rank-1 recognition rate). Despite this, the probability that the correct identity is among the first n subjects is high, such that we can use the binary feature vector for retrieving a short list from the database. Subsequently, we use the real-valued feature vectors for re-sorting the short list.

We show that we can do an 1:N search in 30% of the time compared to an exhaustive search using the real-valued feature vectors only. Additionally, we reduce the chance for false positives.

Q5: Which impact does signal degradation have on the performance of ear recognition systems?
(Addressed in Chapters 10 and 11. Also see [186],[153])
The performance of every biometric system is dependent on the quality of the input images. The quantification of image quality, however, is always dependent on the scenario. We have conducted a series of experiments to learn more about the impact of noise and blur on the performance of ear recognition systems.

We have generated a series of degraded images and computed the Peak Signal to Noise Ratio (PSNR) in order to quantify the quality of the degraded images. The PSNR is defined via the mean squared error between an image of optimal quality and a degraded image and can be regarded as the ground truth information here). We then measured the decline of segmentation and recognition performance with different features for detection [186] and recognition [153]. In general, we noticed that noise has a greater effect on segmentation

and recognition performance than blur.

A similar series of experiments was conducted on images, that are compressed with JPEG and JPEG 2000 (also see Part III 13.1.4). JPEG 2000 - also named after the codec J2K - is a wavelet-based image compression method that, according to the ISO standard, should be preferred over the DCT-based JPEG compression for biometric identification systems. These experiments show, that the detection and recognition performance varies with a decreasing bit rate when compressing with JPEG. We suppose that there is a correlation between the size of the compression artefacts and the radius of the texture descriptors.

Q6: Is it possible to automatically find categories of ear images?
(Addressed in Chapter 12. Also see [152])
Most of the research efforts in ear recognition concentrate on achieving a high recognition performance in closed datasets. The next step towards an operational system is to provide techniques for fast and efficient search in large databases.

We analyse texture feature spaces with respect to cluster tendencies ,which could be exploited for reducing the number of candidates in an $1 : N$ search. We create feature subspaces using linear and non-linear methods and analyze these subspaces for cluster tendencies. We apply different metrics for estimating the goodness of the clustering solutions and show, that the feature subspaces can be organized as convex clusters using K-means.

We show that clustering using 2D ears using K-means, PCA for subspace projection and LPQ for texture feature is possible. For this particular configuration, the search space can be reduced to less than 50% of the database with a chance of 99.01% that the correct identity is contained in the reduced dataset. We also show that a search that is extended to up to three adjacent clusters yields a better performance than a single cluster search. The classes depend on the skin tone of the subjects, but also on the capture settings of the dataset they originally came from. We also observe that feature vectors with a high performance in classification do not necessary yield high recognition rates.

Chapter 2

The GES-3D Project

In our project GES-3D, we develop a demonstrator system for exploring the virtues and limitations of 3D imagery in forensic identification in a semi constrained environment. The project is conducted within a consortium of seven partners, among which are technical partners, consulting partners and the German Criminal Police (BKA) as the stakeholder.

The goal of the project is to develop a fully integrated identification system that uses 3D head models as references and crime scene videos as probes. The reference data is collected under controlled conditions by a police officer, who sends a 3D head model, and a series of photographs to a central database where the data is assembled and stored.

If a subject is to be identified from video material, the identification system automatically extracts face and ear features from a CCTV video and returns a list of the n most likely candidates. In practice, a forensic expert would now further analyze the retrieved images and give an estimation of the similarity between the reference and the probe. Manual analysis is not part of the project though. GES-3D only concentrates on the automated retrieval of the n most likely candidates. After the retrieval process, the 3D model from the database can be used to assist the forensic expert, by offering the opportunity of adjusting the pose of the reference according to the pose on the input video.

For evaluating the system, we develop a test scenario, in which we simulate a typical scenario in the entrance hall of a bank. A person enters the room and walks towards an ATM. Whilst inside the bank, the subject is filmed by different off-the-shelf CCTV cameras from four different viewpoints. The selection of the viewpoints was proposed by BKA, based on their experience with typical surveillance cameras. Figure 2.1 shows a floor plan that illustrates the data collection setup. In existing identification systems and in many research experiments, reference and probe images are collected under the same conditions with the same capture device. One of the main challenges in this project is to compare images from different media (2D, 3D and video), different capture devices and different camera viewpoints. Based on this experiment, we formulate the following requirements for the prospect demonstration system.

- **Usability:** The user interface and the capture device should be easy to use and to understand for police officers, who do not have any image processing or biometrics background.
- **Data protection:** For data protection reasons, all images should be stored in a central database and they should not be distributed to any other third party system.
- **Transparency:** All decisions that are part of of the identity retrieval should be made transparent to the forensic investigator. The forensic investigator should be able to review the search result for every single biometric service provider.
- **Throughput and accuracy:** The search result should be delivered within a reasonable amount of time (several minutes). However, accuracy is still more important than throughput. A lower false accept rate (the probability that the wrong subject gets a high rank in the candidate list) is preferred over a lower false reject rate (the probability that the true perpetrator does not show up in the candidate list), because a false accept would imply that an innocent subject could be accused for a crime.

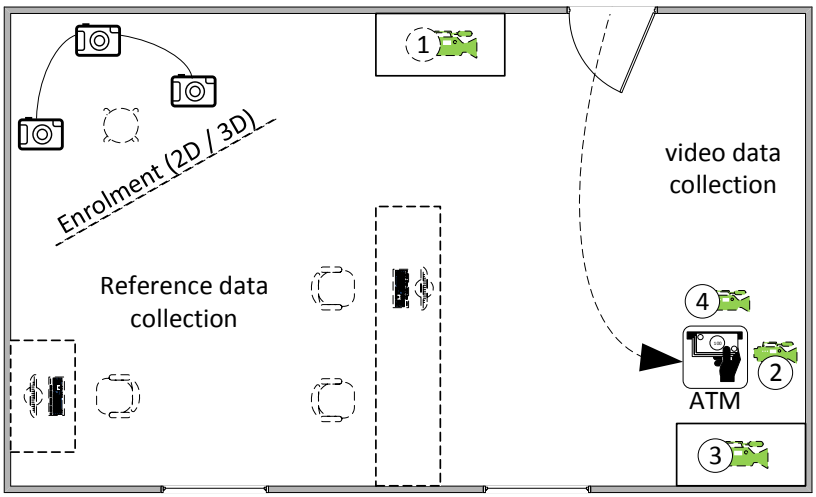


Figure 2.1: Floor plan for the data collection in GES-3D. Reference data is collection with a special setup that consists of 3 depth cameras (left). Probe data is collected by four CCTV cameras.

2.1 General Requirements

GES-3D is a project where partners from research and industry collaborate with the goal of creating a demonstrator system. In order to protect the interests of the industry partners, it is important to make sure that their software can be integrated into the system as a black box. Concurrently running modules should run independently from each other and at the same time, the interface should leave as much freedom as possible for the individual solutions of each project partner (*w.r.t.* algorithms, libraries and programming languages).

In GES-3D, the exchange of data is implemented with a proprietary internal interface. Each project partner delivers a back-end module that provides the functionalities that are specified in this interface (see Figure 2.9). We distinguish between the data capture subsystem [89] and the system back-end. The capture system is operated by a police officer and used for enrolment. The retrieval system is operated by a specialized department of the criminal police. Both parts of the system connect to a central web service, by implementing the previously mentioned interface. Figure 2.2 illustrates the two parts of the system. For Enrolment, we obtain a series of mugshots and a 3D model (left part) and for identification, we obtain one or several videos (right part). The search function also accepted single video frames or images.

The prospective end user of the GES-3D system is the German Federal Criminal Police Office (BKA). BKA is operating the prospect identity retrieval system and will also deploy the image capture system to the local police stations. The process of enrolment is defined by national and international standards and follows a strict protocol [134, 33]. It is crucial that the system fully reflects these standards and seamlessly integrates into the existing work flow. The 3D capture system also collects a series of mugshots (see next section for details) from different poses in addition to the 3D head model. This allows a smooth migration from the existing identity retrieval system (GED-2D) to the new system.

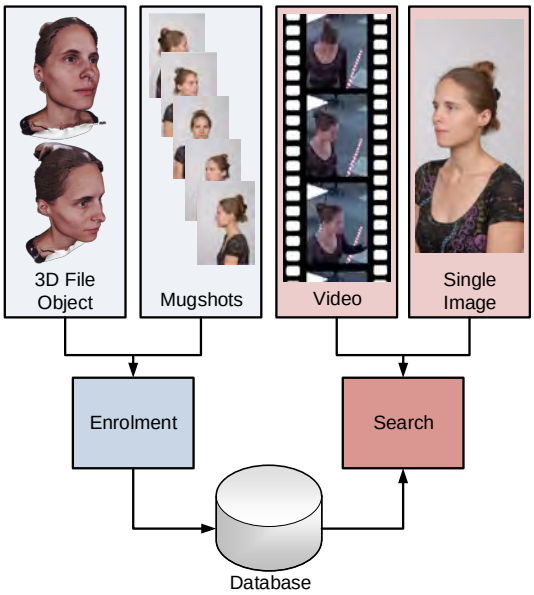


Figure 2.2: Division of tasks with associated media types in the proposed forensic identification system.

2.2 Image Capture System

Interpol and national police agencies in Europe and the US [134] have adopted this standard. The standard is designed to enable the exchange of images between national and international police stations. It is also needed for establishing a quality standard of the images in the forensic databases. According to the standard, it is recommended that at least one frontal image and two profile images with 45 and 90 degrees of pose variation are captured. The profile images should show the left and the right ear. Further, it is recommended that the focal length of a camera should be between 85 mm and 135 mm and that the mid-points of the mouth and of the bridge of the nose should be centered. Shadows should be minimized by using at least two sources of diffuse light. We follow these recommendations and use a data set with five images which represent the previously described poses. The images are captured with a high resolution digital camera under the described environment for each subject in the enrolment and compressed with JPEG compression.

In order to maximize the probability of identifying a subject, the quality of the input media for enrolment should be as high as possible. In particular, there should be no compression artefacts in the 2D images. The 3D models should be as detailed as possible, while containing a low noise level and a minimum of reflection artefacts. 3D face imaging is well proven and tested in various scenarios. 3D ear imaging, however is a new application, which poses some challenges to the arrangement of the capture devices. On one hand, it should be possible to obtain high quality face images and on the other hand, the angles between capture devices should be arranged in a way that we obtain a complete view of the concave and self-occluding surface of the ear.

For obtaining a high quality representation of the outer helix, the capture protocol was designed accordingly. We capture two partial ear models, one with a full profile view and another one that shows the ear from behind at approximately ± 120 degrees (see the two left sub figures in Figure 2.4 for an illustration of these two camera angles). This allows us to obtain a smooth representation of the outer helix. We also get information on the perturbation of the ear.

Figure 2.3: Example images for a collection of mugshots, as they are collected during the enrolment process.

Figure 2.4: Different views of an example for an interpolated 3D head model, as it is collected during the enrolment process with a close-up view of the ear.

2.3 Dataset

The collection of a dedicated dataset was an objective of the GES-3D project. The data was collected in accordance with the standards and recommendations described in the previous section. The GES-3D corpus contains 300 distinct subjects, from which 150 are males and 150 are females. 113 subjects were wearing glasses and 98 subjects wore earrings. The age distribution ranges from 20 until more than 60 year old subjects.

All subjects were asked to enter a room, where an ATM was placed next to a wall. The dotted line in 2.1 represents the expected path of the subjects from the door to the ATM. The room is also equipped with four CCTV cameras, which are marked in green and with numbers 1-4 in Figure 2.1. One camera was placed next to the door, the second one was placed on top of the ATM, a third cameras was placed at the opposite side of the room with respect to the door and a fourth camera was placed next to the ATM's screen. Each video camera is running at a resolution of 1920x1080 pixels and captures 30 frames per second. Example images for each of the camera viewpoints are shown in Figures 2.5, 2.6, 2.7 and 2.8.

The 3D images were captured with a setup of three viSense2 depth cameras (upper left corner in Figure 2.1). The camera uses structured light for capturing the surface of an object at a scanning speed of 0.25 ms. The diameter of the frustum for the 3D acans is 782x582 pixels and the resolution of 2D camera for capturing the texture data is 1624x1234 pixels. Each depth camera is connected with several flashes in order to assure optimal illumination settings. Drop shadows are minimized by using flash diffusers. The partial depth images are registered semi-automatically by an operator. After capturing all input images the operator is asked to annotate three corresponding landmarks in each partial model. These landmarks can be chosen arbitrarily by the user as long as she is able to coarsely point to the same landmarks in two corresponding 3D models. It is recommended to use landmarks that can easily be found in every model, such as the nose tip, the outer corner of the eye

or outer corner of the lip. The registration algorithm performs a coarse alignment based on these manually marked points followed by an automatic fine alignment of the surfaces using Iterative Closest Point Search (ICP) [27]. The output of this is a full 3D model of the subject's head (see Figure 2.4 for an example).

The 2D mugshots are collected simultaneously with the partial 3D models. For each subject, we obtain a left and right full profile view, a left and right half profile view and a frontal view. The mugshots are taken with a high resolution digital camera using JPEG compression (1624x1234). The enrolment data collection setup is separated by a thin wall with a back cloth, in order to obtain mugshots with a uniform background. An example for the mugshots is shown in Figure 2.3.

The reference images are stored together with some meta data, such as a unique identifier for the subject, a unique identifier for the capture sessions, and a time stamp for the 2D and for the 3D images. In our evaluations, we use data from 200 randomly chosen subjects. The remaining 100 subjects are reserved for black box testing.

The collection, storage and processing of the data is subject to the German data and privacy protection regulations, which state that the data may not be published without the explicit consent of each subject and the data may not be used for any other purpose but the GES-3D project. In order to maintain reproducibility and compatibility of our results, we decided to use public datasets rather than the GES-3D data in our scientific publications.

2.4 System Back-End

From a technical perspective, the system back-end consists of five different subsystems, which are the biometric database, the back-end provider module, several biometric service providers (encapsulating the algorithm-specific feature extraction and retrieval methods), the fusion service provider, a connection for the biometric capture device and a module for handling search requests (see Figure 2.9). Together, these subsystems follow the work flow specification for a general biometric system as specified in ISO/IEC SC37 SD11 [89]. The central component is the back-end provider where all intercommunication between the modules is processed. The middleware JBoss Application Server is used to connect the various interfaces between the different modules from the project partners.

For enrolment, the input data is read from the capture device and passed over to each biometric service provider. Triggered by this event, each provider initiates the enrolment process and returns a feature vector. The back-end provider then stores each feature vector in the biometric database, where it is available for future search operations. The processing of the data in our module is written in MATLAB, OpenCV and Java EE with additional external libraries.

The overall system is supposed to mimic the central server that is run by the criminal police. It takes images from the capture device, sends them to the biometric service providers for feature extraction, stores templates and performs search operations upon request. For data retrieval, video sequences are sent to the service provider. Depending on the pose and possible occlusions, the face and ear region are segmented and a template is generated from these regions. Because the database contains a complete 3D model of the head for each subject, pose variations in the input images can be compensated during the comparison process.

Each biometric provider generates a proprietary template. The algorithms for enrolment are developed independently from each other and the extracted feature vectors reflect different properties of the image. A separate fusion component combines the results from each provider and generates a ranked list of identities. The list is sorted according to a descending likelihood that the subject in the probe and the reference sample have the same identity.

Figure 2.5: Example images for camera viewpoint 1. The camera was placed next to the entrance door.

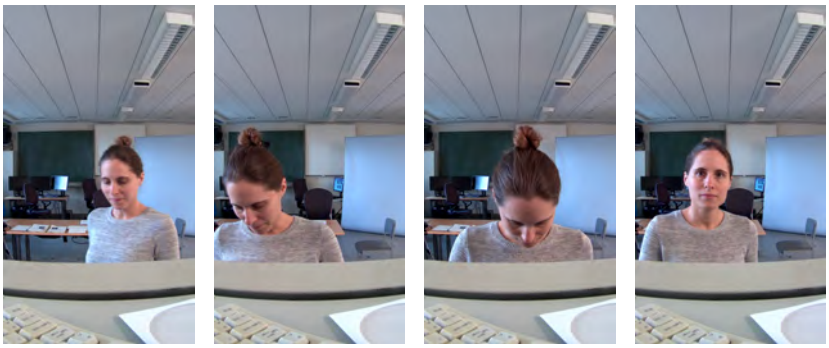


Figure 2.6: Example images for camera viewpoint 2. The camera was placed on top of the ATM

Figure 2.7: Example images for camera viewpoint 3. The camera was placed at the opposite side of the room with respect to the door.

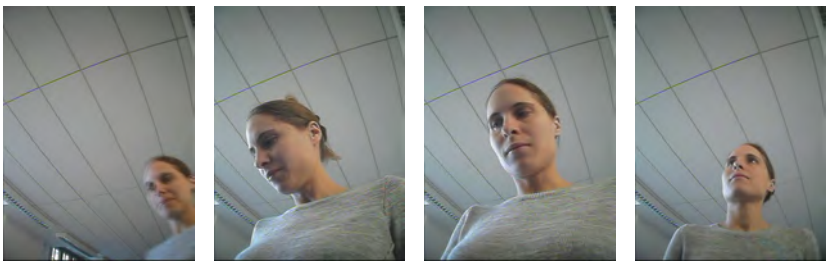


Figure 2.8: Example images camera viewpoint 4. The camera was placed next to the screen of the ATM.

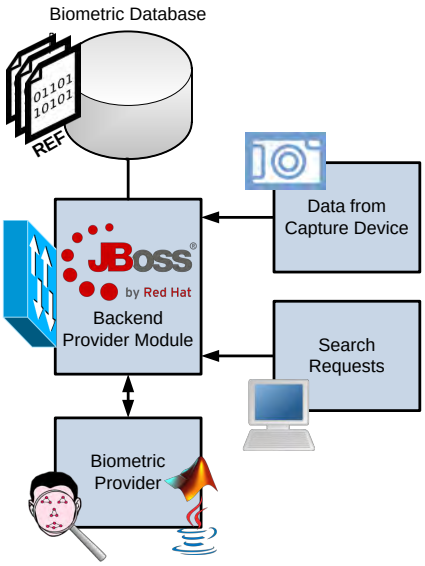


Figure 2.9: Overview of the architecture of the GES-3D identification system with connections to the database back end, biometric service providers and the capture device.

2.5 Workflow of a Biometric Service Provider

A biometric service provider is connected with the back-end provider via a WSDL-API. It combines functionality from OpenCV and MATLAB code with Java and can be deployed on a JBoss 7.1 compatible web service. The WSDL file defines method stubs for the enrolment and the search method, as well as the exchange formats for input media and search results. The interface also provides functionality for retrieving data from the database via the back-end and specifies a unified format for templates and search results. The data retrieval and database access is handled by the back-end provider module. All biometric service providers automatically obtain the same probe data and have fully transparent access to the central database via the back end provider module. Additional libraries are specified in a Maven file and linked to the JBoss environment.

A general overview of the workflow of our provider showing the combination between OpenCV, MATLAB and Java is illustrated in Figure 2.10. For an illustration of overall the segmentation, preprocessing, feature extraction and feature vector generation process, please also refer for Figure 2.13.

2.5.1 Enrolment

The enrolment task takes at least one input medium and returns one feature vector per input medium to the back-end module. Possible media types for enrolment are a series standardized photographies, as described in Section 2.2 or a 3D head model.

2.5.2 Segmentation

The method for ear segmentation depends on the type of the input media. Possible media are a series of photographs, a video stream or a 3D head model.

Series of mugshots: The preprocessing module segments the ear region from the half and full profile images using the Viola-Jones detection method [182] with the implementation from OpenCV (this implementation uses the set of wavelets suggested in [182]). We trained our own haar cascade for detection using manually cropped positive images from the

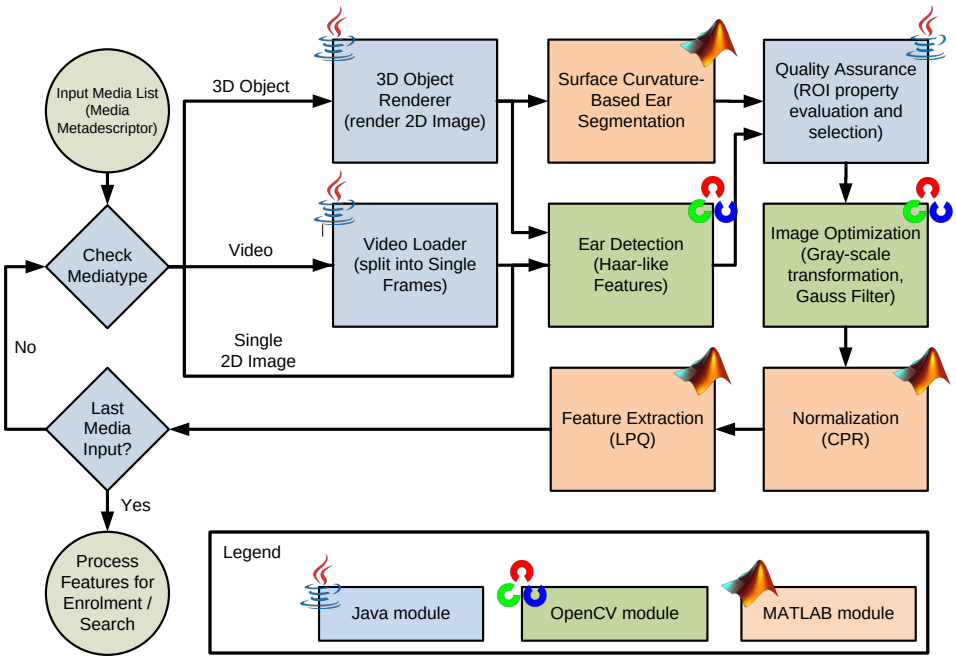


Figure 2.10: Workflow diagram of our biometric service provider. The workflow illustrates the interfaces between different programming languages.

GES-3D training set and negative images that contain randomly cropped parts of the background from the GES-3D scenario. The segmentation step is finalized by a post-processing step, where the most likely region of interest (ROI) is selected from each image. This selection relies on the fact that the process of capturing the image series is highly constrained. These constraints allow us to make assumptions about the ROI size, aspect ratio and its position in the image. Using these assumptions, we check the returned ROIs for plausibility. A ROI is only selected if it complies with the assumptions on size, aspect ratio and position. Otherwise we consider the ROI as a false positive and discard it.

Video streams: For video streams, we first detect a profile face view, as shown in the example image the upper row of Figure 2.13. This is done in two steps: (1) a face profile detector selects frames that contain profile views within a given sub region of the video frame. This constraint is based on the assumption, that the CCTV camera does not move and that each subject, who wants to use the ATM is standing at the same distance from the camera. (2) For each successfully detected profile face, we now run an ear detector within the extended profile face region. Although this means that we discard some frames that could have been used for ear recognition, a sufficiently large number of true positives is left. In case the outer ear is visible, we select the ear region from the profile face ROI. If no ear can be detected, the entire frame is discarded.

3D head models: If the input media is a 3D model, we render pairs of depth and texture images for discrete camera viewpoints. Examples for the left and the right side view of the model are shown in Figure 2.11. The renderer assumes that the null pose of the model is normalized in such a way, that the normal of the nose tip points towards the camera (see Figure 2.11: left column).

For each rendered pose, we segment the ear region by fusing information from the texture and from the depth channel (see Chapter 5).

Figure 2.11: Rendered image with null-pose (left), right pose with a 90 degree rotation (middle) and left pose with a -90 degree rotation (right) from a 3D head model - the left part of the figure shows texture channel and the right part shows the corresponding depth data.

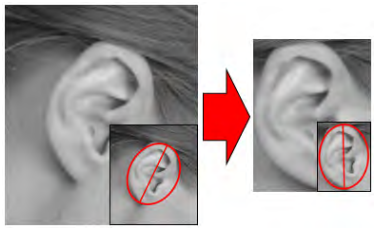


Figure 2.12: Example of a segmented ROI and the result after normalization with CPR.

2.5.3 Normalization

After segmenting the ear region, the orientation of the ear needs to be normalized with respect to rotation and scale. As shown in [148], the correction of these rotations is important for the recognition accuracy, because a misaligned ear will be likely to cause wrong decisions in the comparison process. For the same reason, hair or skin around the ear should be removed before extracting features.

In the normalization step, we use Cascaded Pose Regression (CPR) for removing the margin and to compensate for small rotations [148, 62]. CPR is used for fitting an ellipse on the ear, where the major axis connects the two far-most points of the lobule and the outer helix with each other. We then rotate the ROI, such that the major axis is vertical and then cut off all pixels that are outside of the enclosing rectangle of the ellipse. This leaves us with images, with a minimal margin around the ear and where all ears are oriented in exactly the same way. Finally, all images are resized to 100×100 pixels. This facilitates the subsequent feature extraction step and has no impact on the system accuracy. For video images, we also try to estimate the head pose using a series of tree structured models with a shared pool of landmarks [222]. For pose estimation, the algorithm tries to find the best tree for the given image and returns the corresponding pose. As a side effect, the tree structured model can be used for assuring the quality of the region of interest for the ear region, because it allows us to check whether the position of the ear is plausible.

2.5.4 Feature Extraction

After normalizing the input images, we apply Local Phase Quantization (LPQ) [10] for obtaining a fixed length histogram of local texture features. The concept behind LPQ [10] is to transform the image into the Fourier domain and to only use the phase information in the subsequent steps. For each pixel in the image, we compute the phase within a predefined local radius and quantize the phase by observing the sign of both the real and the imaginary part of the local phase. Similarly to LBP, the quantized neighbourhood of each pixel is encoded as an 8-bit binary string and stored in a code image. The code image hence

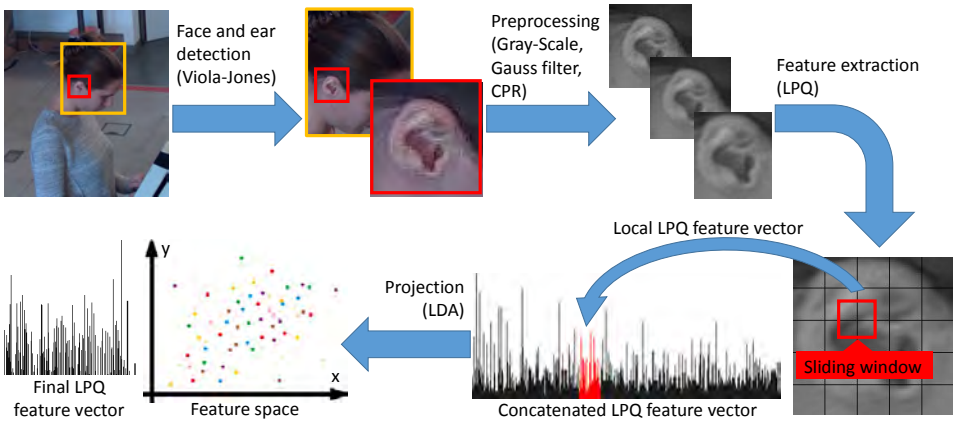


Figure 2.13: Illustration of the single processing steps for computing the features of an image (frame) from a 2D video stream. Processing of other input media only differs in the segmentation step.

contains values between 0 and 255.

We divide the image into equally sized sub-windows and compute the LPQ code image for each window. Every code image is then represented as a histogram with 256 bins. The histograms for all local windows are finally concatenated in order to obtain the overall feature vector. For a 20×20 window size and an overlap of 10 pixels, this results in a 16384 fixed-length dimensional feature vector.

For removing redundant data from this large descriptor, we project the histogram into a lower-dimensional space using Linear Discriminant Analysis (LDA). The projected histogram in LDA space is the final feature vector. This feature vector is stored in a database along with some meta data, such as the pose of the enrolled image, the identifier of the subject, the identifier for the capture process and the input media type. The lower row of Figure 2.13 illustrates the feature extraction process and the feature subspace projection.

2.5.5 Search

The search task takes either a single 2D image or a video sequence and returns a ranked list of identities, where a lower list rank indicates a lower likelihood that the subject in the probe video and the reference are from the same source.

If the input medium is a video, we obtain as many feature vectors as we find video frames with an ear. Each feature vector is compared with the each of the reference images in the database. For comparison, we use a NN-classifier with cosine distance. For each feature vector from the input media, we obtain a list of distances between the feature vector and all the templates in the database. Each of these lists with distances is sorted in ascending order. We only retain the lowest n distances, where n is specified by the user when initiating the search.

Let there be m sorted lists with length n. We fuse these lists by counting the number of occurrences at given ranks for each identity. Note that we distinguish between *identities* and *images*. We may have several images with the same identity (i.e. showing the same subject), and each identity is represented by at least one image. For each valid ROI in the input media, we obtain a sorted list of possible candidates. We iterate through each sorted list and assign a score to each identity according to the position in the list. Identities with higher ranks obtain a higher score than identities with lower ranks. After assigning a score to each identity in each list, we create a fused list by summing up the scores for each identity. Finally we sort the identities in the combined list according to the score in descending order and finally return the n identities with the largest score.

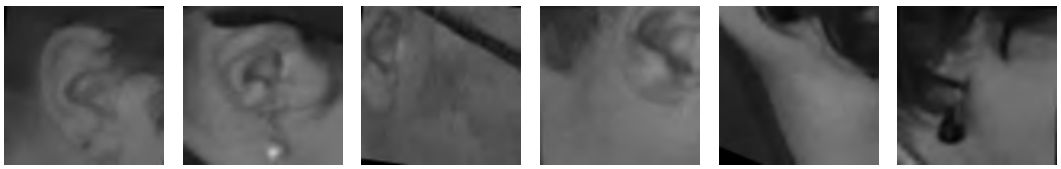


Figure 2.14: Example images for Quality levels 1 (left) until 6 (right) that serves as the basis for the evaluation of the detection accuracy. All of these examples are taken from the video stream.

2.6 System Performance Evaluation

The proposed ear recognition system is evaluated with the dataset described in Section 2.3. In our experiments, we focus on camera pose three, which gives us a profile view of the subjects, while they are using the ATM. Figure 2.2 illustrates the workforce of the system and shows an example for each media type.

We evaluate each step of the system separately, starting with the segmentation step for each media type. In a second experiment, we evaluate the accuracy of the pose estimation module and the third experiment we provide results for the recognition system. Finally, our last experiment provides results on the recognition performance of the complete system, including errors introduced by the segmentation, pose estimation and recognition step.

2.6.1 Ear Detection

The data that we obtain from our project partner is not labelled, such that we do not have a ground truth for evaluating the detection performance, such as in Chapters 4, 5, 6 and 7. We would also like to know more about the typical types of errors that we get and see whether there is any trend toward a certain type of error for a particular media type. We distinguish between six different quality levels for the region of interest (ROI), which are denoted as *L1*, *L2*, *L3*, *L4*, *L5* and *L6*. An example for each of the quality levels can be found in Figure 2.14. Quality levels *L3*,*L4*, *L5* and *L6* represent a failure to capture (FTC) and the failure to extract (FTE). In the following list, the quality levels are defined:

- **L1**: The ear is located in the center of the ROI and the pose is a full profile view.
- **L2**: The ear is fully covered by the ROI, but it is either not centered or the image is off-pose.
- **L3**: Parts of the ear are cut off or occluded, but more than half of the ear is still visible.
- **L4**: Major parts of the ear are cut off. The data is not sufficient for being used in the ear recognition system and should be dropped.
- **L5**: Major parts of the ear are occluded. The data is not sufficient for being used in the ear recognition system and should be dropped.
- **L6**: Something else has been detected (False positives)

The test set for the videos contains 5566 images (single video frames) from 200 subjects. The test sets for the mugshots and for the rendered 3D images contains 400 left and right ears from the same 200 subjects.

For the 3D images and the 2D mugshots, the quality level probabilities are similar, because the capture settings are following the same constraint. For the video sequences, the probabilities for a given quality level differ significantly.

The portion of correctly segmented ears ranges from 40% to 60% of the mugshots and rendered 3D images, which is still unacceptably low, especially for the enrolment set. The

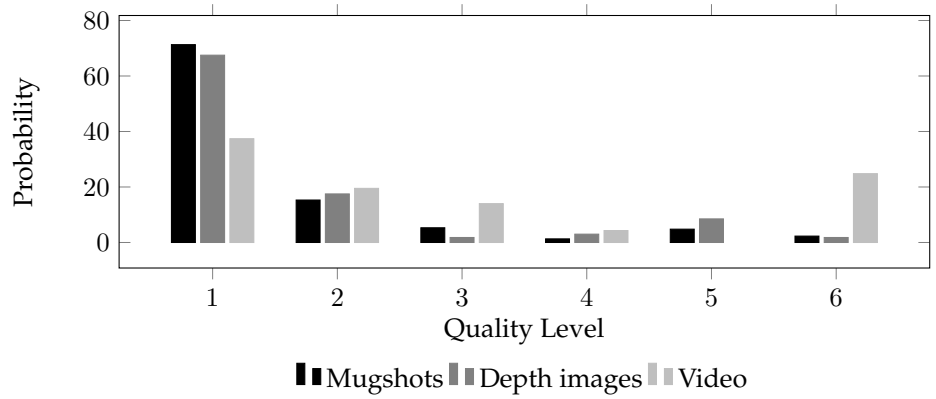


Figure 2.15: Summary of the probability of different segmentation quality levels for each media type.

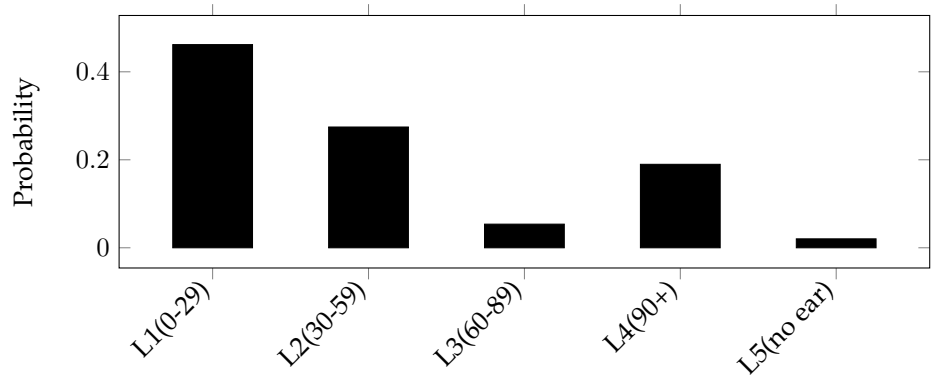


Figure 2.16: Overview of the probability of pose estimation errors in pose estimation from L1 ROIs.

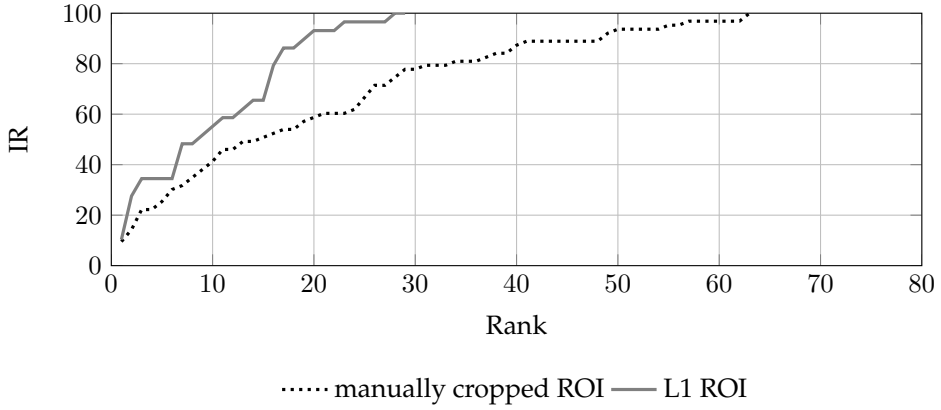


Figure 2.17: Cumulative match characteristic (CMC) of manually cropped ROI and ROIs that have been labelled with *L1* in the segmentation and the normalization step. The rank-1 identification rate for *L1* ROIs is 10.35% and the rank-10 identification rate is 55.17%.

large portion of Failure To Enrol (FTE, also see Appendix B) in the video stream can be compensated for by the high number of frames, such that we must discard a large number of images, but we have a sufficient number of frame. We obtain between between 10 and 30 frames per video that contain a correctly segmented ear.

The main reason for the poor performance is the small number of training images that we have at hand for training the detector for the mugshots and the video streams. During our experiments, we compared different detectors. One of them was the pre-trained haar-cascade from OpenCV [1]. We also trained a detector using the ear data from different publicly available ear datasets. A third detector was trained using positive and negative sample images from the test dataset. As expected, the latter detector outperformed the other two. The detection performance could be improved, if a sufficient number of training subjects for both scenarios would be available.

Further, the detection performance may suffer from the low resolution and high noise in the video streams. The resolution of the video frames could, for instance, be improved by applying a super resolution technique [76].

In the case of the 3D segmentation, we observe a large number of cases where parts the ear are cut off in the ROI. This is a typical limitation of the segmentation algorithm, which we already observed in previous experiments (see chapter 5). We could be able to minimize the number false positive detections by optimizing the parameters of the algorithm to the scale and resolution properties of the test dataset. The parameters for the 3D detection algorithm could be adapted to the particular capture setting in order to improve the detection accuracy.

2.6.2 Pose Estimation

For evaluating the accuracy of the pose estimation step, we only use ROIs of quality level *L1* from the previous segmentation step. Hence, we only have images containing full profile views of a subject (-90 of +90 degrees yaw pose). This knowledge serves as our ground truth in this experiment.

Again, we distinguish between different classes of errors that are characterized by the difference between the estimated viewing angle and the actual viewing angle. In this experiment, we use a pre-trained model that is publicly available for download [2]. This model is optimized for the MultiPie dataset [3] and is supposed to work for face regions lager than 150×150 pixels. We also evaluated models with fewer parts, but we found that the best performance could be achieved with the previously mentioned model. This model is able to locate 41% of the ROIs, which means that we could not estimate a pose for the remaining 59% of the ear images. The results on the accuracy of the pose estimation attempts, where an ear is detected are summarized in Figure 2.6.1.

Similarly to the segmentation experiment, we must conclude that the accuracy of the pose estimation is not satisfactory. Even though we expect to find only full profile images, the pose estimator returns a different (and thus wrong) result in 55% of the cases. Based on these results, we decided not to use the pose estimator at all in the final system. We believe that the pose estimation could be improved, if we would train a detector using the video frames from our scenario.

2.6.3 Algorithm Performance of Ear Recognition

We evaluate the algorithm performance in two evaluation settings that differ by the selection of the input data. In the first evaluation setting we use manually cropped ROIs, which means that we do not have a segmentation error. The second evaluation is using input

[1] http://mozart.dis.ulpgc.es/Gias/MODESTO/DetectDemousingViolaJonesxmlfiles.zip
[2] http://www.ics.uci.edu/~xzhu/face/
[3] http://www.multipie.org/

data, that has been labelled with quality level *L1*. These images do hence contain a normalized image of the ear. We use 3D models as references and video data as probes. The CMC curves of the two evaluation settings are shown in Figure 2.17.

We observe that the performance of the two data sets is similar. The difference in the slope of the curve is only due to the smaller number of samples in the *L1* set. The recognition performance is lowered by differences in the pose and a low image quality of the ear ROIS. As already mentioned in the section on detection performance, super resolution could be applied in order to obtain ROIs with a higher quality (*i.e.* sharper edges, less blur though block artefacts).

When looking at the recognition performance, we must conclude that the high recognition rates from the baseline experiments (see chapter 8), could not be achieved with our test data and the GES-3D system. Further analysis of the system performance in the upcoming section discusses a number of different reasons for this.

2.6.4 System Performance

The system performance experiment treats the entire ear recognition system as a black box. We evaluate the system using videos or mugshots as references and 3D images as probes. We also provide a baseline performance for the existing system, where mugshots are used as references and video stream are used as probes. The error rates reported in this experiment are composed of the failure to enrol (FTE) and the false match/false non-match rate (see Appendix B).

We define the following three test cases. The performance rates for the test cases are summarized in Figure 2.18. These results could be reproduced in the black-box test on unseen data (the remaining 100 subjects), which was conducted by our project partner Fraunhofer IGD.

- **3D references and probes from video**: 200 rendered right ear images subjects from 200 subjects as reference and 200 video files with camera viewpoint 3 (one video per subject with 4 frames per second) as probe images. The remaining three viewpoints were excluded from the experiment, because the ear regions were too small (viewpoint 1) or the quality was reduced through interlacing (viewpoint 4) and motion blur (viewpoint 2).

- **3D references and probes from mugshots**: 400 rendered ear images from left and right ears of 200 subjects as reference and the corresponding 400 mugshot images from 5-part sets (also showing left and right ears of 200 subjects).

- **Mugshots as references and probes from video**: 200 right profile mugshots and from 200 subjects as reference images and 200 video files with camera viewpoint three (one video per subject with 4 frames per second) as probe images. This test case represents the typical search operation as it is done in current forensic identification systems.

In general, the system performance evaluation is consistent with the algorithm performance. The performance rates in all test cases are poor (close or equal to the coin toss), especially for test cases where the rendered 3D images are used as references. We conclude that segmentation errors play a minor role in the overall system performance, though they may have been more influential if the overall system performance was be higher. The GES-3D system is the first of its kind and may be regarded as a reference system for future research projects on the practical application of ear biometrics.

We also observe a noticeable difference between the CMC curve of the test case using 3D images as references (black curve) and the test case using the mugshots as references (dotted curve). This may have several reasons:

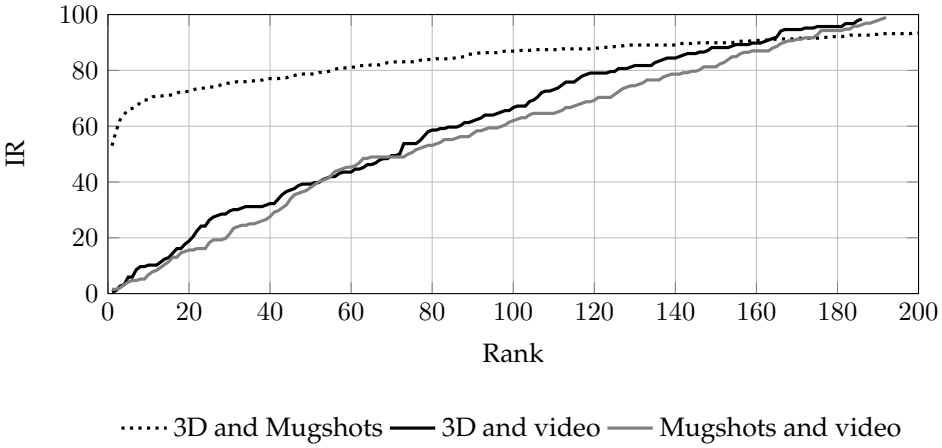


Figure 2.18: Cumulative match characteristic (CMC) of the system performance for different combinations of media types.

- **Normalization errors :** The normalization algorithm may have introduced a small alignment error. In conjunction with the local histogram model, that is used for comparison, these normalization errors have a considerable impact on the comparison scores and hence on the recognition performance.
- **Resolution and quality :** The video quality of a full frame from the surveillance camera has a high quality, however the distance to the subject is so big that the ear region is small. Due to this, the resolution of the ear images from the video is low, which leads to blurry and noisy images. Some frames also contain strong motion blur and compression artefacts. These degradations of the texture information significantly lowers the recognition performance. Rendered 3D models may also contain deficiencies from the 3D representations, such as missing surface patches.
- **Cross media / cross sensor comparisons :** Although the usage of 3D data as references has the potential to improve the robustness to pose variations, different imaging technologies for probe and reference images are used. This means that we have to deal with cross-sensor artefacts (including different image compression techniques). The consequence of this is that feature vectors from rendered 3D texture have different properties than feature vectors from the video stream. We must assume that the feature vectors contain a bias that is introduced by the properties of the imaging technique
- **Pose variations :** Pose variations remain a problem in the current system. The poses in the rendered 3D images and the video frames are slightly different and hence are an important factor that lowers the performance of the system. The impact of pose variations can be clearly quantified when comparing the performance rate of comparisons between mugshots and rendered 3D images on the one hand and 3D images and video frames on the other hand.

2.7 Conclusion

The performance rates obtained with the GES-3D datasets are far behind the biometric performance, we obtained using laboratory data (see 8). Although the term ”real-life data” is frequently used in academic literature, many of the described algorithms and systems are tuned towards the dataset and its underlying constraints. Our results clearly indicate

that the performance rates of ear recognition for a more realistic dataset are far behind the performance that is reported for academic data. Keeping in mind that ear recognition is a valuable amendment for the forensic analysis of facial images, this should be a motivation to continue working on ear recognition in surveillance scenarios. We are aware that our data is also collected in a specified laboratory setting and true "real-life data" is likely to be even more challenging. Compared to the performance achieved by the face recognition modules from our project partners, the ear recognition module is also behind of what is already possible for face recognition. Within the limitations in resolution and pose variations (self-occlusion), the ear recognition module in GES-3D could certainly be improved.

Ear recognition is a promising characteristic, in particular for forensic identification. The results from GES-3D stress the strong connection between capture scenario and the performance that can be expected from the recognition system. Even though we achieved high recognition performance in our experiments using academic datasets, we could not reproduce these results with the GES-3D system. Ear recognition systems from CCTV footage needs further research efforts, which should particularly focus on the question of how pose variations affect the appearance of the outer ear. Moreover, the availability of a suitable dataset would be a valuable contribution to this goal. Ear recognition systems could also benefit from existing technology for face recognition, especially in unconstrained scenarios, where the face recognition community is several steps ahead.

Part II

Research Papers

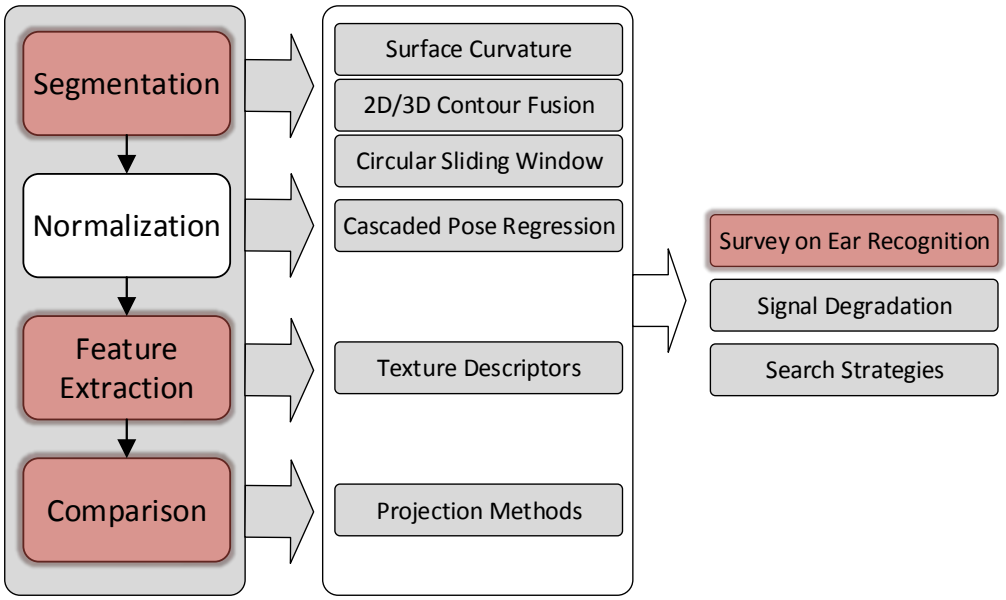
Chapter 3

Ear Biometrics: A Survey of Detection, Feature Extraction and Recognition Methods

This paper provides an elaborate overview of the state of the art in ear recognition in 2012, when this project was launched and it intended to answer research questions **Q0: What is the current state of the art in ear recognition?**

This work gives an overview of available databases and compares a large selection of previous work on segmentation and recognition with respect to the approaches and their performance indicators. It concludes with a section that outlines future challenges in the field. Please refer to Appendix C for an additional Survey of the progress in the field since the publication of this paper.

The paper was published in [147]ANIKA PFLUG, CHRISTOPH BUSCH, Ear Biometrics - A Survey of Detection, Feature Extraction and Recognition Methods, IET Biometrics, Volume 1, Number 2, pp. 114-129.



Abstract

The possibility of identifying people by the shape of their outer ear was first discovered by the French criminologist Bertillon, and refined by the American police officer Iannarelli, who proposed a first ear recognition system based on only seven features.

The detailed structure of the ear is not only unique, but also permanent, as the appearance of the ear does not change over the course of a human life. Additionally, the acquisition of ear images does not necessarily require a person's cooperation but is nevertheless considered to be non-intrusive by most people.

Because of these qualities, the interest in ear recognition systems has grown significantly in recent years. In this survey, we categorize and summarize approaches to ear detection and recognition in 2D and 3D images. Then, we provide an outlook over possible future research in the field of ear recognition, in the context of smart surveillance and forensic image analysis, which we consider to be the most important application of ear recognition characteristic in the near future.

3.1 Introduction

As there is an ever-growing need to automatically authenticate individuals, biometrics has been an active field of research over the course of the last decade. Traditional means of automatic recognition, such as passwords or ID cards, can be stolen, faked, or forgotten. Biometric characteristics, on the other hand, are universal, unique, permanent, and measurable.

The characteristic appearance of the human outer ear (or pinna) is formed by the outer helix, the antihelix, the lobe, the tragus, the antitragus, and the concha (see Figure 3.1). The numerous ridges and valleys on the outer ear's surface serve as acoustic resonators. For low frequencies the pinna reflects the acoustic signal towards the ear canal. For high frequencies it reflects the sound waves and causes neighboring frequencies to be dropped. Furthermore the outer ear enables humans to perceive the origin of a sound.

The shape of the outer ear evolves during the embryonic state from six growth nodules. Its structure, therefore, is not completely random, but still subject to cell segmentation. The influence of random factors on the ear's appearance can best be observed by comparing the left and the right ear of the same person. Even though the left and the right ear show some similarities, they are not symmetric [5].

The shape of the outer ear has long been recognized as a valuable means for personal identification by criminal investigators. The French criminologist Alphonse Bertillon was the first to become aware of the potential use for human identification through ears, more than a century ago [26]. In his studies regarding personal recognition using the outer ear in 1906, Richard Imhofer needed only four different characteristics to distinguish between 500 different ears [83]. Starting in 1949, the American police officer Alfred Iannarelli conducted the first large scale study on the discriminative potential of the outer ear. He collected more than 10 000 ear images and determined 12 characteristics needed to unambiguously identify a person [81]. Iannarelli also conducted studies on twins and triplets, discovering that ears are even unique among genetically identical persons. Even though Iannarelli's work lacks a complex theoretical basis, it is commonly believed that the shape of the outer ear is unique. The studies in [124] and [176] show that all ears of the investigated databases possess individual characteristics, which can be used for distinguishing between them. Because of the lack of a sufficiently large ear database, these studies can only be seen as hints, not evidence, for the outer ear's uniqueness.

Research about the time-related changes in the appearance of the outer ear has shown, that the ear changes slightly in size when a person ages [173][125]. This is explained by the fact that with ageing the microscopic structure of the ear cartilage changes, which reduces the skin elasticity. A first study on the effect of short periods of time on ear recognition [82] shows that the recognition rate is not affected by ageing. It must, however, be mentioned

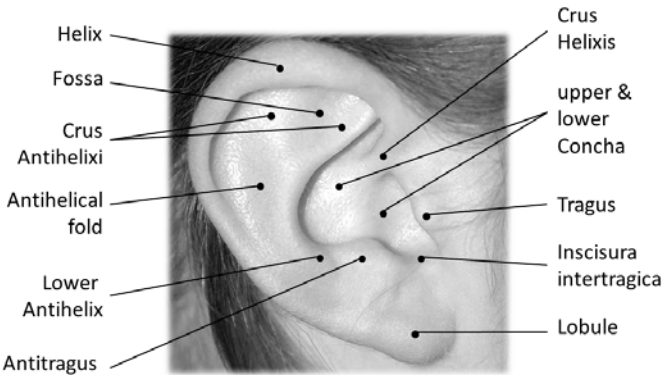


Figure 3.1: Characteristics of the human ear the German criminal police uses for personal identification of suspects

that the largest time elapsing difference in this experiment was only 10 months, and it therefore is still subject to further research whether time has a critical effect on biometric ear recognition systems or not.

The ear can easily be captured from a distance, even if the subject is not fully cooperative. This makes ear recognition particularly interesting for smart surveillance tasks and for forensic image analysis. Nowadays the observation of characteristics is a standard technique in forensic investigation and has been used as evidence in hundreds of cases. The strength of this evidence has, however, also been called into question by courts in the Netherlands [79]. In order to study the strength of ear prints as evidence, the Forensic Ear identification Project (FearID) was initiated by nine institutes from Italy, the UK, and the Netherlands in 2006. In their test system, they measured an EER of 4% and came to the conclusion that ear prints can be used as evidence in a semi-automated system [13]. The German criminal police use the physical properties of the ear in connection with other appearance-based properties to collect evidence for the identity of suspects from surveillance camera images. Figure 3.1 illustrates the most important elements and landmarks of the outer ear, which are used by the German BKA for manual identification of suspects.

In this work we extend existing surveys on ear biometrics, such as [88],[50],[161],[108] or [164]. Abaza *et al.* [6] contributed an excellent survey on ear recognition in March 2010. Their work covers the history of ear biometrics, a selection of available databases and a review of 2D and 3D ear recognition systems. This work amends the survey by Abaza *et al.* with the following:

- A survey of free and publicly available databases.
- More than 30 publications on ear detection and recognition from 2010 to 2012 that were not discussed in one of the previous surveys.
- An outlook over future challenges for ear recognition systems with respect to concrete applications.

In the upcoming Section we give an overview of image databases suitable for studying ear detection and recognition approaches for 2D and 3D images. Thereafter, we discuss existing ear detection approaches on 2D and 3D images. In Section 3.4 we go on to give an overview of ear recognition approaches for 2D images, and in Section 3.5 we do the same for 3D images. We will conclude our work by providing an outlook over future challenges and applications for ear recognition systems.

3.2 Available Databases for Ear Detection and Recognition

In order to test and compare the detection or recognition performance of a computer vision system, in general, and a biometric system in particular, image databases of sufficient size must be publicly available. In this section, we want to give an overview of suitable databases for evaluating the performance of ear detection and recognition systems, which can either be downloaded freely or can be licensed with reasonable effort.

3.2.1 USTB Databases

The University of Science and technology in Beijing offers four collections [1] [2] of 2D ear and face profile images to the research community. All USTB databases are available under license.

- Database I: The dataset contains 180 images in total, which were taken from 60 subjects in 3 sessions between July and August 2002. The database only contains images of the right ear from each subject. During each session, the images were taken under different lighting conditions and with a different rotation. The subjects were students and teachers from USTB.
- Database II: Similarly to database I, this collection contains right ear images from students and teachers from USTB. This time, the number of subjects is 77 and there were 4 different sessions between November 2003 and January 2004. Hence the database contains 308 images in total, which were taken under different lighting conditions.
- Database III: In this dataset 79, students and teachers from USTB were photographed in different poses between November 2004 and December 2004. Some of the ears are occluded by hair. Each subject rotated his or her head from 0 degrees to 60 degrees to the right and from 0 degrees to 45 degrees to the left. This was repeated on two different days for each subject, which resulted in 1600 images in total.
- Database IV: Consisting of 25500 images from 500 subjects taken between June 2007 and December 2008, this is the largest dataset at USTB. The capturing system consists of 17 cameras and, is capable of taking 17 pictures of the subject simultaneously. These cameras are distributed in a circle around the subject, who is placed in the center. The interval between the cameras is 15 degrees. Each volunteer was asked to look upwards, downwards and eyelevel, which means that this database contains images at different yaw and pitch poses. Please note that this database only contains one session for each subject.

3.2.2 UND Databases

The University of Notre Dame (UND) offers a large variety of different image databases, which can be used for biometric performance evaluation. Among them are five databases containing 2D images and depth images, which are suitable for evaluation ear recognition systems. All databases from UND can be made available under license [3].

- Collection E: 464 right profile images from 114 human subjects, captured in 2002. For each user, between 3 and 9 images were taken on different days and under varying pose and lighting conditions.
- Collection F: 942 3D (depth images) and corresponding 2D profile images from 302 human subjects, captured in 2003 and 2004.

[1] http://www1.ustb.edu.cn/resb/en/doc/Imagedb_123_intro_en.pdf
[2] http://www1.ustb.edu.cn/resb/en/doc/Imagedb_4_intro_en.pdf
[3] http://cse.nd.edu/ cvrl/CVRL/Data_Sets.html

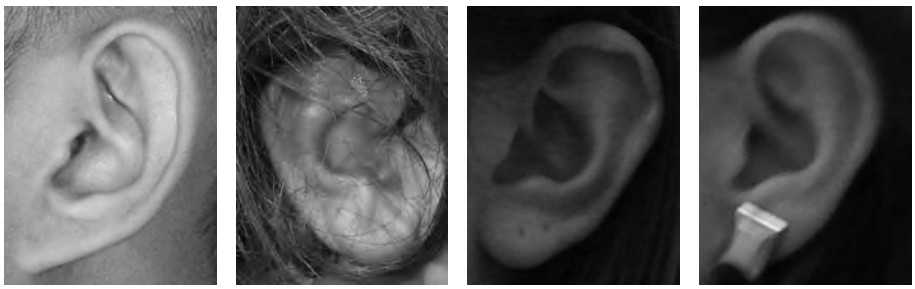


Figure 3.2: Example images from the WPUT ear database [66]. The database contains ear photographs of varying quality and taken under different lighting conditions. Furthermore the database contains images, where the ear is occluded by hair or by earrings.

Figure 3.3: Example images from IIT Delhi ear database [107].

- Collection G: 738 3D (depth images) and corresponding 2D profile images from 235 human subjects, captured between 2003 and 2005
- Collection J2: 1800 3D (depth images) and corresponding 2D profile images from 415 human subjects, captured between 2003 and 2005 [201].
- Collection NDOff-2007: 7398 3D and corresponding 2D images of 396 human subject faces. The database contains different yaw and pitch poses, which are encoded in the file names [64].

3.2.3 WPUT-DB

The West Pommeranian University of Technology has collected an ear database with the goal of providing more representative data than comparable collections [4] [66]. The database contains 501 subjects of all ages and 2071 images in total. For each subject, the database contains between 4 and 8 images, which were taken on different days and under different lighting conditions. The subjects are also wearing headdresses, earrings and hearing aids, and in addition to this, some ears are occluded by hair. In Figure 3.2, some example images from the database are shown. The presence of each of these disruptive factors is encoded in the file names of the images. The database can be freely downloaded from the given URL.

3.2.4 IIT Delhi

The IIT Delhi Database is provided by the Hong Kong Polytechnic University [5] [107]. It contains ear images that were collected between October 2006 and June 2007 at the Indian Institute of Technology Delhi in New Delhi (see Figure 3.3). The database contains 121

[4] http://ksm.wi.zut.edu.pl/wputedb/
[5] http://www4.comp.polyu.edu.hk/čsajaykr/IITD/Database_Ear.htm

Figure 3.4: SCface example images [71]. These images show examples for the photographed pictures, not for the pictures collected with the surveillance camera system.

subjects, and at least 3 images were taken per subject in an indoor environment, which means that the database consists of 421 images in total.

3.2.5 IIT Kanpur

The IITK database was contributed by the Indian Institute of Technology in Kanpur[6] [156]. This database consists of two subsets.

- Subset I: This dataset contains 801 side face images collected from 190 subjects. Number of images acquired from an individual varies from 2 to 10.
- Subset II: The images in this subset were taken from 89 individuals. For each subject 9 images were taken with three different poses. Each pose was captured at three different scales. Most likely, all images were taken on the same day. It is not stated whether subset II contains the same subjects as subset I.

3.2.6 ScFace

The SCface database is provided by the Technical University of Zagreb [7][71] and contains 4160 images from 130 subjects. The aim of the database is to provide a database, which is suitable for testing algorithms under surveillance scenarios. Unfortunately, all surveillance camera images were taken at a frontal angle, such that the ears are not visible on these images. However the database also contains a set of high resolution photographs from each subject, which show the subject at different poses. These poses include views of the right and left profile, as shown in Figure 3.4. Even though the surveillance camera images are likely to be unsuitable for ear recognition studies, the high resolution photographs could be used for examining resistance to pose variations of an algorithm.

3.2.7 Sheffield Face Database

This database was formerly known as the UMIST [8] database and consists of 564 images of 20 subjects of mixed race and gender. Each subject is photographed in a range of different

[6] http://www.cse.iitk.ac.in/users/biometrics/
[7] http://www.scface.org/
[8] http://www.sheffield.ac.uk/eee/research/iel/research/face

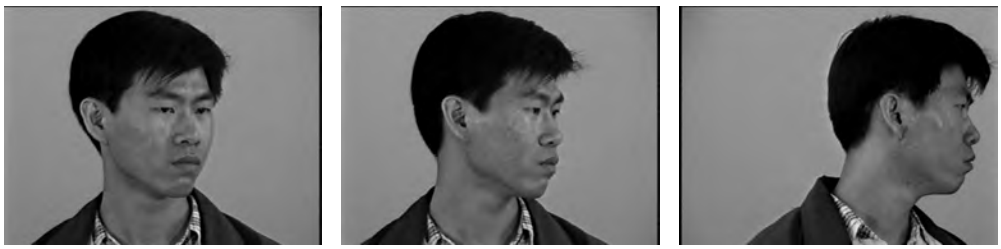


Figure 3.5: Some example images from the NKCU face database, showing the same subject at different angles.

yaw poses, including a frontal view and profile views.

3.2.8 YSU

The Youngston State University collected a new kind of biometric database for evaluation forensic identification systems [11]. For each of the 259 subjects, 10 images are provided. The images are grabbed from a video stream and show the subject in poses between zero and 90 degrees. This means that the database contains right profile images and a frontal view image for each subject. It also contains hand drawn sketches from 50 randomly selected subjects from a frontal angle. However this part of the database is not of interest for ear recognition systems.

3.2.9 NCKU

The National Cheng Kung University in Taiwan has collected an image database, which consists of 37 images for each of the 90 subjects. It can be downloaded from the university's website[9]. Each subject is photographed in different angles between -90 degrees (left profile) and 90 degrees (right profile) in 5 degree steps. In Figure 3.5 some examples are displayed. Such a series of images is collected at two different days for each of the subjects. All images were taken under the same lighting conditions and with the same distance between the subject and the camera.

As this data was originally collected for face recognition, some of the ears are partly or fully occluded by hair, which make this data challenging for ear detection approaches. Consequently, only a subset of this database is suitable for ear recognition.

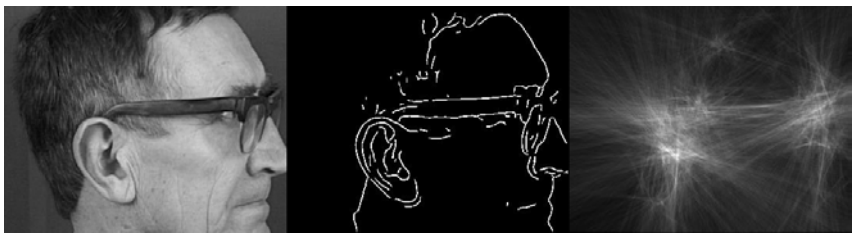
3.2.10 UBEAR dataset

The dataset presented in [165] contains images from the left and the right ear of 126 subjects. The images were taken under varying lighting conditions and the subjects were not asked to remove hair, jewelry or headdresses before taking the pictures. The images are cropped from video stream, which shows the subject in different poses, such as looking towards the camera, upwards or downwards.

Additionally, the ground truth for the ear's position is provided together with the database, which makes it particularly convenient for researches to study the accuracy of ear detection and to study the ear recognition performance independently from any ear detection.

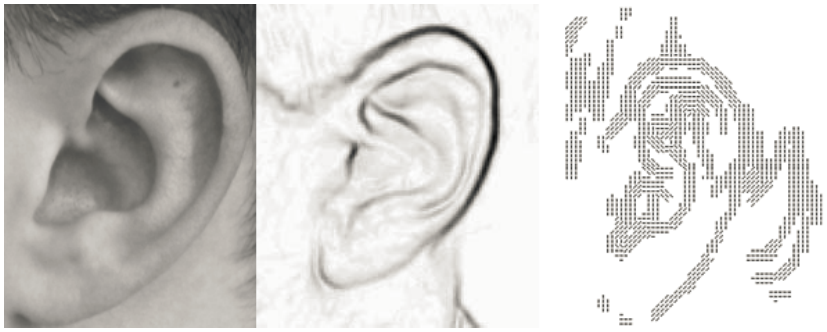
Table 3.1: Summary of automatic ear detection methods for 2D and 3D images

Publication	**Detection Method**	**Database**		**Perf.**
		# Img	**Type**	
Chen & Bhanu [49]	Shape model and ICP	700	3D	87.71%
Chen & Bhanu [47]	Helix Shape Model	213	3D	92.6%
Zhou *et al.* [218]	Histograms of Categorized Shapes	942	3D	100%
Prakash & Gupta [157]	connectivity graph	1604	3D	99.38%
Abaza *et al.* [4]	Cascaded adaboost	940	2D	88.72%
Ansari and Gupta [16]	Edge detection and curvature estimation	700	2D	93.34%
Alvarez *et a.l* [14]	Ovoid model	NA	2D	NA
Arbab-Zavar & Nixon [17]	Hough Transform	942	2D	91%
Arbab-Zavar & Nixon [18]	Log-Gabor filters and wavelet transform	252	2D	88.4%
Attarchi *et al.* [20]	Edge detection and line tracing	308	2D	98.05%
Chen & Bhanu [48]	Template Matching with Shape index histograms	60	2D	91.5%
Cummings *et al.* [55]	Ray transform	252	2D	98.4%
Islam *et al.* [84]	Adaboost	942	2D	99.89%
Jeges & Mate [97]	Edge orientation pattern	330	2D	100%
Kumar *et al.* [106]	Edge clustering and active contours	700	2D	94.29%
Liu & Liu [114]	Adaboost and skin color filtering	50	2D	96%
Prakash & Gupta [158]	Skin color and graph matching	1780	2D	96.63%
Shih *et al.* [174]	Arc-Masking and AdaBoost	376	2D	100%
Yan & Bowyer [201]	Concha Detection and active contours	415	2D	>97.6
Yuan & Mu [209]	CAMSHIFT and a contour fitting	Video	2D	NA



Original image, edge image and Hough transform [17]

Original image and ray transform [55]



Original image, edge enhanced image and corresponding edge orientation model [97].

Figure 3.6: Examples for different ear detection techniques

3.3 Ear Detection

This section summarizes the state of the art in automatic ear detection in 2D and 3D images respectively. Basically all ear detection approaches are relying on mutual properties of the ears morphology, like the occurrence of certain characteristic edges or frequency patterns. Table 3.1 gives a short overview of the ear detection methods outlined below. The upper part of the table contains algorithms for 3D ear localization, whereas the lower part lists algorithms designed for ear detection in 2D images.

Chen and Bhanu propose three different approaches for ear detection. In the approach from [48] Chen and Bhanu train a classifier, which recognizes a specific distribution of shape indices, which are characteristic for the ear's surface. However this approach only works on profile images and is sensitive to any kind of rotation, scale and pose variation. In their later ear detection approaches Chen and Bhanu detected image regions with a large local curvature with a technique they called step edge magnitude [47]. Then a template, which contains the typical shape of the outer helix and the anti-helix, is fitted to clusters of lines. In [49] where Chen and Bhanu narrowed the number of possible ear candidates by detecting the skin region first before the helix template matching is applied on the curvature

[9] http://robotics.csie.ncku.edu.tw/Databases/FaceDetect_PoseEstimate.htm

lines. By fusing color and curvature information, the detection rate could be raised to 99.3% on the UCR dataset and 87.71% on UND collection F and a subset of collection G. The UCR dataset is not publicly available and is hence not covered in Section 3.2. For a description of this dataset see [6].

Another example for ear detection using contour lines of the ear is described by Attrachi *et al.* [20]. They locate the outer contour of the ear by searching for the longest connected edge in the edge image. By selecting the top, bottom, and left points of the detected boundary, they form a triangle with the selected points. Further the barycenter of the triangle is calculated and selected as reference point for image alignment. Ansari *et al.* also use an edge detector in the first step of their ear localization approach [16]. The edges are separated into two categories, namely convex and concave. Convex edges are chosen as candidates for representing the outer contour. Finally the algorithm connects the curve segments and selects the figure enclosing the largest area for being the outer ear contour. It should be noted that the IITK database and USTB II already contain cut-out ear images. Hence it can be put into question, whether the detection rates of 93.34% and 98.05% can be reproduced under realistic conditions.

A recent approach on 2D ear detection using edges is described by Prakash and Gupta in [158]. They combine skin segmentation and categorization of edges into convex and concave edges. Afterwards the edges in the skin region are decomposed into edge segments. These segments are composed to form an edge connectivity graph. Based on this graph the convex hull of all edges, which are believed to belong to the ear, is computed. The enclosed region is then labeled as the ear region. In contrast to [20], Prakash and Gupta prove the feasibility of edge-based ear detection on full profile images, where they achieved a detection rate of 96.63% on a subset of the UND-J2 collection. In [157] propose the same edge connectivity for ear recognition on 3D images. Instead of edges, they use discontinuities in the depth map for extracting the initial edge image and then extract the connectivity graph. In their experiments, they use the 3D representations of the same subset as in [158] and report a detection rate of 99.38%. Moreover they show that the detection rate of their graph-based approach is not influence by rotation and scale.

Jedges and Mate propose another edge-based ear detection approach, which is likely to be inspired by fingerprint recognition techniques. They train a classifier with orientation pattern, which were previously computed from ear images. Like other naive classifiers, their method is not robust against rotation and scale. Additionally the classifier is likely to fail under large pose variations, because this will affect the appearance of the orientation pattern.

Abaza *et al.* [4] and Islam *et al.* [86] use weak classifiers based on Haar-wavelets in connection with AdaBoost for ear localization. According to Islam *et al.*, the training of the classifier takes several days, however once the classifier is set up, ear detection is fast and effective. Abaza *et al.* use a modified version of AdaBoost and report a significantly shorter training phase. The effectiveness of their approach is proved in evaluations on five different databases. They also include some examples of successful detections on images from the internet. As long as the subject's pose does not change, weak classifiers are suitable for images which contain more than one subject. Depending on the test set Abaza *et al.* achieved a detection rate between 84% and 98.7% on the Sheffield Face database. On average, their approach successfully detected 95% of all ears.

Yan and Bowyer developed an ear detection method which fuses range images and corresponding 2D color images [201]. Their algorithm starts by locating the concha and then uses active contours for determining the ear's outer boundary. The concha serves as the reference point for placing the starting shape of the active contour model. Even though the concha is easy to localize in profile images, it may be occluded if the head pose changes or if a subject is wearing a hearing aid or ear phones. In their experiments Yan and Browyer only use ear images with minor occlusions where the concha is visible; hence it could neither be proved nor disproved whether their approach is capable of reliably

detecting ears if the concha is occluded.

Yuan and Mu developed a method for real-time ear tracking in video sequences by applying Continuously Adaptive Mean Shift (CAMSHIFT) to video sequences [209]. The CAMSHIFT algorithm is frequently used in face tracking applications and is based on region matching and a skin color model. For precise ear segmentation, the contour fitting method based on modified active shape models, which have been proposed by Alvarez *et al.* is applied [14]. Yuan and Mu report a detection rate of 100%, however the test database only consisted of two subjects. Nevertheless their approach appears to be very promising for surveillance applications but needs to be further evaluated in more realistic test scenarios.

Shih *et al.* determine ear candidates by localizing arc-shaped edges in an edge image. Subsequently the arc-shaped ear candidates are verified by using an Adaboost classifier. They report a detection rate 100% on a dataset, which consists of 376 images from 94 subjects.

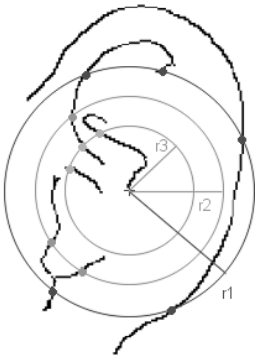
Zhou *et al.* train a 3D shape model in order to recognize the histogram of shape indexes of the typical ear [218]. Similarly to the approaches of Abaza *et al.* and Islam *et al.*, a sliding window of different sizes is moved over the image. The ear descriptor proposed by Zhou *et al.* is built from concatenated shape index histograms, which are extracted from subblocks inside the detection window. For the actual detection, an SVM classifier is trained to decide whether an image region is the ear region or not. As far as we know, this is the first ear detection approach, which does not require having corresponding texture images in addition to the range image. Zhou *et al.* evaluated their approach on images from the UND collections and report a detection rate of 100%. It should be noted that this approach was not tested under rotation, pose variations and major occlusions, but under the impression of the good performance, we think this is an interesting task for future research.

Ear detection methods based on image transformations have the advantage of being robust against out-of-plane rotations. They are designed to highlight specific properties of the outer ear, which occur in each image where the ear is visible no matter in which pose the ear has been photographed. In [17] the Hough transform is used for enhancing regions with a high density of edges. In head profile images, a high density of edges especially occurs in the ear region (see Figure 3.3). In [17] it is reported that the Hough transform based ear detection gets trapped when people wear glasses since the frame introduces additional edges to the image. This especially occurs in the eye and nose region. The ear detection approach based on Hough transform was evaluated on the images in the XM2VTS database (see [6] for a detailed database description), where a detection rate of 91% was achieved.

The ray transform approach proposed in [55] is designed to detect the ear in different poses. Ray transform uses a light ray analogy to scan the image for tubular and curved structures like the outer helix. The simulated ray is reflected in bright tubular regions and hence these regions are highlighted in the transformed image. However, the ray transform also highlights straight edges and edges from other objects, such as hair and glasses (see Figure 3.3). Using this method Alastair *et al.* achieved am impressive recognition rate of 98.4% on the XM2VTS database. Hence, the ray transform approach by Cummings *et al.* outperforms Hough transform, most likely because it is more robust against disruptive factors such as glasses or hair.

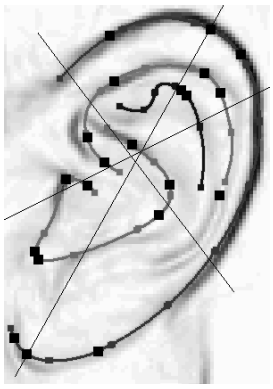
A recent approach for 2D ear detection is described in [106]. Kumar *et al.* propose to extract ears from 2D images by using edge images and active contours. They evaluate their approach on a database, which consists of 100 subjects with 7 images per subject. A special imaging device was used for collecting the data. This device makes sure that the distance to the camera is constant and that the lighting conditions are the same for all images. Within this setting a detection rate of 94.29% is reported.

When putting ear detection into practice, robustness against pose variation and occlusion is of great importance. Nevertheless, most of the ear detection methods described above were not tested with realistic occlusion scenarios, such as occlusion by hair, jew-



Concentric Circles [52]

SIFT features [19]



Active Contour [97]

Force Field [80]

Figure 3.7: Examples for feature extraction for 2D ear images.

ellery or headdresses. A possible reason for this may be the lack of databases containing appropriate images, but this gap has been filled recently by different working groups, who contributed appropriate datasets (see Section 3.2). Furthermore, to our best knowledge there are no investigations on the effect of occlusion in 3D ear images.

3.4 2D Ear Recognition

Each ear recognition system consists of a feature extraction and a feature vector comparison step. In this survey we divide ear recognition approaches into four different subclasses namely holistic approaches, local approaches, hybrid approaches and statistical approaches.

In Tables 3.2 and 3.3 all 2D ear recognition approaches mentioned in this paper are summarized in chronological order.

3.4.1 Holistic Descriptors

Another approach, which has gained some popularity is the Force Field Transform by Hurley [80]. The Force Field transformation approach assumes that pixels have a mutual attraction proportional to their intensities and inversely to the square of the distance between them rather like Newton's universal law of gravitation. The associated energy field takes the form of a smooth surface with a number of peaks joined by ridges (see Figure 3.4.1).

Table 3.2: Summary of approaches for 2D ear recognition, part 1. Unless stated differently, performance always refers to rank-1 performance.

Publication	Summary	Database		Perf.
		# Subj	# Img	
Burge and Burger [35]	Vornoi Distance Graphs	NA	NA	NA
Yuan and Mu [207]	Full Space LDA with Outer Helix Feature Points	79	1501	86.76%
Hurley [80]	Force Field Transform	63	252	99%
Moreno *et al.* [128]	Geometric features with Compression Network	28	268	93%
Yuizono *et al.* [210]	Genetic Local Search	110	660	99%
Victor *et al.* [180]	PCA	294	808	40%
Chang *et al.* [46]	PCA	114	464	72.7%
Abdel-Mottaleb and Zhou [7]	Modified Force Field Transform	29	58	87.9%
Mu *et al.* [129]	Geometrical measures on edge images	77	308	85%
Abate *et al.* [1]	General Fourier Descriptor	70	210	88%
Lu *et al.* [119]	Active Shape Model and PCA	56	560	93.3%
Yuan *et al.* [206]	Non-Negative Matrix Factorization	77	308	91%
Arbab-Zavar *et al.* [19]	SIFT points from ear model	63	252	91.5%
Jedges and Mate [97]	Distorted Ear Model with feature points	28	4060	5.6% EER
Liu *et al.* [115]	Edge-based features from different views	60	600	97.6%
Nanni and Lumini [132]	Gabor Filters and SFFS	114	464	80%
Rahman *et al.* [163]	Geometric Features	100	350	87%
Sana *et al.* [169]	Haar Wavelets and Hamming Distance	600	1800	98.4%
Arbab-Zavar and Nixon [18]	Log-Gabor Filters	63	252	85.7 %
Choras [52]	Geometry of ear outline	188	376	86.2%

Table 3.3: Summary of approaches for 2D ear recognition, part 2. Unless stated differently, performance always refers to rank-1 performance.

Publication	Summary	Database		Perf.
		# Subj	# Img	
Dong and Mu [63]	Force Field Transform and NKFDA	29	711	75.3%
Guo and Xu [72]	Local Binary Pattern and CNN	77	308	93.3%
Nasseem *et al.* [133]	Sparse representation	32	192	96.88%
Wang *et al.* [192]	Haar Wavelets and Local Binary Patterns	79	395	92.41%
Xie and Mu [198]	Locally Linear Embedding	79	1501	80%
Yaqubi *et al.* [204]	HMAX and SVM	60	180	96.5%
Zhang and Mu [215]	Geometrical Features, ICA and PCA with SVM	77	308	92.21
Badrinath and Gupta [21]	SIFT landmarks from ear model	106	1060	95.32%
Kisku *et al.* [102]	SIFT from different Color Segments	400	800	96.93%
Wang and Yuan [189]	Low-Order Moment Invariants	77	308	100%
Alaraj *et al.*. [12]	PCA with MLFFNNs	17	85	96%
Bustard *et al.* [39]	SIFT Point Matches	63	252	96%
De Marisco *et al.* [60]	Partitioned Iterated Function System (PIFS)	114	228	61%
Gutierrez *et al.* [73]	MNN with Sugeno Measures and SCG	77	308	97%
Wang *et al.* [191]	Moment Invariants and BP Neural Network	NA	60	91.8%
Wang and Yuan [190]	Gabor Wavelets and GDA	77	308	99.1%
Prakash and Gupta [156]	SURF and NN classifier	300	2066	2.25% EER
Kumar *et al.* [106]	SIFT	100	700	95% GAR, 0.1% FAR
Wang and Yan [193]	Local Binary Pattern and Wavelet Transform	77	308	100%
Kumar and Wu [107]	Phase encoding with Log Gabor filters	221	753	95.93%

Using this method, Hurley *et al.* achieved a rank-1 performance of more than 99% on the XM2VTS database (252 images). Building on these results, Abdel-Mottaleb and Zhou use a 3D representation of the force field for extracting points lying on the peak of the 3D force field [7]. Because the force field converged at the outline of the ear, the peaks in the 3D representation basically represent the ear contour. Nonetheless, the force field method is more robust against noise than other edge detector, such as Sobel or Canny. Using this approach, Abdel-Mottaleb and Zhou achieved a rank-1 performance of 87.93% on a dataset with consists of 103 ear images from 29 subjects.

Dong and Mu [63] add pose invariance to the edges, which are extracted by using the force field method. This is achieved with null space kernel fishier discriminant analysis (NKFDA), which has the property of representing non-linear relations between two datasets. Dong and Mu conducted experiments on the USTB IV dataset. Before feature extraction, the ear region was cropped out manually from the images and the pose is normalized. For pose variations of 30 degrees they report a rank-1 recondition rate of 72.2%. For pose variations of 45 degrees the rank-1 performance dropped to 48.1%.

In a recent publication of Kumar and Wu [107] they present an ear recognition approach, which uses the phase information of Log-Gabor filters for encoding the local structure of the ear. The encoded phase information is stored in normalized grey level images. In the experiments, the Log-Gabor approach outperformed force field features and a landmark-based feature extraction approach. Moreover, different combinations of Log-Gabor filters were compared with each other. The rank-e performance for the Log-Gabor approaches ranges between 92.06% and 95.93% on a database which contains 753 images from 221 subjects.

The rich structure of the outer ear results in specific texture information, which can be measured using Gabor filters. Wang and Yuan [190] extract local frequency features by using a battery of Gabor filters and then select the most distinctive features by using general discriminant analysis. In their experiments on the USTB II database, they compared the performance impact of different settings for the Gabor filters. Different combinations of orientation and scales in the filter sets are compared with each other and it was found that neither the number of scales nor the number of orientations has a major impact on the rank-1 performance. The total rank-1 performance of Wang and Yuan's approach is 99.1%. In a similar approach Arbab-Zavar and Nixon [18] measured the performance of Gabor filters in the XM2VTS database where they report a rank-1 performance of 91.5%. A closer look at the Gabor filter response showed that the feature vectors are corrupted by occlusion or other disruptive factors. In order to overcome this, a more robust comparison method is proposed, which resulted in an improved recognition rate of 97.4%.

Abate *et al.* [1] use a generic Fourier descriptor for rotation and scale invariant feature representation. The image is transformed into a polar coordinate system and then transformed into frequency space. In order to make sure, that the centroid of the polar coordinate system is always at the same position, the ear images have to be aligned before they can be transformed into the polar coordinate system. The concha serves as a reference point for the alignment step, such that the center point of the polar coordinate system is always located in the concha region. The approach was tested on a proprietary dataset, which contains 282 ear images in total. The images were taken on two different days and in different roll and yaw poses. The rank-1 performance of the Fourier descriptor varies depending on the pose angle. For 0 degrees pose variation the rank-1 performance is 96%, but if different poses are included in the experiments, it drops to 44% for 15 degrees and 19% for 30 degres.

In the work of Fooprateepsiri and Kurutach exploit the concepts of multi-resolution Trace transform and Fourier transform. The input images from the CMU PIE database are serialized by using the trace transform and stored in a feature vector. The advantage of the trace transform is that the resulting feature vector is invariant to rotations and scale. Furthermore Fooprateepsiri and Kurutach show that their descriptor is also robust against

pose variations. In total they report a rank-1 performance of 97%.

Sana *et al.* use selected wavelet coefficients extracted during Haar-Wavelet compression for feature representation [169]. While applying the four level wavelet transform several times on the ear image, for each iteration they store one of the derived coefficients in a feature vector. The reported accuracy of their algorithm is 96% and was achieved on the basis of the IITK database and on the Saugor database (350 subjects).

A feature extraction system called PIFS is proposed by De Marisco *et al.* [60]. PIFS measures the self-similarity in an image by calculating affine translations between similar sub regions of an image. In order to make their system robust to occlusion, De Marisco *et al.* divided the ear image into equally large tiles. If one tile is occluded, the other tiles still contain a sufficiently distinctive set of features. De Marisco *et al.* could show that their approach is superior to other feature extraction methods under the presence of occlusion. The experiments of De Marisco *et al.* have been conducted in order to assess the system performance in different occlusion scenarios. The basis for these tests was the UND collection E and the first 100 subjects of the FERET database. If occlusion occurs on the reference image, a rank-1 performance of 61% (compared to 40% on average with other feature extraction methods) is reported. Without occlusion, the rank-1 performance is 93%.

Moment invariants are a statistical measure for describing specific properties of a shape. Wang *et al.* [191] compose six different feature vectors by using seven moment invariants. They also show that each of the moment invariants is robust against changes in scale and rotation. The feature vectors are used as the input for a back propagation neural network which is trained to classify the moment invariant feature sets. Based on a proprietary database of 60 ear images, they report a rank-1 performance of 91.8%. In [189] Wang and Yuan compare the distinctiveness of different feature extraction methods on the USTB I database. They compare the rank-1 performance of Fourier descriptors, Gabor-Transform, Moment Invariants and statistical features and come to the conclusion that the highest recognition rate can be achieved by using moment invariants and Gabor transform. For both feature extraction methods Wang and Yuan report a rank-1 performance or 100%.

3.4.2 Local Descriptors

Scale invariant Feature Transform (SIFT) is known to be a robust way for landmark extraction even in images with small pose variations and varying brightness conditions [118]. SIFT landmarks contain a measure for local orientation; they can also be used for estimating the rotation and translation between two normalized ear images. Bustard *et al.* showed that SIFT can handle pose variations up to 20 degrees [39]. However it is not a trivial to assign a SIFT landmark with its exact counterpart, especially in the presence of pose variations. In highly structured image regions, the density and redundancy of SIFT landmarks is so high, that exact assignment is not possible. Hence the landmarks have to be filtered before the actual comparison can start. Arbab-Zavar *et al.* [19] as well as Badrinath and Gupta [21] therefore train a reference landmark model, which only contains a small number of non-redundant landmarks. This landmark model is used for filtering the SIFT landmarks, which were initially detected in the probe and reference ear. Having the filtered landmarks it is possible to assign each of the landmarks with its matching counterpart. Figure 3.7 shows an example for SIFT landmarks extracted from ear images, which were used as training data for the reference landmark model in the work of Arbab-Zavar *et al.*. Because Arbab-Zavar *et al.* also used the XM2VTS database for evaluation, their results can be directly compared to the rank-1 performance reported by Bustard and Nixon. Arbab-Zavar *et al.* achieved a rank-1 performance of 91.5%. With the more recent approach by Bustard and Nixon the performance could be improved to 96%. Using the IIT Delhi database Kumar *et al.* report a GAR of 95% and a FAR of 0.1% when using SIFT feature points.

Kisku *et al.* address the problem of correct landmark assignment by decomposing the ear image into different color segments [102]. SIFT landmarks are extracted from each seg-

ment separately, which reduces the chance of assigning SIFT landmarks that are not representing the same features. Using this approach, Kisku *et al.* achieve a rank-1 performance of 96.93%.

A recent approach by Prakash and Gupta [156] fuses Speeded Up Robust Features (SURF) [23] feature points from different images of the same subject. They propose to use several input images for enrolment and to store all SURF feature points in the fused feature vector, which could be found in the input images. These feature sets are then used for training a nearest neighbor classifier for assigning two correlated feature points. If the distance between two SURF feature points is less than a trained threshold, they are considered to be correlated. The evaluation of this approach was carried out on the UND collection E and the two subsets of the IIT Kanpur database. Prakash and Gupta tested the influence of different parameters for SURF features and for the nearest neighbor classifier. Depending on the composition of the parameters the EER varies between 6.72% and 2.25%.

Choras proposes a set of geometric feature extraction methods inspired by the work of Iannarelli [52]. He proposes four different ways of feature location in edge images. The concentric circles method uses the concha as reference points for a number of concentric circles with predefined radii. The intersection points of the circles and the ear contours are used as feature points (see Figure 3.7.). An extension of this is the contour tracing method, which uses bifurcations, endpoints and intersecting points between the ear contours as additional features. In the angle representation approach, Choras draws concentric circles around each center point of an edge and uses the angles between the center point and the concentric circles intersecting points for feature representation. Finally the triangle ratio method determines the normalized distances between reference points and uses them for ear description. Choras conducted studies on different databases where he reported recognition rates between 86.2% and 100% on a small database off 12 subjects and a false reject rate between 0% and 9.6% on a larger database with 102 ear images.

Similar approaches which are using the aspect ratio between reference points on the ear contours are proposed by Mu *et al.* with a rank-1 performance of 85% on the USTB II database [129] and Rahman *et al.* [163]. Rahman *et al.* evaluated their approach on a database, which consists of 350 images from 100 subjects. They report a rank-1 performance of 90%. For images, which were taken on different days the rank-1 performance dropped to 88%.

Local binary patterns (LBP) are a technique for feature extraction on the pixel level. LBP encode the local neighborhood of a pixel by storing the difference between the examined pixel and its neighbors. Guo *et al.* extract LBP from the raw ear images and create histograms describing the distribution of the local LBP. Then a cellular neural network is trained to distinguish between the LBP of different subjects in the USTB II database [72].

In the by Wang and Yan [193] the dimensionality of the feature vector is reduced with linear discriminant analysis before a Euclidean distance measure quantifies the similarity of two feature vectors. Wang and Yan evaluated their approach on the USTB II dataset and report a rank-1 performance of 100%.

3.4.3 Hybrid Approaches

The approach of Jedges and Mate is twofold [97]. In a first feature extraction step they generate an average edge model from a set of training images. These edges represent the outer helix contour as well as the contours of the antihelix, the fossa triangularis and the concha. Subsequently each image is enrolled by deforming the ear model until it fits the actual edges displayed in the probe ear image. The deformation parameters, which were necessary for the transformation, are the first part of the feature vector. The feature vector is completed by adding additional feature points lying on intersections between a predefined set of axes and the transformed main edges. The axes describe the unique outline of ear. Figure 3.7 shows the edge enhanced images with fitted contours together with the

additional axes for reference point extraction. They report an EER of 5.6% using a database with cropped images and without pose variations.

Liu *et al.* combine front and backside view of the ear by extracting features using the triangle ratio method and Tchebichef moment descriptors [115]. Tchebichef moments are a set of orthogonal moment functions based on discrete Tchebichef polynomials and have been introduced as a method for feature representation in 2001 [131]. The backside of the ear is described by a number of lines that are perpendicular to the longest axis in the ear contour. These lines measure the local diameter of the auricle at predefined points. The rank-1 performance of this combined approach is reported to be 97.5%. If only the front view is used, the rank-1 performance is 95% and for the backside images, Liu *et al.* report 86.3% rank-1 performance.

Lu *et al.* [119] as well as Yuan and Mu [207] use the active shape model for extracting the outline of the ear. Lu *et al.* are using manually cropped ear images from 56 subjects in different poses. A feature extractor stores selected points on the outline of the ear together with their distance to the tragus. Before applying a linear classifier, the dimensionality of the feature vectors is reduced by principal component analysis (PCA). Lu *et al.* compare the rank-1 performance of pipelines where only the left or the right ear was used for identification and also show that using both ears increases the rank-1 performance from 93.3% to 95.1%. In the USTB III database Yuan and Mu report a rank-1 performance of 90% if the head rotation is lower than 15 degrees. For rotation angles between 20 degrees and 60 degrees the rank-1 performance drops to 80%.

3.4.4 Classifiers and Statistical Approaches

Victor *et al.* were the first research group to transfer the idea of using the Eigen space from face recognition to ear recognition [180]. They reported that the performance of the ear as a feature is inferior to the face. This may be due to the fact that in their experiments Victor *et al.* considered the left and the right ear to be symmetric. They used the one ear for training and the other ear for testing, which could have lowered the performance of PCA in this case. The reported rank-1 performance is 40%. With a rank-1 performance of 72.2% in the UND collection E, Chang *et al.* [46] report a significantly better performance than Victor *et al.*. Alaraj *et al.* [12] published another study, where PCA is used for feature representation in ear recognition. In their approach a multilayer feed forward neural network was trained for classification of the PCA based feature components. The observed a rank-1 performance of 96%, and hence improved the previous results by Victor *et al.* and Chang *et al.*. However it should be noticed that this result is only based on a subset of one of the UND collections, which consists of 85 ear images from 17 subjects.

Zhang and Mu conducted studies on the effectiveness of statistical methods in combination with classifiers. In [215] they show that independent component analysis (ICA) is more effective on the USTB I database than PCA. They first used PCA and ICA for reducing the dimensionality of the input images and then trained an SVM for classifying the extracted feature vectors. Furthermore the influence of different training set sizes on the performance was measured. Depending on the size of the training set the rank-1 performance for PCA varies between 85% and 94.12%, whereas the rank-1 performance for ICA varies between 91.67% and 100%.

Xie and Mu [198] propose an improved locally linear embedding (LLE) algorithm for reducing the dimensionality of ear features. LLE is a technique for projecting high-dimensional data points into a lower dimensional coordinate system while preserving the relationship between the single data points. This requires the data points to be labeled in some way, so that their relationship is fixed. The improved version of LLE by Xie and Mu eliminated the problem by using a different distance function. Further Xie and Mu show, that LLE is superior to PCA and Kernel PCA, if the input data contains pose variations. Their studies were conducted on the USTB III database showed that the rank-1 performance of regular

LLE (43%) is improved significantly by their method to 60.75%. If the pose variation is only 10 degrees, the improved LLE approach achieved a rank-1 performance of 90%.

In their approach Nanni and Lumini [132] propose to use Sequential Forward Floating Selection (SFFS), which is a statistical iterative method for feature selection in pattern recognition tasks. SFFS tries to find the best set of classifiers by creating a set of rules, which best fits the current feature set. The sets are created by adding one classifier at a time and evaluating its discriminative power with a predefined fitness function. If the new set of rules outperforms the previous version, the new rule is added to the final set of rules. The experiments were carried out on the UND collection E and the single classifiers are fused by using the weighted sum rule. SFFS selects the most discriminative sub-windows which correspond to the fittest set of rules. Nanni and Lumini report a rank-1 recognition rate of 80% and a rank-5 recognition rate of 93%. The EER varies between 6.07% and 4.05% depending on the number of sub-windows used for recognition.

Yiuzono *et al.* consider the problem of finding corresponding features in ear images as an optimization problem and apply genetic local search for solving it iteratively [210]. They select local sub windows with varying size as the basis for the genetic selection. In [210] Yiuzono *et al.* present elaborated results, which describe the behavior of genetic local search under different parameters, such as different selection methods and different numbers of chromosomes. On a database of 110 subjects they report a recognition rate of 100%.

Yaqubi *et al.* use features obtained by a combination of position and scale-tolerant edge detectors over multiple positions and orientations of the image [204]. This feature extraction method is called HMAX model and is inspired by the visual cortex of primates and combines simple features to more complex semantic entities. The extracted features are classified with an SVN and a kNN. The rank-1 performance on a small dataset of 180 cropped ear images from 6 subjects varies between 62% and 100% depending on the kind of basis features.

Moreno *et al.* implement a feature extractor, which locates seven landmarks on the ear image, which correspond to the salient points from the work of Iannarelli. Additionally they obtain a morphology vector, which describes the ear as a whole. These two features are used as the input for different neural network classifiers. They compare the performance of each of the single feature extraction techniques with different fusion methods. The proprietary test database is composed of manually cropped ears from 168 from 28 subjects. The best result of 93% rank-1 performance was measures using a compression network. Other configurations yielded error rates between 16% and 57%.

Gutierrez *et al.* [73] divide the cropped ear images into three equally sized parts. The upper part shows the helix, the middle part shows the concha and the lower part shows the lobule. Each of these sub images is decomposed by wavelet transform and then fed into a modular neural network. In each module of the network a different integrators and learning functions was used. The results of each of the modules are fused in the last step for obtaining the final decision. Depending on the combination between integrator and learning function, the results vary between 88.4% and 97.47% rank-1 performance on the USTB I database. The highest rank-1 performance is achieved with Sugeno measure and conjugate gradient.

In [133] Nasseem *et al.* propose a general classification algorithm based on the theory of compressive sensing. They assume that most signals are compressible in nature and that any compression function results in a sparse representation of this signal. In their experiments in the UND database and the FEUD database, Nasseem *et al.* show that their sparse representation method is robust against pose variations and varying lighting conditions. The rank-1 performance varied between 89.13% and 97.83%, depending on the dataset used in the experiment.

Table 3.4: Summary of approaches for 3D ear recognition. Performance (Perf.) always refers to rank-1 performance.

Publication	Comparison Method	Database # Subj	# Img	Perf.
Cadavid *et al.* [40]	ICP and Shape from shading	462	NA	95%
Chen and Bannu [49]	Local Surface Patch	302	604	96.36%
Chen and Bhanu [47]	ICP Contour Matching	52	213	93.3%
Liu and Zhang [116]	Slice Curve Matching	50	200	94.5%
Islam *et al.* [85]	ICP with reduced meshes	415	830	93.98%
Islam *et al.* [87]	Local Surface Features with ICP-Matching	415	830	93.5%
Passalis *et al.* [144]	Reference ear model with morphing	525	1031	94.4%
Yan and Bowyer [200]	ICP using voxels	369	738	97.3%
Yan and Bowyer [201]	ICP using Model Points	415	1386	97.8%
Zheng *et al.* [211]	Local Binary Patters	415	830	96.39%
Zhou *et. al.* [219]	Surface Patch Histogram and voxelization	415	830	98.6%, 1.6% EER

3.5 3D Ear Recognition

In 2D ear recognition pose variation and variation in camera position, so-called out-of-plane-rotations, are still unsolved challenges. A possible solution is using 3D models instead of photos as references, because a 3D representation of the subject can be adapted to any rotation, scale and translation. In addition to that, the depth information contained in 3D models can be used for enhancing the accuracy of an ear recognition system. However, most 3D ear recognition systems tend to be computationally expensive. In Table 3.4 all 3D ear recognition systems described in this section are summarized.

Although ICP is originally designed to be an approach for image registration, the registration error can also be used as a measure for the dissimilarity of two 3D images. Because ICP is designed to be a registration algorithm, it is robust against all kinds of translation or rotations. However ICP tends to stop too early, because it gets stuck in local minima. Therefore ICP requires the two models to be coarsely pre-aligned before fine alignment using ICP can be performed. Chen and Bhanu extract point clouds from the contour of the outer helix and the register these points with the reference model by using ICP [47]. In a later approach Chen and Bhanu use local surface patches (LSP) instead of points lying on the outer helix [49]. As the LSP consist or fewer points than the outer helix, this reduces the processing time while enhancing the rank-1 performance from 93.3% with the outer helix points to 96.63 % with LSP.

Yan and Browyer decompose the ear model into voxels and extract surface features from each of these voxels. For speeding up the alignment process, each voxel is assigned an index in such a way that ICP only needs to align voxel pairs with the same index [200] (see

Figure 3.8). In [201] Yan and Browyer propose the usage of point clouds for 3D ear recognition. In contrast to [47] all points of the segmented ear model are used. The reported performance measures of 97.3% in [200] and 97.8% in [201] is similar but not directly comparable, because different datasets were used for evaluation.

Cadavid *et al.* propose a real-time ear recognition system, which reconstructs 3D models from 2D CCTV images using the shape from shading technique [41]. Thereafter the 3D model is compared to the reference 3D images, which are stored in the gallery. Model alignment as well as the computation of the dissimilarity measure is done by ICP. Cadavid *et al.* report a recognition rate of 95% on a database of 402 subjects. It is stated in [41] that the approach has difficulties with pose variations. In [219] Zhou *et. al.* use a combination of local histogram features voxel-models. Zhou *et. al.* report that their approach is faster and with an EER of 1.6% it is also more accurate than the ICP-based comparison algorithms proposed by Chen and Bhanu and Yan and Browyer.

Simlarly to Cadavid *et al.*, Liu *et al.* reconstruct 3D ear models from 2D views [115]. Based on the two images of a stereo vision camera, a 3D representation of the ear is derived. Subsequently the resulting 3D meshes serve as the input for PCA. However Liu *et al.* do not provide any results concerning the accuracy of their system but since they did not publish any further results on their PCA mesh approach, it seems that it is no longer pursued.

Passalis *et al.* go a different way for comparing 3D ear models in order to make comparison suitable for a real-time system [144]. They compute a reference ear model which is representative for the average human ear. During enrolment, all reference models are deformed until they fit the reference ear model. All translations and deformations, which were necessary to fit the ear to the reference model are then stored as features. If a probe for authentication is given to the system, the model is also adapted to the annotated ear model in order to get the deformation data. Subsequently the deformation data is used to search for an associated reference model in the gallery. In contrast to the previously described systems, only one deformation has to be computed per authentication attempt. All other deformation models can be computed before the actual identification process is started. This approach is reported to be suitable for real-time recognition systems, because it takes less than 1 milliseconds for comparing two ear templates. The increased computing speed is achieved by lowering the complexity class from $O(n)^2$ for ICP-based approaches to $O(n)$ for their approach. The rank-1 recognition rate is reported to be 94.4%. The evaluation is based on non-public data, which was collected using different sensors.

Heng and Zhang propose a feature extraction algorithm based on slice curve comparison, which is inspired by the principles of computer tomography [116]. In their approach the 3D ear model is decomposed into slices along the orthogonal axis of the longest distance between the lobule and the uppermost part of the helix. The curvature information extracted from each slice is stored in a feature vector together with an index value indicating the slice's former position in the 3D model. For comparison the longest common sequence between two slice curves with similar indexes is determined. Their approach is only evaluated on a non-public dataset, which consists of 200 images from 50 subjects. No information about pose variations or occlusion during the capturing experiment is given. Heng and Zhang report a rank-1 performance of 94.5% for the identification experiment and 4.6%EER for the verification experiment.

Islam *et al.* reconnect point clouds describing 3D ear models to meshes and iteratively reduce the number of faces in the mesh [85]. These simplified meshes are then aligned with each other using ICP and the alignment error is used as the similarity measure for the two simplified meshes. In a later approach Islam *et al.* extract local surface patches as shown in Figure 3.9 and use them as features [86]. For extracting those LSP, a number of points is selected randomly from the 3D model. Then the data points which are closer to the seed point than a defined radius are selected. PCA is then applied to find the most descriptive features in the LSP. The feature extractor repeats selecting LSP until the desired number of features has been found. Both approaches were evaluated using images from UND. The

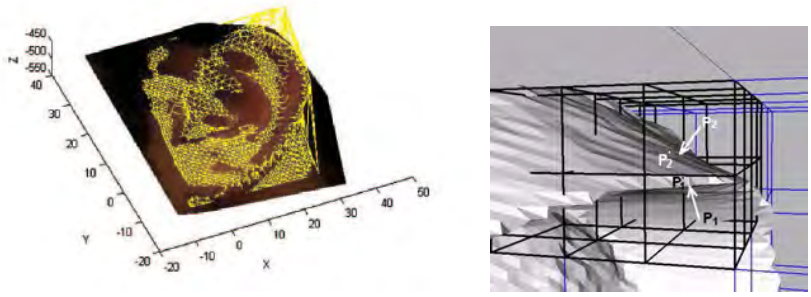


Figure 3.8: Examples for surface features in 3D ear images. The left image shows an example for ICP-based comparison as proposed in [47], whereas the right figure illustrates feature extraction from voxels as described in [200].

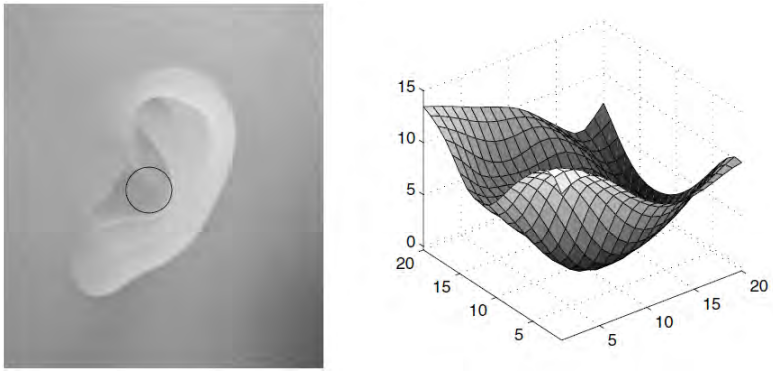


Figure 3.9: Example for local surface patch (LSP) features as proposed in [86]

recognition rate reported for [85] is 93.98% and the recognition rate reported for [86] is 93.5%. However, none of the approaches has been tested with pose variation and different scaling.

Zheng *et al.* extract the shape index at each point in the 3D model and use it for projecting the 3D model to 2D space [211]. The 3D shape index at each pixel is represented by a grey value at the corresponding position in the 2D image. Then SIFT features are extracted from the shape index map. For each of the SIFT points a local coordinate system is calculated where the z-axis correspondents to the feature point's normal. Hence the z-values of the input image are normalized according to the normal of the SIFT feature point they were assigned to. As soon as the z values have been normalized, they are transformed into a grey level image. As a result, Zheng *et al.* get a local grey level image for each of the selected SIFT features. Next LBP are extracted for feature representation in each of these local grey level images. Comparison is first performed by coarsely comparing the shape indexes of key pints and then using Earth mover's distance for comparing LBP histograms from the corresponding normalized grey images. Zheng *et al.* evaluated their approach on a subset of the UND-J2 Collected and achieved a rank-1 performance of 96.39%.

3.6 Open challenges and future applications

As the most recent publications on 2D and 3D ear recognition show, the main application of this technique is personal identification in unconstrained environments. This includes applications for smart surveillance, such as in [40] but also the forensic identification of perpetrators on CCTV images or for border control systems. Traditionally these application

fields are part of face recognition systems but as the ear is located next to the face, it can provide valuable additional information to supplement the facial images.

Multi modal ear and face recognition systems can serve as a means of achieving pose invariance and more robustness against occlusion in unconstrained environments. In most public venues surveillance cameras are located overhead in order to capture as many persons as possible and to protect them from vandalism. In addition, most of the persons will not look straight into the camera, so in most cases no frontal images of the persons will be available. This fact poses serious problems to biometric systems, using facial features for identification. If the face is not visible from a frontal angle, the ear can serve as a valuable additional characteristic in these scenarios.

Because of the physical proximity of the face and the ear, there are also many possibilities for the biometric fusion of these two modalities. Face and ear images can be fused on the feature level, on the template level and on the score level. Against the background of this application, there are some unsolved challenges, which should be addressed by future research in this field.

3.6.1 Automatic Ear Localization

The fact that many systems presented in literature use pre-segmented ear images shows, that the automatic detection of ears especially in real-life images is still an unsolved problem. If ear recognition systems should be implemented in automatic identification systems, fast and reliable approaches for automatic ear detection are of importance. As a first step towards this goal, some research groups have published data collections, which simulate typical variations in uncontrolled environments such as varying lighting conditions, poses and occlusion. Based on these datasets, existing and future approaches to ear recognition should be tested under realistic conditions in order to improve their reliability.

Moreover, 3D imaging systems become increasingly cheap in the last years. Consequently 3D ear recognition becomes important and with it the need of locating ears in depth images or 3D models. Currently, only one approach for ear detection in depth has been published, which is a first step towards ear detection in 3D images.

3.6.2 Occlusion and Pose Variations

In contrast to the face, the ear can be partially or fully covered by hair or by other items such as headdresses, hearing aids, jewelry or headphones. Because of the convex surface of the outer ear, parts of it may also be occluded if the subject's pose changes. In some publications, robustness against occlusion is explicitly addressed, but there are no studies on the effect of the effect of certain types of occlusion like hair or earrings on the recognition rate of an ear recognition system. Once more, the availability of public databases which contain occluded ear images is likely to foster the development of solutions for pose invariant and robust algorithms for ear detection and feature extraction.

Moreover to our best knowledge there are no studies about the visibility of the outer ear in different public environments. In order to develop algorithms for ear detection and recognition, further information about commonly occluded parts of the ear is needed.

Occlusion due to pose variations is another unmet challenge in ear recognition system. Similarly to face recognition, parts of the ear can become occluded if the pose changes. Recently, some feature extraction methods have been proposed, which are robust against pose variations to some degree. However, this issue is not fully solved yet. Another possibility compensating for pose variations could be the usage of 3D models instead of depth images of photographs.

3.6.3 Scalability

Currently available databases only consist of less than 10 000 ear images. The only exception is the USTB IV collection, which has not been released for the public yet. In realistic environments the size of the database will be significantly larger, which makes exhaustive search in identification scenarios infeasible. Therefore, not only the accuracy but also the comparison speed of ear recognition systems will be interesting for future research.

In order to make ear recognition applicable for large scale systems, exhaustive searches should be replaced by appropriate data structures allowing logarithmic time complexity during the search. This could for example be achieved by exploring the possibilities of organizing ear templates in search trees.

3.6.4 Understanding Symmetry and Ageing

Because ear recognition is one of the newer fields of biometric research, the symmetry of the left and the right ear has not been fully understood yet. A study by Abaza and Ross [5] indicates that there is some degree of symmetry between left and right ears, that could be exploited when comparing left and right ears. Their result encourage more research on the symmetry constraints between the left and the right ear.

The studies of Iannarelli indicate that some characteristics of the outer ear can be inherited and ageing slightly affects the appearance of the outer ear. Both assumptions could be confirmed in more recent studies, but because of a lack of sufficient data, the effect of inheritance and ageing on the outer ear's appearance is not fully understood yet. Furthermore, there are no large scale studies of the symmetry relation between the left and the right ear yet.

Therefore another interesting field for future research could be, to gain a deeper understanding of the effect of inheritance any symmetry on the distinctiveness of biometric template. Moreover, long term studies on the effect of time on ear templates are needed in order to get a better understanding of the permanence of this characteristic.

3.7 Summary

We have presented a survey on the state of the art in 2D and 3D ear biometrics, covering ear detection and ear recognition systems. We categorized the large number of 2D ear recognition approaches into holistic, local, hybrid and statistical methods, discussed their characteristics and reported their performance.

Ear recognition is still a new field of research. Although there is a number of promising approaches, none of them has been evaluated under realistic scenarios which include disruptive factors like pose variations, occlusion and varying lighting conditions. In recent approaches, these factors are taken under account, but more research on this is required until ear recognition systems can be used in practice. The availability of suitable test databases, which were collected under realistic scenarios, will further contribute to the maturation of the ear as a biometric characteristic.

We have collected a structured survey of available databases, existing ear detection and recognition approaches and unsolved problems for ear recognition in the context of smart surveillance system, which we consider to be the most important application for ear biometrics. We think that this new characteristic is a valuable extension for face recognition systems on the way to pose invariant automatic identification.

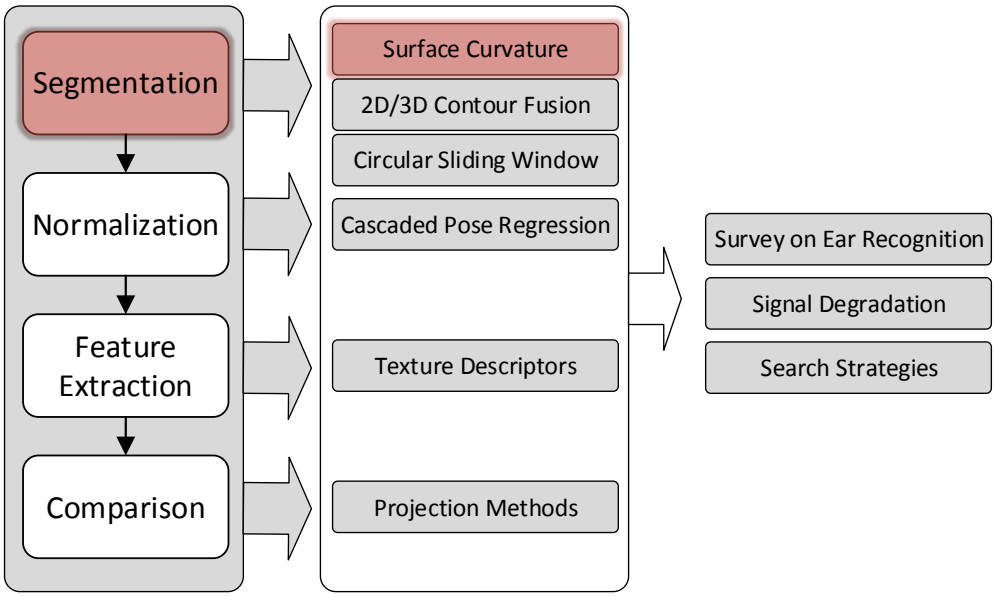
Chapter 4

Ear Detection in 3D Profile Images based on Surface Curvature

In this chapter, we give the first part of the answer to research question **Q1: How can the outer ear be automatically detected from 2D and 3D images?** In this work, we focus on the detection of ears from left profile depth images by using clusters of maximum curvature points.

When talking about 3D ear images, we frequently refer to depth images instead of full 3D representations. In this work, we propose a method for segmenting the outer ear from depth images. Our method makes use the fact that the ear region contains unique shape information that consists of diverse convex and concave structures. We reconstruct the ear outline, by combining these structures to a shape. Our method selects the most likely ear outline, which is the combined shape that fulfils a number of criteria that are typical for the shape of the human ear.

The paper was published in [154]ANIKA PFLUG, ADRIAN WINTERSTEIN, CHRISTOPH BUSCH, Ear Detection in 3D Profile Images Based on Surface Curvature, International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP), 2012



Abstract

Although a number of different ear recognition techniques has been proposed, not much work has been done in the field of ear detection. In this work we present a new ear detection approach for 3D profile images based on surface curvature and semantic analysis of edge-patterns. The algorithm applies edge-based detection techniques, which are known from 2D approaches, to a 3D data model. As an additional result of the ear detection, the outline of the outer helix is found, which may serve as a basis for further feature extraction steps. As our method does not use a reference ear model, the detector does not need any previous training. Furthermore, the approach is robust against rotation and scale. Experiments using the 3D images from UND-J2 collection resulted in a detection rate of 95.65%.

4.1 Introduction

Referring back to the first large-scale study on the suitability of the ear as a biometric characteristic by Iannarelli in 1964 [81], several automated and semi-automated ear recognition systems have been proposed in literature. Since then the ear has been highly valued in forensic image analysis as an individual structure with a high distinctive potential.

The outer ear (also referred to as auricle or pinna) is a richly structured body part, which is composed of cartilage covered by a thin skin layer. Its appearance is not completely random, but rather subject to the somewhat predictable process of cell segmentation. In his work on 'earology', Iannarelli was able to show that the ear is not only unique, but also stable over its entire lifetime. Ear recognition is also more acceptable than face recognition, as people feel more comfortable when a photograph of their ear, as opposed to their face, is taken [51]. In public opinion, ear recognition, unlike fingerprints, is not associated with criminal investigations, and the fact that it does not require any physical contact with the sensor, further contributes to its acceptability.

In the field of 3D ear recognition, a large number of descriptors for ear models have been proposed. Despite this, the problem of ear detection is not addressed by many authors as they are using manually pre-segmented images. Possible solutions to the problem of ear detection have been proposed by Chen and Bhanu [29] and by Yan and Bowyer [201]. Both approaches require the availability of 2D texture images as well as corresponding 3D images. The idea is to reduce complexity by coarsely pre-segmenting the image in 2D space before locating the ear in the 3D model with a helix-template. However, Zhou *et al* showed that the ear can also be localized efficiently without the help of additional 2D texture information [218].

In this work, we introduce a technique for ear detection from a 3D profile image that detects ears from profile images without using color information or making any assumptions about orientation and scale. Our approach is inspired by edge-based 2D ear detection approaches such as [20] or [16] and relies on the fact that edges we see in 2D images are a result of the interaction between extreme curvature values on the object's surface and reflections of ambient light. The algorithm is similar to the first bottom-up ear detection approaches based on depth images by Chen and Bhanu [29]. Our approach, however, exceeds their reported detection rate of 92.4%. In contrast to Chen and Bhanu's method, our work is based on curvature values and uses a number of constraints that provide a general description of the ear.

The next section presents a detailed description of our 3D ear detection approach and will be split into four subsections. Each of these subsections covers one step of the detection algorithm. Subsequently, section 4.3 presents the results we obtained during the examination of the algorithm's performance and will also cover the strengths and weaknesses of the approach. Finally, in section 4.4, the findings of this paper are summarized and suggestions for further improvements are made.

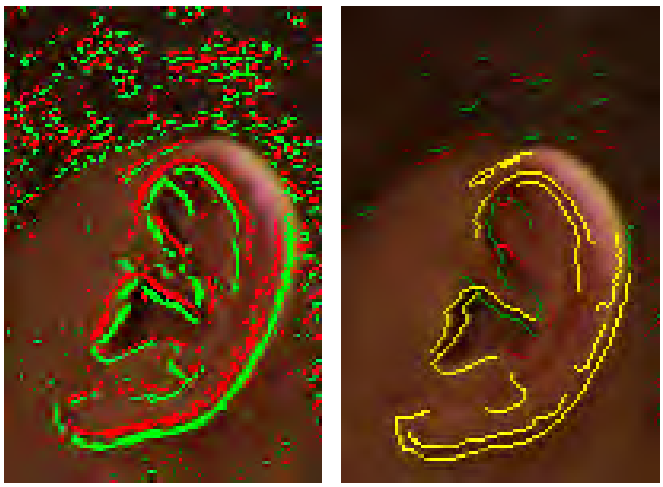


Figure 4.1: Projected curvature lines after threshold (a) and helix candidates after smoothing, thinning and removing small components (b).

4.2 Ear Detection Approach

The ear detection algorithm outlined in this paper can be divided into three subsequent steps, namely the curvature calculation and a preprocessing step, the closing of small gaps in the helix contour, and finally the evaluation of helix candidates. In each step, the algorithm reconstructs and combines lines in the image in such a way as to satisfy a number of conditions that determine if the line is part of the ear. In the upcoming subsections, each of these steps will be explained in more detail.

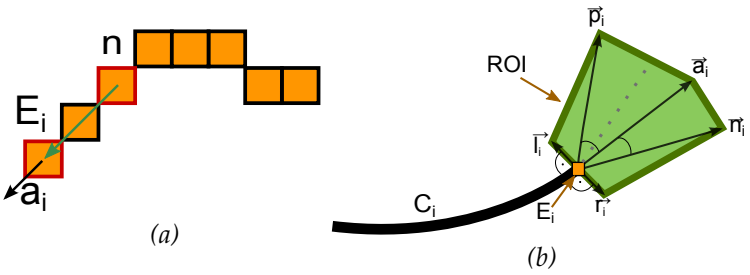
4.2.1 Mean curvature and binarization

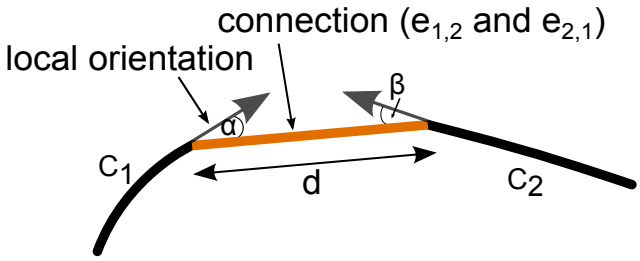
Our detection approach is based on the assumption, that the ear region, with its rich and curved structure, can easily be distinguished from other regions by looking at the local curvature. Hence we need an appropriate measure for quantifying the surface curvature, and we need to assign a curvature value for each point in the 3D model. In our approach, we use the mean curvature H, which is defined as the mean value of the minimum and the maximum principal curvature k_{min} and k_{max} at a point on the surface [28].

$$H = \frac{1}{2}(K_{min} + k_{max}) \tag{4.1}$$

After calculating a curvature value for each point in the 3D model, all mean curvature values between a minimum value t_1 and a maximum value t_2 are removed from the set of points in the model. These threshold values should be defined according to the actual curvature values that occur in the image, such that enough points are left in the model for later analysis. Generally speaking, it is better to have too many, rather than too few points, left in the model. For UND-J2 we chose $t_1 = -0.5$ and $t_2 = 0.5$. This step removes all points with smooth curvature values, and only leaving points with large curvature $H < t_1$ and $H > t_2$ for further processing and in the point set P. As the ear is a structure with extreme surface curvature, the points representing the ear's outline will be a subset of P.

After applying the threshold, all points in P are projected on a 2D binary image according to their x and y coordinates. The depth information, which was contained in the z value, is not used in this step. Instead the points are divided into two categories, as shown in 4.2. Maximum curvature ($H < t_1$) points are depicted in red and minimum curvature ($H > t_2$) are depicted in green. For the rest of this paper, the term curvature always refers to curvature of lines in 2D space, not surface curvature in 3D.





(c)

Figure 4.2: Subfigure *(a)* illustrates the calculation of the local orientation vector. Subfigure *(b)* shows the definition of the search space used when searching for components to connect with, and subfigure *(c)* shows an example of a connection between two components with distance d and the two angles α and β, respectively.

The line-like shapes in the projected 2D image are then smoothed by applying a Gaussian filter and thinned to the width of one pixel. As shown in figure 4.2, the ear contours are prominent larger structures surrounded by smaller ones, which are likely to be noise artifacts. Hence, the last step of the binarization process removes all components that are smaller than a fixed minimum value from the image. For the images in UND-J2, the minimum size for a component was set to 5.

4.2.2 Reconstructing the helix contour

After having made a coarse selection of the components, likely to be part of the ear contour in the previous processing step, the next step reconstructs contour lines from the remaining components in the image. The goal of this step is to reconstruct components from the image that likely belong to the same contour, but are not connected in the 2D binary image. These missing connections are often a result of occlusion by hair or cluttering by other objects, such as scarves or earrings, as large values for surface curvature cannot be found in occluded regions.

The basis for connecting lines in the image is the calculation of the vector $\vec{a_i}$, which represents the local orientation at an endpoint E_i in the image. The local orientation vector $\vec{z_i}$ is determined by the difference between the x and y coordinates of the endpoint E_i and the x and y coordinates of the nth pixel along the currently examined line (See Figure 4.2.2). As the minimum size of a line is 6 pixels (all shorter lines were removed during the preprocessing step), we set the maximum search depth $n = 5$.

$$\vec{z_i} = \begin{pmatrix} x_{E_i} \\ y_{E_i} \end{pmatrix} - \begin{pmatrix} x_n \\ y_n \end{pmatrix} \tag{4.2}$$

Finally, we get the orientation vector $\vec{a_i}$ by normalizing $\vec{z_i}$

$$\vec{a_i} = \frac{\vec{z_i}}{\|\vec{z_i}\|} \tag{4.3}$$

For narrowing the search space of possible connections, a region of interest (ROI) is defined for each line ending E_i of a component C_i. The ROI is a polygon spanned by the output vector $\vec{a_i}$ of E_i as well as two vectors $\vec{p_i}$ and $\vec{n_i}$ for the angular tolerance and two vectors $\vec{l_i}$ and $\vec{r_i}$ in an orthogonal direction to $\vec{a_i}$. The vector $\vec{p_i}$ is the angular tolerance in the direction of the curvature of C_i, whereas $\vec{n_i}$ is the angular tolerance in the opposite direction. As we want to construct lines that do not change their curvature, the angular tolerance for $\vec{p_i}$ is larger than the tolerance for $\vec{n_i}$. The length of the vectors $\vec{a_i}$, $\vec{p_i}$ and $\vec{n_i}$ is the maximum search distance value d_{max}, which is a predefined fixed value. The length of $\vec{l_i}$ and $\vec{r_i}$ is set to $\frac{d}{2}$ (see Figure 4.2 *(b)*). This specific shape of the ROI covers all points near the endpoint E_i while ignoring all points with a large distance. The orthogonal tolerance vectors are necessary because the ROI should also cover lines that are nearby E_i but not directly in front of it. Without the orthogonal offset, either the angular tolerance must be very broad or the maximum distance must be very large, which increases the chances for false connections during the connection step.

For each pair of possible connections, there are two angles α and β, such that $\alpha = e_{1,2} - a_1$ and $\beta = e_{2,1} - a_1$, where $e_i j$ is the vector between the endpoints and d is the distance between the endpoints. Because the connection is only evaluated if the corresponding endpoint is inside the ROI, d is always smaller than d_{max}. In case the two angles α and β are smaller than a given maximum angle, the connection is added to the list of plausible connections. Furthermore, if this list contains more than one possible connection between two endpoints of the components C_1 and C_2, all plausible connections between these points are ranked by using a quality score Q. The likelihood for a connection between C_1 and C_2 increases if α and β and the distance between the endpoints d are small. Moreover, for small values of α and β, a larger value for d may be preferred to a connection with a small d but high values for α and β (see Figure 4.2 *(c)*).

$$Q = d(\alpha + \beta) \tag{4.4}$$

A connection between two components C_1 and C_2 is only established if C_2 has an endpoint in the ROI of E_1 (which is an endpoint of C_1) and their score Q is the smallest score of all possible connections in the Region of interest. Moreover, Q must not be larger than a given maximum score Q_{max} and it must not have any intersection points with itself.

After reconnecting components in the image, the ear's outline is now among the largest components. Because of the specific shape of the outer helix, all lines that consist of both positively and negatively curved parts are discarded, as they are unlikely to represent the helix. The algorithm only selects lines that are curved in only one direction. Therefore, the set of possible components representing the ear's outline can once again be reduced by selecting the ten largest components with a single direction of curvature as helix-candidates. In Figure 4.1, the highlighted lines are the ten helix-candidates selected from the set of lines after the reconstruction step. Note that many of the selected lines are already part of the ear. Other long and prominent lines that are frequently selected as helix candidates are hair or hair ties, the outlines of glasses or clothing. In the next step, the helix-candidates are further combined and evaluated to make sure that the only lines kept are those that form an ear.

4.2.3 Combining Segments and Evaluation of Helix Candidates

In the previous step, each component in the image was examined separately by selecting components due to their size and their curvature. It is, however, unlikely that the ear outline is reflected by only a single component. The algorithm therefore combines helix-candidates and selected additional components, and then estimates the likelihood that the

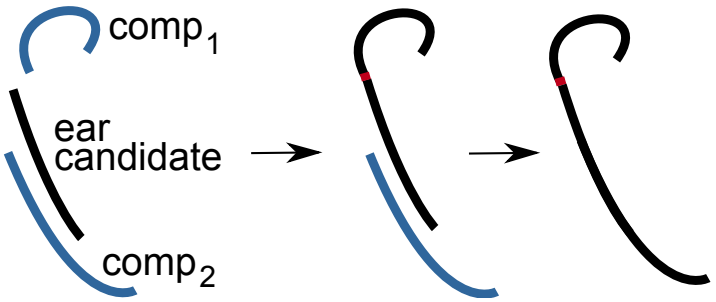


Figure 4.3: Combination of single shapes to create an ear candidate

resulting shape is an ear or not. Figure 4.3 shows an example of the combination of three components C_1, C_2 and C_3 to form a common structure, which reflects the ear contour better than each component alone.

One specific property of the helix contour with respect to the 2D binary image plane is, that is usually consists of a convex and a concave line, which are parallel in the projected 2D image. Therefore, the first combination step is to search for parallel components of helix-candidates in the image. These parallel components do not necessarily have to be helix-candidates themselves but can be chosen from all reconstructed components in the projected curvature image. The algorithm then combines non-parallel helix-candidates that are close to each other. For each of the possible combinations, a score is calculated which describes the likelihood that the combined shape represents the outline of an ear. If the score drops below a certain threshold, no ear is detected in the image. Otherwise, the combined shape with the largest score is chosen and is marked with a bounding box.

There are two different categories of criteria for evaluating the combined shape. The first category is that of absolute criteria, which are calculated individually for each combined shape. The second category consists of relative criteria, which enable the algorithm to weigh different ear candidates and mark the most likely ear region.

4.2.3.1 Absolute Criteria

Absolute criteria are used to make a coarse selection of possible combinations and exclude every combined shape that does not fulfill each of them. These criteria are the proportion B_{P_i}, the cumulative curvature B_{K_i}, the ratio of parallel shapes B_{R_i} and the number of corners B_{C_i}. Figure 4.3 illustrates the absolute criteria with the regards to a sample ear contour. The total score, which estimates the likelihood that the combined shape i is an ear, is denoted as A_i, and is the sum of each of the criteria.

$$A_i = \frac{1}{4}(B_{P_i} + B_{K_i} + B_{R_i} + B_{C_i}) \tag{4.5}$$

In order to prevent a combined shape from having a high score because it satisfies only one criterion extremely well, no score for a single criterion may be larger than 150% of the lowest score among the four. In this case, A_i is set to the value of the lowest score.

$$A_i = \min(B_{P_i}, B_{K_i}, B_{R_i}, B_{C_i}) \tag{4.6}$$

The ideal proportion of an ear is measured by calculating the ratio between the major and the minor axis of an ellipse that encloses the current combined shape. This ratio should be between $^2/_1$ and $^3/_1$. Any deviation from these ratios decreases the total proportion score B_{P_i} for a combined shape i.

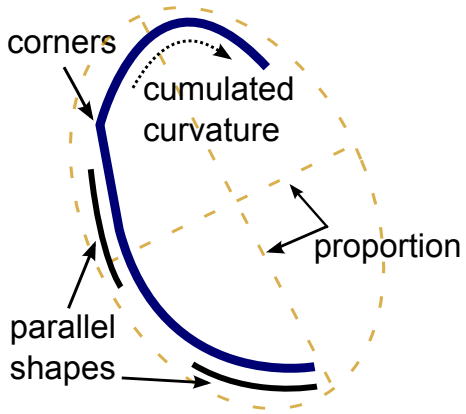


Figure 4.4: Visualization of the criteria for absolute score computation

$$B_{P_i} = \begin{cases} 1 - \frac{(2 - proportion_i)^2}{4} & 0 < proportion_i < 2 \\ 1 - \frac{(proportion_i - 3)^2}{4} & 3 < proportion_i < 5 \\ 1 & 2 \leq proportion_i \leq 3 \\ 0 & else \end{cases} \tag{4.7}$$

Additionally, a complete outline of an ear should have an accumulated curvature of approximately 2π. Values below 2π indicate that the outline is not complete, whereas values larger than 2π indicate that the currently evaluated shape is not an ear. The accumulated curvature K_i of a combined shape i is defined as the sum of all curvature values c_j of the n pixels on the shape's outline. The curvature values on the outline of the shape are calculated by using the method proposed in [78].

$$curvSum_i = \sum_{j=1}^{n} |c_j| \tag{4.8}$$

Then the total curvature score is defined as

$$B_{K_i} = \begin{cases} 1 - \frac{|2\pi - curvSum_i|}{1,5\pi} & 0,5\pi < curvSum_i < 3,5\pi \\ 0 & else \end{cases} \tag{4.9}$$

The third criterion B_{R_i} measures the number of pixels on the outline of the combined shapes, that have parallel shapes in their neighborhood. As stated earlier, the presence of parallel lines in the projected image is an important property for the ear outline, due to the outer ear's unique shape. The ratio of parallel components is the number of pixels contained in components that have a parallel line $p_{parallel}$ divided by the total number of pixels in the current combined shape p_{total}.

$$B_{R_i} = \frac{p_{parallel}}{p_{total}} \tag{4.10}$$

The last absolute criterion is the number of corners in the combined shape. This criterion is used to exclude jagged and noisy lines, which often represent hair or clothes. The outline of the helix should be smooth and hence, no corners should be present in an optimal combined shape representing an ear, and their contribution to the total curvature of the shape is 0. In order to find corners in a given combined shape, we use the corner detector described in [78]. In order to determine B_{C_i}, the ratio between the total angle accumulated

in corners θ_i and the total accumulated angle of the whole combined shape Θ_i is calculated.

$$B_{C_i} = 1 - \left(\frac{\theta_i}{\Theta_i}\right)^2 \tag{4.11}$$

4.2.3.2 Relative Criteria

Due to the ear's self-similarity, the absolute criteria can also be fulfilled by combined shapes that only cover part of the ear. To overcome this, the algorithm additionally compares combined shapes, using relative criteria in order to select the largest and most complete combined shape. The relative score is composed of a bonus for large shapes l_i, which is the total number of pixels the connected shape consists of, and two non-linear penalty scores. g_i reflects the total distance in pixels, that had to be bridged during the reconstruction step, and m_i is the distance, in pixels, between the single components the combined shape is composed of. The exponent λ adapts the penalty score to the resolution of the projected 2D curvature image and can be a value between $[1, 2]$. For the models in UND-J2 a value of 1.2 proved to be a good choice for λ.

The non-linear weighting of the distance penalty makes sure that small distances are strongly preferred to high distances between the components. This makes the algorithm pick the largest but also most compact component in the image. Taking all these factors into account, the relative score N_i for each combined shape is

$$N_i = l_i - g_i^\lambda - m_i^\lambda \tag{4.12}$$

In order to be able to compare the relative scores with each other and to use them in the final evaluation of the combined shape, N_i is normalized to a value between $[0\ 1]$. $max(N)$, here, is the maximum score among all relative scores N_i for the current combined shape i.

$$R_i = \begin{cases} \frac{N_i}{max(N)} & N_i > 0 \\ 0 & N_i \leq 0 \end{cases} \tag{4.13}$$

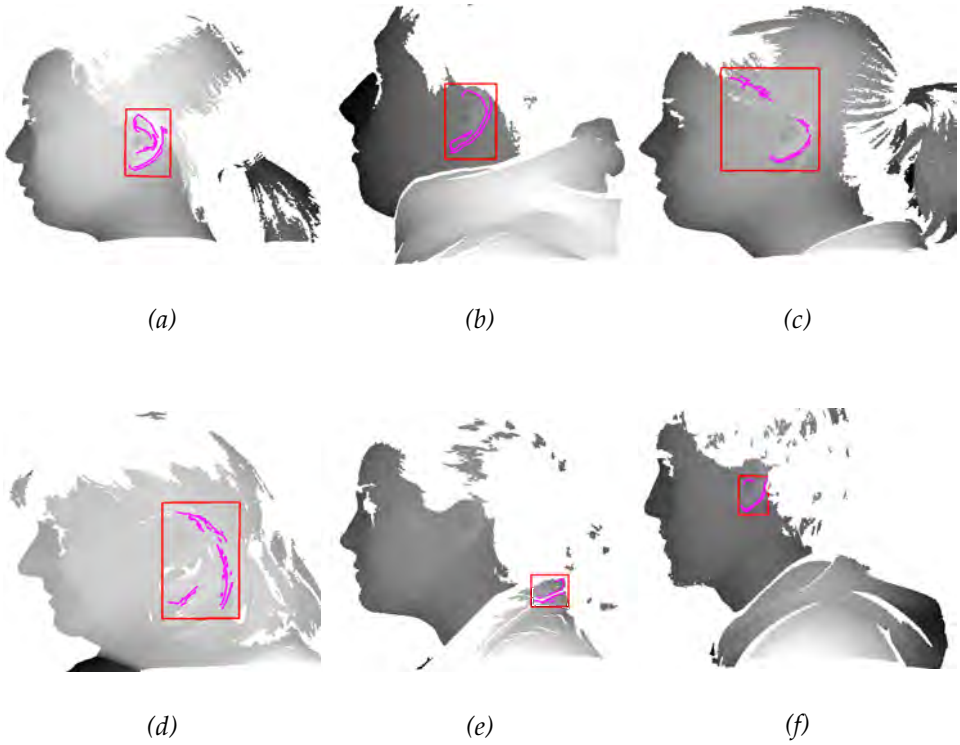
Finally, each combined shape i is evaluated by combining its absolute score A_i and the relative score R_i to create the final score S_i. A_i and R_i can be weighted to change the influence of each of the scores. A_i is more sensitive in differentiating between ear-like and other combined shapes, whereas R_i makes the algorithm choose the largest shape with its components close to each other. We assigned larger weight to A_i as this criterion appeared to be the stronger of the two.

$$S_i = \omega_1 A_i + \omega_2 \frac{1}{4} R_i \tag{4.14}$$

4.3 Detection Performance

We tested our approach on the 3D models contained in the UND-J2 collection [201]. The dataset consists of 404 subjects with up to 22 samples (3D models) per subject, with a total number of 2414 images. Figure 4.5 shows some examples for successful and unsuccessful detection results using our algorithm. As a result of our tests, we measured a detection rate of 95.65%.

In our test setup, the detection rate is defined as the percentage of overlapping pixels between a ground truth and the ear region marked by the algorithm. The ground truth was generated by manually cropping the ear region from the images. If more than 50% of the pixels in the ground truth and the detected ear region from the algorithm overlap, we consider the detection as successful. In Figure 4.6 the detection rate is plotted against the minimum grade of overlap between the ground truth and the detected region. Even if the



(a) *(b)* *(c)*

(d) *(e)* *(f)*

Figure 4.5: Examples for successful(subfigures *(a)* and **(b)**) and unsuccessful detections.

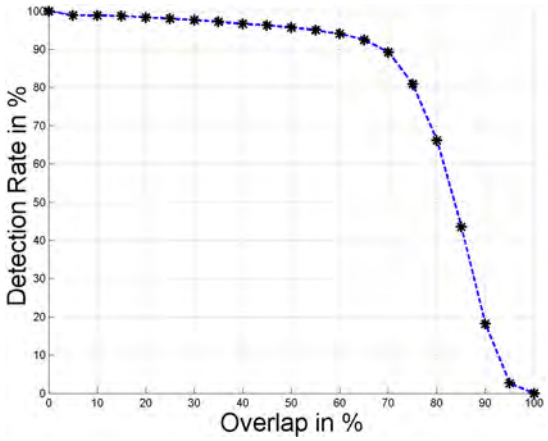


Figure 4.6: Detection rates for different grades of overlap between ground truth and detected region. Occlusion (if any) is reflected by the ground truth.

constraint for a positive detection were set to an overlap of at least 70%, 90% of all images would still comply.

In Figure 4.5 *(a)* and *b*, two examples of successfully detected ears are given. In general, our algorithm successfully detected the ear region, even when it was partially occluded by hair. Our approach, however, had problems distinguishing long hair, located near the ear region from the actual outline of the auricle (see Figure 4.5). In the binarized image, long hair is reflected by prominent and smoothly curved lines, which look similar to the shape of the outer helix. For example, in Figure 4.5 *(c)*, the segmented region is too large because a strain of hair is included in the set of shapes that represent a typical ear. This behavior is

caused by the relative score, which tends to favor the largest ear-like shape in the image. If there is a strain of hair, that fits the absolute criteria and expands the marked region, the ear detection considers it as a part of it.

In some cases the algorithm considers clothes, such as scarves and collars, as the ear region. This happens if the auricle's outline is not sufficiently represented in the curvature image, which is mostly caused by occlusion or missing data. In these cases, the detector incorrectly uses smoothly curved and parallel lines from other parts of the image, such as long hair strains or cloth.

4.4 Conclusion

Our method for ear detection based on surface curvature delivered promising results and outperformed comparable edge-based methods, such as the one proposed by Chen and Bhanu. The ear detection works reliably and is also capable of delivering valuable information for a subsequent feature extraction step. We are optimistic, that the performance of our method can be further improved by including the Tragus location and edge images, which are calculated from the corresponding 2D image.

Furthermore we plan to refine the measure for the accumulated curvature for more accurate curvature estimation. We will also conduct more experiments on additional databases in order to evaluate the robustness against pose variations. In addition, a feature extractor which uses the helix outline, which is returned by our ear detection algorithm, will be developed.

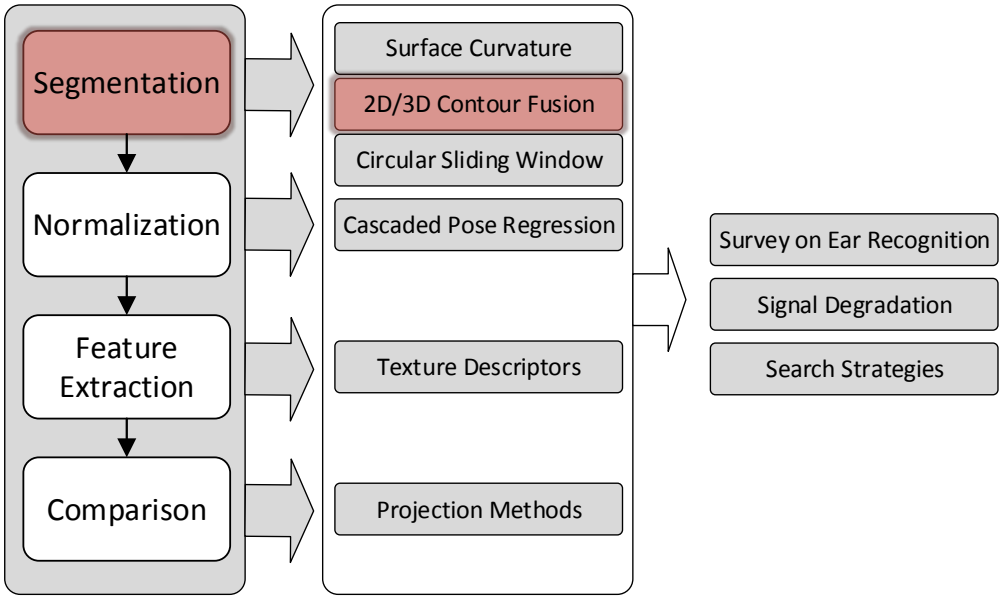
Chapter 5

Robust Localization of Ears by Feature Level Fusion and Context Information

Based on the promising performance and the finding in the previous chapter, we extend our method to use the texture and the depth channel with the goal of further improving the detection performance and give the second part of the answer to the research question **Q1: How can the outer ear be automatically detected from 2D and 3D images?**

We combine the texture and the depth channel in the image space by considering edges in the texture images and curvature in the depth images. Our extended detection method also contains an improved scoring method for the ear shapes, which also considers the relative size and position of the shape with the size of the silhouette of the head.

The paper was published in [155]ANIKA PFLUG, ADRIAN WINTERSTEIN, CHRISTOPH BUSCH, Robust Localization of Ears by Feature Level Fusion in Image Domain and Context Information, 6th IAPR International Conference on Biometrics (ICB), 2013



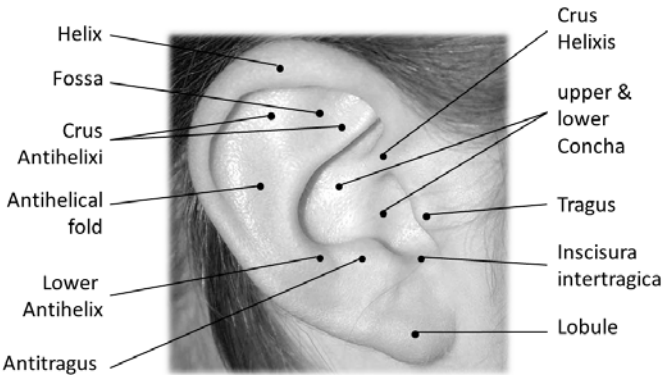


Figure 5.1: Morphology of the outer ear.

Abstract

The outer ear has been established as a stable and unique biometric characteristic, especially in the field of forensic image analysis. In the last decade, increasing efforts have been made for building automated authentication systems utilizing the outer ear. One essential processing step in these systems is the detection of the ear region.

Automated ear detection faces a number of challenges, such as invariant processing of both left and right ears, as well as the handling of occlusion and pose variations. We propose a new approach for the detection of ears, which uses features from texture and depth images, as well as context information. With a detection rate of 99% on profile images, our approach is highly reliable. Moreover, it is invariant to rotations and it can detect left and right ears. We also show, that our method is working under realistic conditions by providing simulation results on a more challenging dataset, which contains images of occluded ears from various poses.

5.1 Introduction

The observation of the outer ear, which is frequently referred to as auricle, is an emerging biometric method, which has drawn the attention of research during the last years. In forensic investigation, ear prints on doors and windows can be collected and are used as a means for identifying perpetrators [13].

The outer ear is believed to be just as unique as the face. An extensive study of Iannarelli [81] and a more recent study from India [176] show, that the outer ear possesses numerous characteristics, which make each ear unique. In Fig.5.1, we have annotated a number of features, which are used by the German Criminal Police Office for identifying a subject.

Forensic investigators do not only value the uniqueness of the outer ear, but also its permanence. In contrast to the face, the structure of the outer ear remains stable after the 6th month of life [81]. A recent study by Ibrahim *et al.* [82] confirms that the recognition rate of a biometric system is not affected considerably over eleven months.

Due to the proximity of the observed physiological characteristics, ear recognition may be considered as a valuable extension to face recognition systems. Ear and face images can be collected with the same hardware. Especially in unconstrained scenarios, such as video surveillance, the outer ear can contribute additional features, which increases the chance of identifying a person in off-pose images.

The contribution of this paper is a novel ear detection algorithm, which uses texture and depth images for localizing the ear in full profile as well as in off-pose images. We utilize the rich details on the ear surface and of edge images for determining the ear outline in an image. We present a set of flexible rules, which allow us to distinguish between the

Table 5.1: Comparison between the proposed method and previously achieved results on UND-datasets.

Author	Performace	Remarks
Yan and Bowyer[201]	97.38% rank-1	UND-J2 collection, depth and color images, no details about detection accuracy
Chen and Bhanu [49]	87.11%	UND collection F and subset of collection G, depth and color images
Prakash *et al*. [157]	99.38%	Subset of UND-J2 collection, depth images only, 10% of images excluded
Zhou *et al*. [218]	98.22%	Results on UND-J2 based on re-implementation in [146], depth images only
Islam *et al*. [84]	100%	203 profile images from UND-J2 database, 2D images only
Pflug *et al*. [154]	95.65%	UND-J2 collection, depth images only
Proposed method	99%	UND-J2 collection, depth and color images

ear outline and other objects in the image. These rules describe an abstract ear model and include context information. Our algorithm is invariant to rotations and it can detect left and right ears with the same parameter set. Moreover it is robust to pose variations and occlusion. The feasibility of the proposed ear detection system is shown by providing simulation results on the UND-J2 database [201] as well as UND-NDOff-2007 [64]. A detection rate of 99% on the UND-J2 dataset shows that our approach is outperforming other recent work. Moreover, we also show that our method has the ability to detect ears under realistic conditions, where it has to handle occlusions and pose variation.

In the upcoming section, we describe the state of the art in ear detection in 3D images and describe related work. Subsequently, we introduce the proposed algorithm. In section 5.4 we point out the experimental setup and present simulation results using the previously mentioned datasets. Finally, conclusions are drawn and an outlook on future work is given in Sec. 5.5.

5.2 Related Work

In contrast to 3D meshes, depth images have a matrix-like data structure and they can be acquired along with the texture image with a single capture process. If depth images are recorded under controlled conditions, they are co-registered with the texture image, which makes it easy to combine texture and depth information. This fact has inspired a number of researchers to come up with different approaches, which use depth and texture information for ear detection and recognition. Many of these early methods are developed and tested on the public dataset collected by the University of Notre Dame, such as the collections F and G, and especially collection UND-J2 [201] which is the largest publicly available database for ear recognition. It consists of texture and depth images of left profiles. In Tab. 5.1, we compare the detection rates of recent ear detection algorithms, which were tested on the UND-J2 dataset.

Yan and Bowyer [201] propose a full biometric ear recognition system based on the profile images of the UND-J2 collection. They first locate the nose tip and then use skin color from the texture image to narrow the search space for the ear region. Subsequently the lower concha is detected by using curvature information from the corresponding depth image. The final ear region contour is fixed by using an active contour approach.

Chen and Bhanu [49] also combine texture and depth images in their detection approach. First, they create different image masks from skin color and depth information. Then this mask is used for selecting edges from the depth and the texture image. The ear detection is completed by clustering these edges and selecting the largest cluster of edges in the superimposed image.

In the ear recognition system proposed in [84], Islam *et al*. use Haar-like features for building an ear classifier for 2D images. Because the texture and the depth images are co-registered, the detected ear region in the texture image and in the depth image have the same position.

More recently, several ear detection approaches were proposed, which exploit the properties of the detailed surface structure of the auricle. Zhou *et al*. [218] extract local histograms of categorized shapes from a sliding window and use a SVM for deciding whether a local histogram represents an ear or not. Subsequently the largest cluster of positive detections is selected as the ear region. In their paper, Zhou *et al*. only provide results on a subset of UND collection F, where this algorithm achieved a detection rate of 100%. However, the detection rate drops significantly, when the ear is rotated by more than 10 degrees [146].

Another class of ear detection algorithms are approaches, which use specific line-like patterns for localizing the outer ear. Prakash *et al*. [157] define edges in the 3D image as regions with maximum depth variation. Based on his assumption, an edge map is created from the depth image. Subsequently, the local orientations of the edges are approximated with vectors. These vectors serve as the edges of an edge connectivity graph. Subsequently their algorithm generates potential ear candidates from these graphs and then selects the final ear region by comparing the each candidate with a reference template.

In [154] ears are detected using the specific distribution of surface curvatures in the ear region. This results into convex and concave edges, which are then processed in a number of subsequent steps for combining the multiple neighbored edges to ear candidates. According to its proportion, size, redundancy and cumulated slope, a score is assigned to each candidate. The final ear region is defined by the circumference of the ear candidate with the highest score. This approach, however, has some limitations if the depth image is noisy or if the ear contour is on the verge of the depth image. In these cases, no meaningful curvature information can be extracted from the depth image and the detection fails. In Tab.5.1, the reported detection accuracy of the cited previous work is summarized and compared to the proposed detection method. With a detection accuracy of almost 99% on the UND-J2 dataset, the proposed ear detection approach is not only competitive to the graph based approach by Prakash *et al*. with respect to the detection performance - moreover the proposed method has no need to exclude challenging samples from the evaluation as it was done in [157].

5.3 Ear Detection System

The proposed ear detection algorithm utilizes the fact that the surface of the outer ear has a delicate structure with high local curvature values. In some depth images, however, some parts of the ear are missing, because curvature can only be measured between neighboring points. However, many depth images have holes next to the outer helix, which results in missing curvature values [154] (see Fig. 5.2 for an example). We solve this problem by fusing co-registered edge and depth images.

Our ear detection approach consists of four different steps, which are illustrated in Fig.5.2. We start with a preprocessing step, where edges and shapes are extracted from the texture and the depth image. Subsequently, the edges and shapes are fused in the image domain. In the next step, the components are combined with each other to ear candidates and a score for each ear candidate is computed. Finally, the enclosing rectangle of the best ear candidate is returned as the ear region.

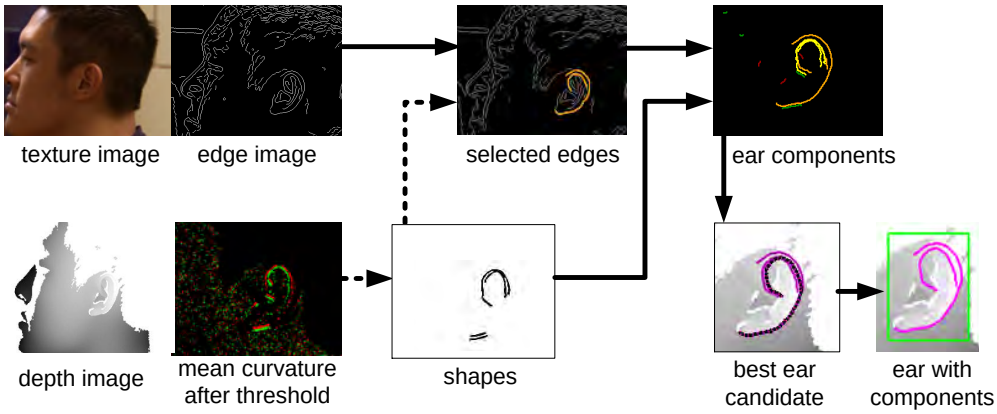


Figure 5.2: Illustration of the processing steps of the proposed ear detection system (database images taken from [201]).

5.3.1 Preprocessing

The outcome of the preprocessing step is an edge image from the texture, which is created using a Canny edge detector [42]. Furthermore, we determine a number of shapes from the depth image, which serve as the basis for the subsequent fusion step.

We first calculate the mean curvature [28] of the depth image. The key points on the ear surface have large convex and concave curvature values. Therefore, we apply a threshold to the mean curvature image in order to keep only large convex and concave curvature values. We now delete all connected components from the image, which are smaller than 3 pixels. As a result, we get a number of blobs, which are mainly located in the ear region. These blobs are thinned to a width of 1 pixel and subsequently re-connected using the method proposed in [154]. After the reconnecting step, a limited number of lines is left from the mean curvature image. For the remainder of this paper, these lines are referred to as shapes.

5.3.2 Fusion of Texture and Depth Information

In the fusion step, we select a number of edges from the edge image, based on the vicinity to the position of the most prominent shapes and other criteria. First, we select the ten longest shapes from the mean curvature image.

If the sign of the local orientation of one of the selected shapes changes, the shape is divided into two shorter shapes. Shapes can also be split up if they contain a corner. The two resulting shapes may be omitted, if there exists another shape in the curvature image, which is longer than any of the two divided shapes. The result of this procedure is a set of the longest shapes, which are smoothly curved.

Each shape in the set is dilated with a circular structuring element. After the dilation, each shape represents a region, which is now used for selecting appropriate edges from the edge image. As shown in Fig.5.3, we select all edges, that exceed a predefined threshold with their intersecting points in the dilated region. For edges, which have an endpoint inside the dilated region, the minimum number of intersecting pixels should be smaller than for other edges. This is due to the fact, that edges with intersecting endpoints play a special role in the upcoming combination step and also lead to better results in the scoring step.

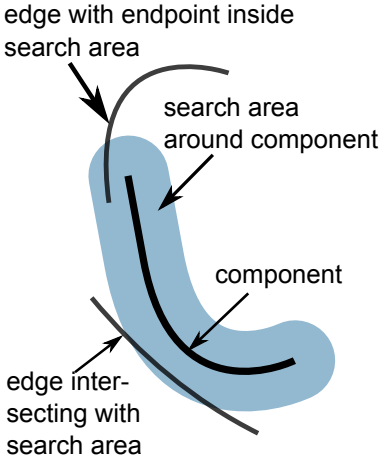


Figure 5.3: Fusion of 2D and 3D data. Shapes from the depth image are dilated and intersecting edges are selected.

All intersecting edges are added to the set of components. In case, an intersecting edge has a corner or if the sign of the local orientation changes, it is divided into two smaller edges. A divided edge that does not have an intersecting point with the dilated region, is removed from the set of components.

5.3.3 Combination of Components

In the combination step, all components are combined with each other or smaller shapes nearby in order to obtain complete ear outlines. In a first combination round, we only combine components with each other. In a second round, we also allow other shapes and edges that were not selected as components to be combined with ear candidates. This is done by picking a component and then combine it with other components. A component that has been combined with other components is considered an ear candidate. Thus we create a new ear candidate, each time a component is linked with one of the existing ear candidates.

During the creation of new ear candidates, a component can be adapted to the ear candidate in a number of ways. The component can either be translated, pruned or interpolated in order to fit to the ear candidates. In Fig. 5.4, an example for the combination of an ear candidate and two components is shown. When $comp_1$ is combined with the ear candidate, we have to translate the contributing component $comp_1$ and fill remaining gaps. In the second step, $comp_2$ is translated and pruned.

Each time, a new ear candidate is created, we compute a score, that describes the fitness of this candidate (see Sec. 5.3.4.3 for more details on the scoring system). However, without any additional constraints, we would have to combine all components with each other. Many of the created ear candidates would be redundant and the detection would be inefficient. Therefore, we introduce a terminating criterion, which prevents the algorithm from doing an exhaustive search through all possible combinations between ear candidates and components.

Let $maxScore$ be the best score, that has been achieved by any ear candidate for all components. A newly created candidate will only be used in subsequent combination steps, if its score is higher than $0.7 * maxScore$. An ear candidates with a lower score will be discarded and is not used in subsequent iterations. In the first iteration, it is very likely that an ear candidate will satisfy this condition and a large number of new candidates will be created. However, as the combination step proceeds further, an increasing number of new

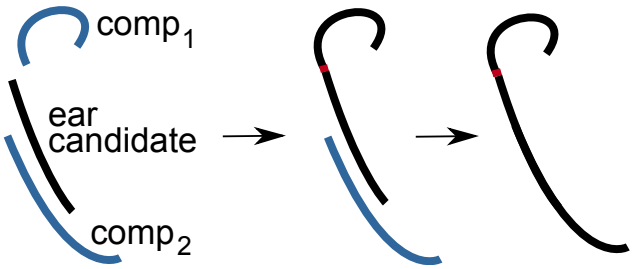


Figure 5.4: Iterative creation of an ear candidates though stepwise combination and adaptation of components.

candidates will be discarded. The more iterations have been completed, the higher the probability, that an ear candidate with a high score has been created and the more candidates are dropped.

5.3.4 Scoring system for ear candidates

Each time an ear candidate and a component are combined, we assign a fitness score to the newly created ear candidate. The score reflects the similarity of the ear candidate with an idealized ear outline. This similarity is expressed through a number of criteria, a good ear candidate has to comply with. The fitness score is composed of three different components, which reflect different properties. We distinguish between the individual score that is computed for each ear candidate, a relative score and a context score.

After the combination step, we select the ear candidate with the highest score return its bounding box as the detected ear region. If we cannot find an ear candidate with a larger score than 0.5, we consider the ear to be occluded and do not return an ear region.

Let I_i be the individual score of the ith ear candidate, R_i the relative sore and C_i the context score. Furthermore, let ω_1, ω_2 and ω_3 be weighting factors for each of the scores. The total fitness score of the ith ear candidate, denoted by S_i can be expressed as follows:

$$S_i = \omega_1 I_i + \omega_2 R_i + \omega_3 C_i \tag{5.1}$$

The values for ω should be adapted according to the variance of each partial score I, R and C. The larger the variance between single ear candidates, the higher the weight of the according partial score.

5.3.4.1 Individual Score

The individual scores consists of three components, which measure the cumulated sum of local orientations, the contribution to the sum of local orientations in corners and the proportion.

The sum of local orientations reflects the expectation, that an ear candidate should be convex and it should have a gap (connecting line between $p1$ and $p2$ in Fig. 5.5). We first compute the convex hull of an ear candidate and define two points p_1 and $p2$. The points p_1 and $p2$ are the points in $A \cup B$ with the maximal distance, such that all pixels on the connecting line $\overline{p_1 p_2}$ between them are on the convex hull but not in $A \cup B$.

As shown in Fig. 5.5, p_1 and $p2$ are on the opposite side of the outer helix. We can define a third point p_3, which divides the ear candidate pixels on the convex hull into two equally sized subsets A and B. Finally, we distinguish between those pixels of the ear candidate that are on the convex hull and those that are not. The set of pixels, that are not on the convex hull is denoted as C.

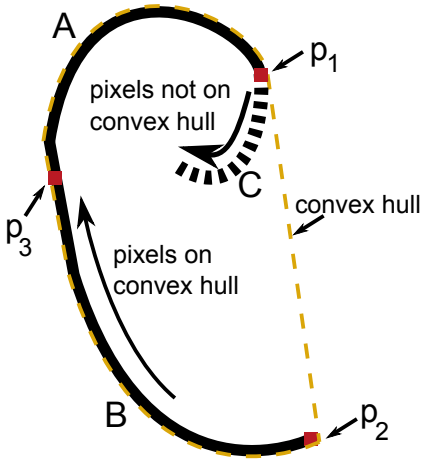


Figure 5.5: Calculation of the sum of local orientations using the convex full of an ear candidate.

Let sum_{hull} be the sum of all local orientations in $A \cup B$. If we have a good ear candidate, this sum is expected to be larger than 1.5π.

$$onHull = \begin{cases} \frac{sum_{hull}}{1.5\pi} & sum_{hull} \leq 1.5\pi \\ 1 & otherwise \end{cases} \tag{5.2}$$

We also expect a good and complete ear candidate to be symmetric. This can be expressed by comparing the sum of local orientations in A, denoted as sum_A and B, denoted as sum_B, with each other. The division by π is necessary for getting a normalized result between 0 and 1.

$$ratio = \begin{cases} \frac{|sum_A - sum_B|}{\pi} & |sum_A - sum_B| \leq \pi \\ 1 & otherwise \end{cases} \tag{5.3}$$

Based on the curvature sum of all pixels in $A \cup B$, we also calculate a weighting factor, denoted as λ, for the score contribution of $onHull$ and $ratio$.

$$\lambda = \begin{cases} \frac{1}{onHull} & onHull \leq \frac{1}{3} \\ 1 & otherwise \end{cases} \tag{5.4}$$

The outer helix score I_o, that measures the fitness of an ear candidate with respect to its cumulated sum of local orientations is composed of the previously defined coefficients $onHull$ and $ratio$, the weighting factor λ and a fourth component, which is a penalty score. For a good ear candidate, the sum of local orientations in C, denoted as sum_C, should be as small as possible. We hence subtract $\frac{sum_C}{20}$ from the total fitness score. The value of the denominator has been obtained empirically.

$$I_o = (1 - \lambda) * onHull + \lambda * ratio - \frac{sum_C}{20} \tag{5.5}$$

The larger the difference between sum_A and sum_B, the higher the influence of $ratio$ and the lower the influence of $onHull$. This reflects the fact, that incomplete ear candidates with a small sum of local orientations should get a better score than those with sum of local orientations close to 2π. The algorithm will hence be less strict with incomplete ear candidates than with complete ones.

In addition to the measure of the distribution of local orientations on the ear candidate, we also require, that a good ear candidate should have as few corners as possible. Corners are an indication for jagged components or failures during the combination step. We hence compare the ratio between the sum of local orientations in $A \cup B \cup C$, denoted as sum_{ABC}, and the sum of local orientations at all corners of the component θ. For a good ear candidate, θ should be as small as possible. In order to increase the impact of this criterion in cases where θ is large, we use the quadratic function.

$$I_c = 1 - \left(\frac{\theta}{sum_{ABC}} \right)^2 \tag{5.6}$$

In the last criterion for the individual score, we assume that the ratio between the major and the minor axis of a surrounding ellipse should be between $2:1$ and $3:1$. Let ρ be the ratio between the major and the minor axis of an ear candidate. The proportion score I_p decreases faster with larger deviations from the ideal ratio.

$$I_p = \begin{cases} 1 - \frac{(\rho-2)^2}{4} & 0 < \rho < 2 \\ 1 & 2 \leq \rho \leq 3 \\ 1 - \frac{(\rho-3)^2}{4} & 3 < \rho < 5 \\ 0 & otherwise \end{cases} \tag{5.7}$$

All components of the individual score are normalized values between 0 and 1, whereas higher scores represent better ear candidates. The individual score for the ith ear candidate I_i is the mean value of the three components for this candidate I_{o_i}, I_{c_i} and I_{p_i}.

5.3.4.2 Relative Score

The relative score compares different ear candidates with each other and is calculated in two steps. This score rewards ear candidates, if they are composed of long neighbored components. First, we calculate a base score, called η. The base sore is centred on the total length of the ear candidate in pixels l. Because this candidate was built by reconstruction of the ear outline, we subtract the number of pixels that had to be filled in during the combination step g and the sum of all distances between all components the current ear candidate is composed of. This sum is denoted as m.

$$\eta = l - g - m \tag{5.8}$$

Subsequently, we normalize this score by dividing it by the maximum value of η for all the ear candidates in the image. The usage of the quadratic function ensures, that ear candidates, that only have a slightly smaller score η than the maximum η are rejected. The relative score for the ith ear candidate R_i in the fused image is defined as follows:

$$R_i = \begin{cases} (\frac{\eta_i}{\max(\eta)})^2 & \eta_i > 0 \\ 0 & otherwise \end{cases} \tag{5.9}$$

5.3.4.3 Context Score

The scoring system is completed with an estimation of the ear candidate's size in relation to the size of the silhouette of the head in the depth image. We assume that a good ear candidate is located in the head region of the image and that the ratio between the height of the ear and the diameter of the face should be within fixed bounds. These bounds depend on the image scale and should be set individually for each capture device. We denote the lower bound for the ratio between h and d as τ and the upper bound as υ, respectively.

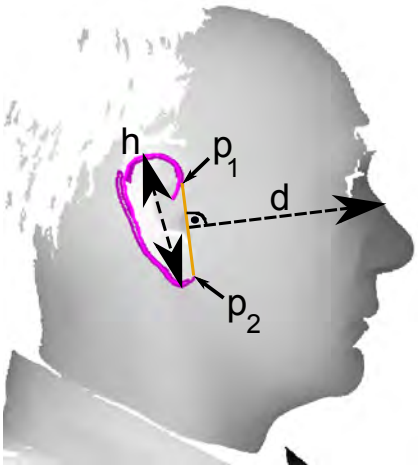


Figure 5.6: Estimation of optimal size using context information from the silhouette (Depth image taken from [201]).

Moreover, let h be the largest distance between any two points of the ear candidate. Furthermore, d is the distance between the edge of the silhouette and the point in the middle of line segment $\overline{p_1p_2}$. As shown in Fig.5.6, the face diameter d is measured orthogonally to the line segment $\overline{p_1p_2}$. For all yaw poses, the outer ear is located at the back of the head and d should be pointing towards the nose tip.

We hence assume that for a good ear candidate, d_i should be among the longest diameters for all ear candidates, denoted by D. If d_i is smaller than the mean value of the larger half on D, the ear candidate is rejected. If not, we compute the ratio between h_i and d_i.

$$cr_i = \frac{h_i}{mean(\{d_i \in D \quad | \quad d_i > 0.5\max(D)\})} \tag{5.10}$$

Dependent on cr_i, we can now compute the context score C_i. Note that the context score decreases faster if cr_i is too small. This reflects that the outer ear can be relatively large, compared to the face diameter. We hence prefer to keep large ear candidates. If the ear candidate is too small, however, it should receive a low score.

$$C_i = \begin{cases} 1 & \tau \leq cr_i \leq \upsilon \\ 1 - 2(\tau - cr_i)^2 & cr_i > \tau - \frac{1}{2} \\ 1 - \sqrt{cr_i - \upsilon} & cr_i < \upsilon + 1 \\ 0 & otherwise \end{cases} \tag{5.11}$$

5.4 Experimental Setup and Results

For obtaining the detection performance of our approach, we conducted two experiments on two different datasets. In the first experiment, we evaluate the impact of the image domain fusion. This experiment is conducted on the UND-J2 dataset [201] and on the UND-NDOff-2007 dataset [64]. In the second experiment, we show the robustness of our approach to rotation and flip.

The UND-J2 collection contains 1776 unique left profile images from 404 different subjects (-90 degrees yaw pose). Four images had to be removed from the database, because their texture and the associated depth images did not belong to the same subject. The

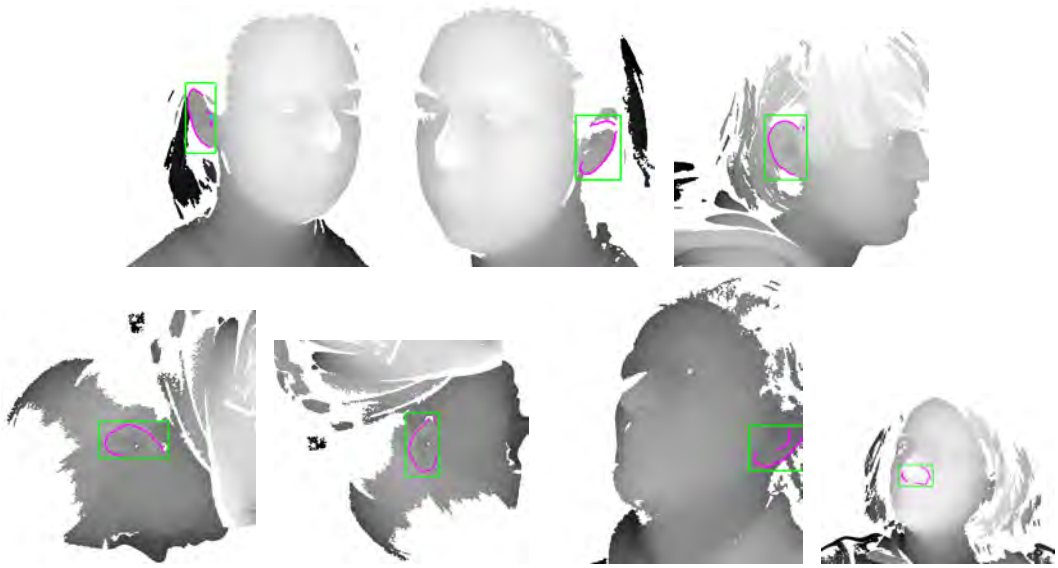


Figure 5.7: Examples for successful detections for left and right ears in images with pose variation, partial occlusion, missing depth data and rotation. We also see an example, of an ear candidate with missing components a detection failure (Original images were taken from[201] and [64])

.

UND-J2 dataset also contains some images, which are off pose. However, we did not exclude these images from our test set.

The UND-NDOff dataset was originally collected for the purpose of face recognition research. For an ear detection system, it represents a more realistic, but also more challenging scenario, than the profile views of UND-J2. We selected 2785 images with yaw poses between ± 90 and $\pm$ 30 degrees, whereas yaw poses of ± 90 are profile images and yaw poses of ± 30 are half profiles. In some cases the data collection contains different combinations between yaw and pitch poses. If different pitch poses are available for a given yaw pose, all pitch poses were included to the test set. The UND-NDOff-2007 dataset contains images where the ear is partly or fully occluded by hair and some subjects are wearing earrings. There is also a small number of images, were the subject has moved during the acquisition, which results in poor image quality.

The detection rates are calculated on the basis of manually created ground truth data. For creating the ground truth, we marked the ear region with a surrounding rectangle, and stored the coordinates of the upper left and the lower right corner. A detection for a given image pair is considered to be successful, if the overlap O between the ground truth pixels G and the pixels in the detected region R is at least 50%.

$$O = \frac{2|G \cap E|}{|G| + |E|} \tag{5.12}$$

5.4.1 Impact of Image Domain Fusion

In the first experiment, we show the impact of the image domain fusion and the context score on the detection results. In Fig. 5.7, some examples for detected ears with partial occlusion and different poses are shown. The detection rates in Tab. 5.2 show, that the proposed ear detection algorithm is robust to pose variation. Although the detection rate drops, with increasing deviation from ± 90 degrees, it still detects more than 75% of the ears

Table 5.2: Comparison between the detection rates with and without image domain fusion on UND-J2 [201] and UND-NDOff-2007 [64].

dataset	yaw pose	depth only, no context	fused with context
UND-J2			
	-90	92,9%	99%
UND-NDOff-2007			
	-90	86,9%	96,5%
	-60	70,9%	83,5%
	-45	50,5%	76,5%
	-30	23,7%	58,9%
	30	19,8%	42,7%
	45	49,4%	76%
	60	86,1%	85%
	90	91,8%	93,5%

correctly, if the yaw pose is $\pm$45 degrees. This also includes images, where the ear is partially occluded, as shown in Fig.5.4 and images, where the algorithm correctly recognizes, that the ear is occluded.

Image domain fusion and the usage of context information generally improve the detection rate of the proposed method. The improvement gets more significant with larger deviations from $\pm$90 degrees yaw pose. We can extract good ear candidates form profile images, even without image domain fusion. Moreover, the number of ear candidates, that get rejected though the context score is small. With larger pose variations, the probability increases, that the 3D data in the ear region is of low quality and hence that many shapes from other image regions are selected. Further, we get an increasing number of false ear candidates from the depth image. By using image domain fusion and context information, we can give preference to the correct ear candidate. From this we can conclude, that the usage of context information substantially contributes to the algorithm's robustness to pose variation.

In some cases, the detected ear region is too small (see Fig.5.4), because the algorithm fails to find all necessary ear components. This happens, when the number of shapes from the depth image is not sufficient or if the edges in the texture images are interrupted. This issue, however, can be addressed by allowing the algorithm to choose more shapes from the depth image before starting the image domain fusion.

Especially for images of yaw poses $\pm$30 degrees, there is another common type of error. Fig.5.4 illustrates an example. This error is mainly responsible for the drop in the detection rate at $\pm$30 and occurs if the ear is not visible in the image. In these cases, the algorithm selects shape from the nose or the eye region and creates ear candidates from them. Often, these ear candidates are rejected because of their low context score, but it happens that they are good enough for not being rejected. If there is no better ear candidate available, the algorithm will then mark it as the ear region.

5.4.2 Rotation Invariance

For evaluating the rotation invariance, we rotated the images from the UND-J2 dataset by 90 degrees clockwise and anticlockwise. Furthermore, we have run simulations on images rotated by 180 degrees and on images that have been mirrored (also referred to as vertical

Table 5.3: Results on rotated and flipped images from UND-J2 [201].

image orientation	**detection rate**
No rotation	99%
180 degrees	98,6%
Mirrored	99%
90 degrees clockwise	99%
90 degrees anticlockwise	98,8%

flip). As it can be seen in Tab.5.3, the detection rate stays stable for all rotations. Two examples for successfully detected ears in rotated images are shown in Fig.5.4 and 5.4.

The detection rates for left and right profiles (yaw pose $\pm$90 degrees) in Tab.5.2 in connection with the results on the mirrored images, stresses, that the proposed method can be used for left and right ears, without changes in the parameter set.

5.5 Conclusion

In this paper, we have presented a new approach to ear detection, which proposes a scoring system for ear components that are derived from co-registered texture and depth images. The proposed method utilizes the distribution of local orientations, the length of components and context information for detection of the outer ear in images from multiple poses and for left and right ears. Moreover, we have shown that our algorithm is invariant to rotation and robust to partial occlusion, while maintaining the same detection accuracy as previous work.

Our algorithm does not only localize ears, but also estimates their orientation, which is important for normalization. In the future, we plan to use the ear outlines, which are a side-product of the localization approach as a basis for normalization and feature extraction. At the same time, we are planning to improve the robustness to pose variation, by conducting more experiments on context information. Additionally, we plan to improve the throughput by exploiting the fact, that the combination step can easily be parallelized.

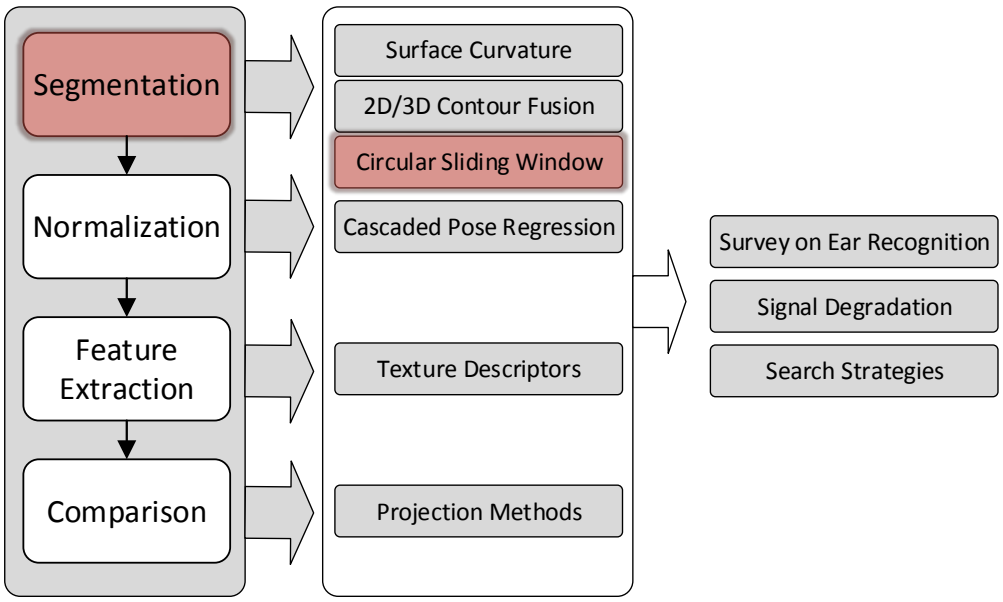
Chapter 6

Towards Ear Detection that is Robust Against Rotation

This chapter provides the third and last aspect of research question **Q1: How can the outer ear be automatically detected from 2D and 3D images?** In this work, we focus on the question, if sliding window detection approaches can be made invariant to in-plane rotations by using a circular detection window.

In this work, we extend an existing approach for segmenting ears in depth images by making it invariant to in-plane rotations. The original approach is using a rectangular sliding window from which a block-wise histogram descriptor is created. We propose to use a circular detection window instead and show that the circular window makes the ear detection invariant to in-plane rotations. The limitation of this approach is, that the projection into the polar coordinate system causes a loss of image data though merging or stretching single pixels in order to assign them to the correct radius. A suitable countermeasure could be a weighting function that assigns weights to each block according to the relative number of pixels it is representing.

The paper was published in [146]ANIKA PFLUG, PHILIP MICHAEL BACK, CHRISTOPH BUSCH, Towards making HCS Ear detection robust against rotation, International Carnahan Conference in Security Technology (ICCST), 2012



Abstract

In identity retrieval from crime scene images, the outer ear (auricle) has ever since been regarded as a valuable characteristic. Because of its unique and permanent shape, the auricle also attracted the attention of researches in the field of biometrics over the last years. Since then, numerous pattern recognition techniques have been applied to ear images but similarly to face recognition, rotation and pose still pose problems to ear recognition systems.

One solution for this is 3D ear imaging. The segmentation of the ear, prior to the actual feature extraction step, however, remains an unsolved problem. In 2010 Zhou at al. have proposed a solution for ear detection in 3D images, which incorporates a naive classifier using Shape Index Histogram. Histograms of Categorized Shapes (HCS) is reported to be efficient and accurate, but has difficulties with rotations.

In our work, we extend the performance measures provided by Zhou et al. by evaluating the detection rate of the HCS detector under more realistic conditions. This includes performance measures with ear images under pose variations. Secondly, we propose to modify the ear detection approach by Zhou et al. towards making it invariant to rotation by using a rotation symmetric, circular detection window. Shape index histograms are extracted at different radii in order to get overlapping subsets within the circle. The detection performance of the modified HCS detector is evaluated on two different datasets, one of them containing images in various poses.

6.1 Introduction

Ear Recognition is an emerging biometric, which has drawn increasing attention of researches during the last years. Similarly to the face, the outer ear (auricle) is believed to be unique. Additionally, the outer ear is particularly stable over a lifetime, easy to capture and widely accepted. Even though, there are no long-term studies yet, observations by Iannarelli [81] and Meijermann et al. [125] indicate, that the auricle is stable throughout a longer period.

The auricle has a delicate and richly structured surface, which has been used for manual identification by criminal investigators. In a biometric authentication scenario, the ear could extend facial features in order to enhance the accuracy of this system. Moreover, the usage of the outer ear can help to build biometric authentication and identification systems in unconstrained scenarios. The main advantages of the ear are its permanence and its acceptability. The auricle's shape remains stable after the first moths of a human life. Additionally it is not affected by facial expressions. A survey by Choras et al. [50] could also confirm, that many people do not have concerns about their appearance during the capture process. Hence, ear recognition systems are believed to be more convenient and more acceptable than face recognition stems.

In order to build automated ear recognition systems, using 3D ear images, it becomes necessary to reliably segment the ear from the input image. Chen and Bhanu as well as Yan and Browyer suggest using 2D photographs, which are generated when using a stereo vision device.

In [201] they detect the skin region and the nose tip in profile images in order to reduce the search space. Then, a modified active contours method is used to finally locate the ear.

Chen and Bhanu [49]locate the ear in profile images with a reference ear model, which reflect the contour of the outer helix and the concha region. They exploit the fact, that high curvature values can be measured between the edge of the outer helix and the surface of the head. For locating the ear among several candidates, the template is fitted to the candidates and the candidate with the lowest registration error is picked.

In [157], Prakash and Gupta an ear detection technique for 3D profile image, which makes use of connected components, which are constructed by using the edges of the depth image. These edges are combined to a connectivity graph, which is then matched with a shape distribution template.

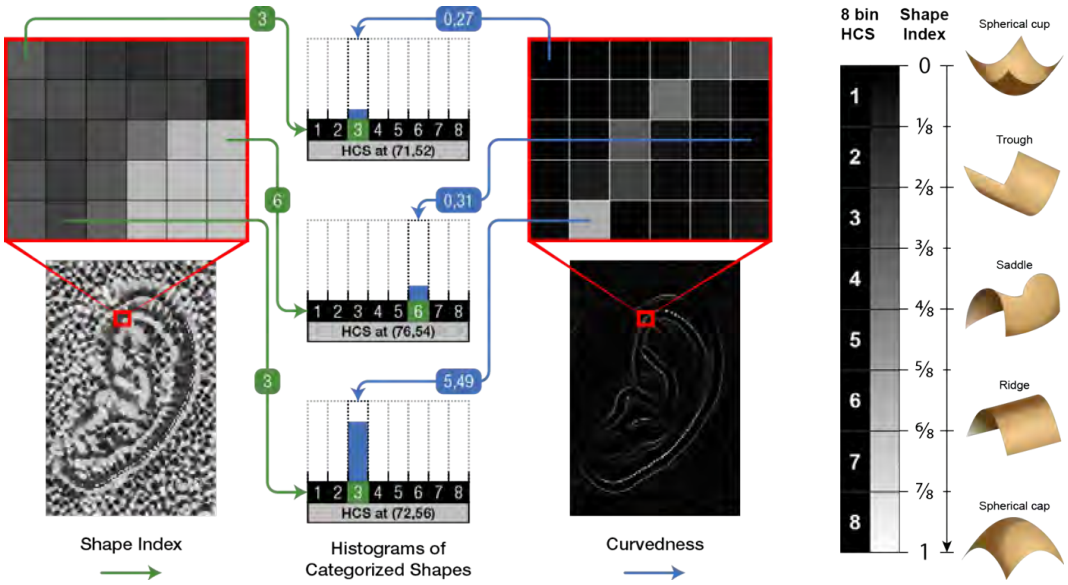


Figure 6.1: Three HCS examples illustrating the transformation of shape index and curvedness values into HCS. The shape index decides the shape category and the curvedness decides the value to increment with. For each block inside the detection window, a separate HCS is calculated. The final descriptor is generated by concatenating all HCSs for all blocks within the detection window.

Pflug et al. [154] propose to project high curvature values of the mean curvature to a 2D binary image. After applying some morphological operations for removing artifacts, they combine these projected structures with each in order to find structures, which are most likely to represent the characteristic shape of an ear. The combined shape with the highest likelihood is then marked as the ear region.

Finally, Zhou et al. [218] proposed a promising technique, called Histograms of Categorized Shapes (HCS), which uses the characteristic distribution of shape index values for ear detection in 3D profile images. Although, it was only tested on a small dataset, Zhou et al. achieved promising results. HCS ear detection is based on a sliding detection window and hence it is not invariant to rotation. We observe that for larger angles the HCS ear detector suffers from an increasing number of false positive detections. This causes many incorrectly marked ear regions (see Figure 6.4). In this work, we provide additional performance measures of the HCS descriptor on the UND-J2 [201] and the UND-NDOff-2007 [64] dataset. Furthermore, we propose to make the HCS detector invariant to rotation by using a circular, rotation invariant detection window.

In the upcoming section, we summarize the approach of Zhou et al. and explain the particularities of our implementation. Subsequently, we introduce the concept of the circular HCS detector and point out, how the original feature vector generation has to be adapted in order to use a circular detection window. In section 6.4 we present the results of our measurements on the detection performance of the circular as well as the rectangular HCS detector. Finally, we conclude with an outlook on future work in Section 6.5

6.2 The HCS Detector

Histograms of Categorized Shapes (HCS) was introduced by Zhou et al. in [218]. This approach for ear detection in 3D profile images exploits the characteristic, richly curved

structure of the outer ear for segmenting it. In experiments on UND Collection F [201], Zhou et al. reported a detection rate of 95.8% at low computational costs.

The first step of HCS is to quantize the shape index s and the curvedness value c into a number of shape categories for describing the distribution of shapes, i.e. characteristics. According to [28], the shape index s and the curvedness c of a surface patch are defined as follows.

$$s = \frac{1}{2} - \frac{1}{\pi} \arctan\left(\frac{k_{max} + k_{min}}{k_{max} - k_{min}}\right) \tag{6.1}$$

$$c = \sqrt{\frac{k_{max}^2 + k_{min}^2}{2}} \tag{6.2}$$

Here, k_{min} is the minimum principal curvature and k_{max} is the maximum principal curvature, respectively. The composition of the previously obtained shape index and curvedness values is used to vote each pixel into a histogram of categorized shapes (HCS) in order to obtain the HCS descriptor. Figure 6.1 demonstrates how three pixel locations are voted into HCSs. Pixels in the shape index image determine the shape category, i.e. a HCS bin, whereas pixels in the curvedness image determine the value to increment with in the determined shape category. We found that a special treatment for pixels with non-defined values (background pixels, which are not part of the model) is necessary. Simply ignoring these pixels and adding empty HCSs caused a large number of false detections. So we decided to introduce a penalty value for each undefined pixel in a block. In our implementation, each undefined pixel is subtracted from all bins in the histogram. Thus, a block which only contains undefined pixels will result in histogram with negative values for each bin.

The final HCS feature vector however is composed of overlapping blocks inside the current detecting window. All HCSs from each of these blocks are concatenated in order to obtain the descriptor for the whole detection window (see Figure 6.1).

The generation of the feature vector can be implemented in an efficient way by using the concept of integral images, which was introduced by Viola and Jones [182]. The usage of the integral image accelerates the summation of pixels, that are located inside a certain area of an image i. For a pixel $p(x, y)$, the integral image I contains the sum of pixels, which lie inside the square with the boundaries of x and y.

$$I(x, y) = \sum_{x'=1}^{x} \sum_{y'=1}^{y} i(x', y') \tag{6.3}$$

Using the integral image approach, the HCSs are generated up for each block inside the current detection window. This means, that the magnitude of each HCS bin inside a block is determined by calculating the integral image for each of the eight bins values.

Because the blocks can be of different size, the summed curvedness values in large block HCSs are significantly larger than the curvedness values in small block HCSs. In order to equalize the weights of HCSs from different blocks, the HCSs are normalized. In order to identify the optimal normalization scheme, we evaluated different normalization methods and were able to reproduce the results of Zhou et al. who found that the L2-Norm yields a slightly better performance than L1-Norm, L1-sqrt and L2-hys. Based on these results, we chose the L2-norm as the default scheme for all experiments using the rectangular detection window. The HCSs feature vector serves as the input for a Support vector Machine (SVM), which is trained to distinguish between HCSs, that were extracted from the ear region and other HCSs. The classifier was trained with the ear images of the first 100 subjects of the UND-2 Database.

The positive training set was generated by manually cropping the ear region from the input images. Subsequently, the cropped window is resized in order to have the same size

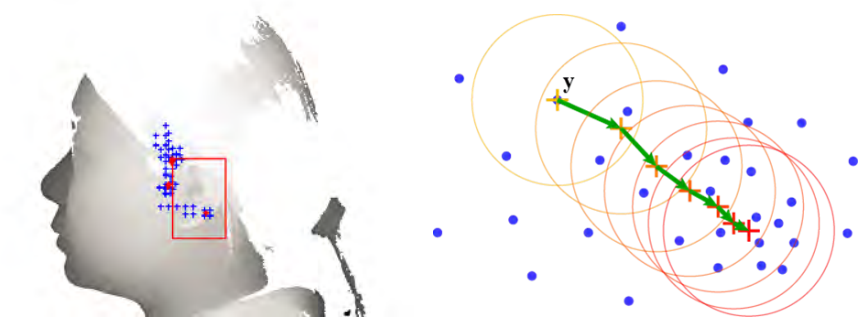


Figure 6.2: The left figure shows an example for ear candidate detections and a detected ear region in a profile image from UND-J2. The crosses represent the ear candidates and the circles represent the center points of the clusters after mean shift. The right figures illustrates how mean shift converges to the central point in a cluster of ear candidates.

as the detection window. Once the training images have been adjusted in size, four additional ear samples are obtained for each of the training images. The purpose of these additional samples is to cover small variations of the ear, such as small rotations and translations. The four additional samples are generated by shifting the cropping window twice by $\pm$ 5 pixels and by rotating the ear by $\pm$ 5 degrees.

For obtaining negative examples, arbitrary parts of the training images were cropped with the restriction, that they have less than 10% overlap with the ear region. For each positive training image, 15 negative examples are generated. Using the previously trained SVM classifier, a sliding window of varying scales is now moved across the image and its content is classified as ear or non-ear. In Figure 6.2 the results of the SVM classifier are marked with a cross. It can be seen, that the ear region is detected several times, which leads to a cluster of detection is this region. Furthermore, there is a number of false positive detections in the collar region. It is, however, assumed, that the ear actual region can be found in the densest cluster of detections. Hence, the last step for detecting the ear is to select the densest region of detection by using mean shift [56].

Mean Shift is an iterative algorithm, which locates the point of the highest density of detection points. It starts at some point y in the image and analyses the neighborhood of y in a predefined radius σ. In each iteration y is shifted to the point inside σ, with the highest local density of detection points. As y keeps moving towards the highest density region of the current cluster, the step size in each shift of y decreases until it falls below a predefined value. Once this is the case, the algorithm stops and the last position of y is the point with the highest density of ear candidates. This point is the top left corner of the detection window containing the ear.

6.3 HCS Using a Circular Detection Window

Because of the usage of variable mean shift, the HCS detector is invariant to scale up to some degree. Invariance to rotation, however, remains an unsolved issue for many naive classifiers. Moreover, the order of the blocks inside the detection window is important, which means that an additional classifier will be necessary for detection left and right areas from images. Any additional classifier, however, doubles the processing time for the detection.

In order to overcome these issues, we propose to replace the rectangular detection window with a circular one. The blocks inside the detection window are ring-shaped, which makes the detector invariant to rotations and gives it the ability to detect left and right ears with the same classifier.

We achieve this by adapting the existing HCS approach as follows. First, we start by determining the shape index and curvedness values of the input image, as defined in equations 6.2 and 6.2. To obtain the feature vector, we first crop a square region from the image and use the pixels of its in-circle, which is transformed into a polar coordinate system. The transformed detection window is now square again and any rotations of the input image will become visible as a horizontal shift. The ring-shaped blocks can now be easily obtained from the rows of the transformed images.

For the regions at the center of the circle, the number of input pixels per row is smaller than the number of input pixels from regions at the circle's boundary. In order to obtain roughly the same number of input pixels per block size, the radii which define the inner and the outer boundary of the ring-shaped block must be adapted. For blocks in the center of the circle, the difference between the outer boundary radius r_0 and the inner boundary radius must be larger than for blocks at the detection window's edges. The area of a block A is

$$A = \pi r_0^2 - \pi(r_0 - b)^2 \tag{6.4}$$

b stands for the block size. The area of each block is always calculated at the outer boundary of the detection window. Hence, for a detection window with a radius of 48 pixels, r_0 would be 548. In our experiments, the values for the block sizes b are set to 2 pixels, 4 pixels and 8 pixels.

The block extraction starts from the outer boundary of the circle and moves towards its center. We define r_{i-1} as the outer boundary of the current block and r_i as its inner boundary. The inner boundary of the previous block automatically becomes the outer boundary of the subsequent one. According to 6.3, the relation between r_i and r_{i-1} is

$$A = \pi r_{i-1}^2 - \pi r_i^2 \tag{6.5}$$

By solving this equation, we obtain the inner radius r_i for the outer radius r_{i-1}.

$$r_i = \sqrt{\frac{\pi r_{i-1}^2 - A}{\pi}} \tag{6.6}$$

It is preferable to introduce some redundancy into the feature vector, in order to get better detection results. This is done by repeating the block extraction for block size b with a different initial radius denoted as r_0'. Based on the initial radius r_0, r_0' is simply

$$r_0' = r_0 - \frac{b}{2} \tag{6.7}$$

Similarly to the original HCS approach, the feature vector is obtained by concatenation all HCSs from all blocks, as illustrated in Figure 6.3. The HCSs were computed by using the same voting scheme as in Figure 6.1. Moreover, we used the L2-norm for normalizing the HCSs from each block before adding them to the feature vector.

The ear detection itself is performed in the same way as for the HCSs using a rectangular detection window. The training set consists of the same number of positive and negative examples, which the difference that we do not generate additional samples by rotating them by ± 5 degrees. The negative examples are, again, obtained by using randomly selected regions from the training images, having less than 10% overlap with the ear region.

During the detection phase, a square detection window is moved across the image. The in-circle of the detection window is cropped and the local feature vector is obtained after the contents of the in-circle are transformed into a polar coordinate system. Subsequently, the feature vector is obtained from the transformed image and the SVN decides, whether this feature vector may be an ear or not. The final ear region is selected by locating the densest cluster of ear candidates from all scales using mean shift.

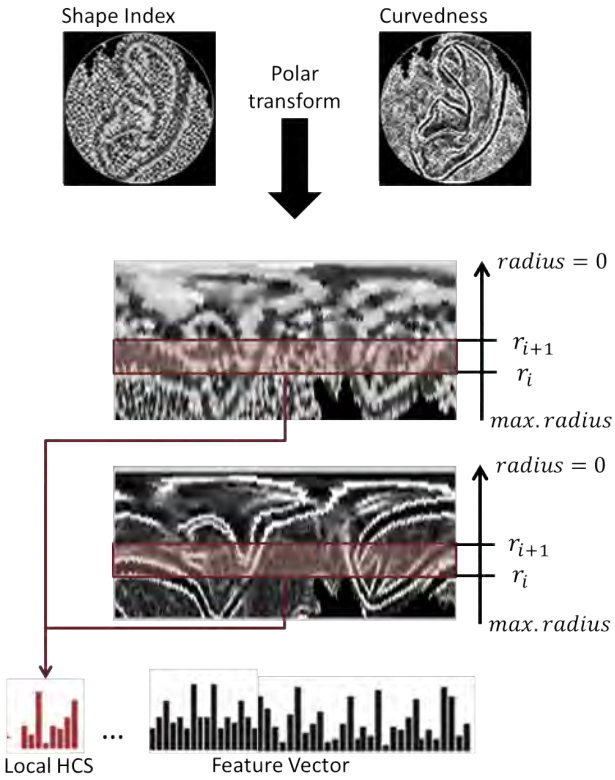


Figure 6.3: Illustration of the feature vector computation using a circular detection window and variable radii.

6.4 Experimental Results

Our experiments cover tests on the HCS descriptor with the rectangular and the circular detection window and are based on two different datasets. We used the UND-J2 collection [201] as well as the UND-NDOff-2007 collection [64] for testing HCS ear detection for robustness against rotation, occlusion and pose variations.

The first series of tests is based on the UND-J2 collection. It consists of 2414 left profile images from 404 subjects. The images do not contain any pose variation and only minor occlusions (mostly because of earrings). Variations in the pitch pose were simulated by rotating the images by various angles. In our experiments, we used a subset of UND-J2, which consists of 1287 images from 253 subjects.

A second experiment is conducted on the basis of UND-NDOff-2007 collection, which consists of 7386 images from 296 subjects. We selected a subset of this database, which covers all images with yaw poses between 30 and 90 degrees and -30 -90 respectively. This subset contains fully and partly occluded ears. Moreover, some subjects are wearing earrings. For each yaw pose, UND-NDOff-2007 contains different pitch poses between -60 and + 45 degrees, where positive pitch values indicate and elevated gaze. UND-NDOff does not contain arbitrary combinations between yaw and pitch poses. The main reason for this is, that the database was initially collected for evaluation face recognition approaches and not for ear detection. Hence it mainly contains frontal views, where a large part of the face is still visible. In Table 6.2 we have marked the missing combinations with "NA". For each available combination between yaw and pitch, the database contains roughly 170 images per subject.

The performance of the detection approaches is measured by comparing the detected

Table 6.1: Detection rates for different rotation angles for UND-J2

Pitch	rectangular	circular
−60°	38.78%	82.14%
−40°	64.13%	79.37%
−30°	93.17%	89.59%
−20°	94.07%	89.9%
−10°	97.28%	93.08%
0°	98.22%	94.17%
+10°	94.43%	93.18%
+20°	75.44%	92.08%
+30°	49.06%	74.98%
+40°	43.07%	79.02%
+60°	26.01%	82.12%

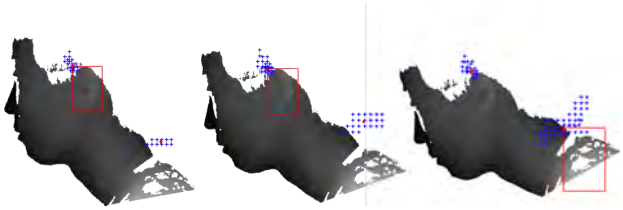


Figure 6.4: Detections of the rectangular HCS detector under rotations of +10, +20 and +30 degrees. It can be noticed that the number of false positive detection increases with the rotation angle.

ear region with a ground truth. The ground truth was generated with a small tool, where humans can draw bounding boxes around the ear region. We consider the detection result as correct if at least 50% of the detected ear region and the ground truth overlap.

6.4.1 Experiments on UND-J2

The goal of the experiments using the UND-J2 dataset is twofold. Firstly, we would like to reproduce the results from [218] in order to prove the correctness of our implementation. Secondly, the behavior of the HCS detector under rotations (which ware variations in the pitch pose) is observed. In order to achieve this, we measured the detection rate of HCS with a rectangular and a circular detection window with different rotations between plus and minus 60 degrees of the image.

The results of the first experiment are summarized in Table 6.1. For non-rotated images, we measured a detection rate of 98.22%, which means that we could almost reproduce the results of Zhou et al. in [218]. The detection rate, however, drops when the rotation angle is increased. It is striking that the HCS detector appears to be more robust against clockwise rotations than anti-clockwise rotations. At the first sight, this behavior seems to be surprising. A closer look at the detection results in Figure 6.4, however, shows the reason.

The detection rate for clockwise and anti-clockwise rotations is different, because the number of false positives is different for each rotation. For anti-clockwise rotations, the number of false positives increases faster than for clockwise rotations. This is a problem, because mean shift places the final detection window in the image region with the densest

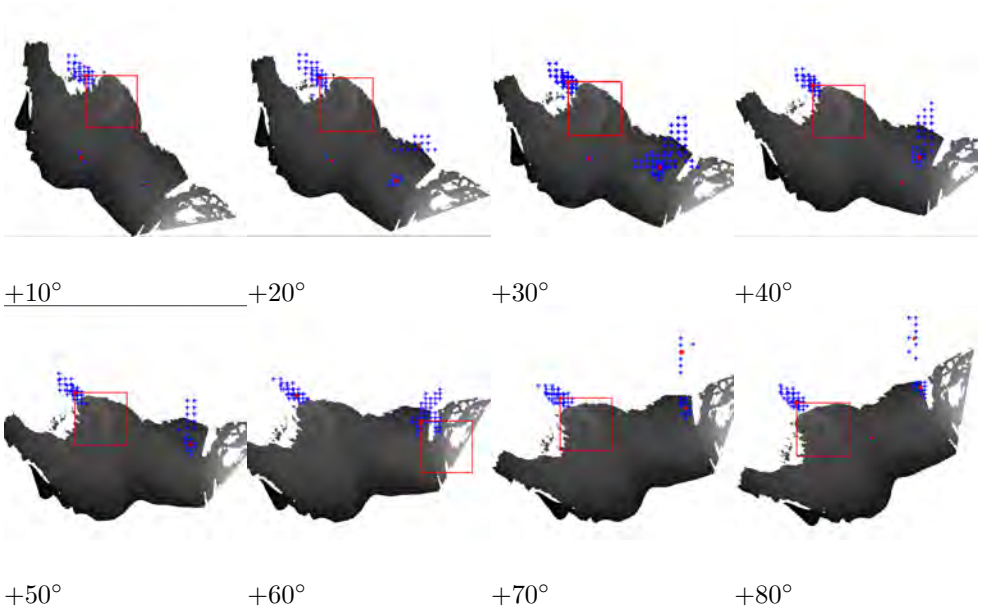


Figure 6.5: Detection results under different rotations using the HCS detector with a circular detection window.

cluster of positive detections. If the number of false positives is too large, mean shift converges towards the wrong image region and hence returns a misplaced detection window.

The overall detection rate of the circular descriptor is lower than the detection rate for the rectangular one. This may be due to the loss of one dimension and hence the shorter feature vector, which roughly has half the length of the feature vector for the rectangular window.

The detection rate still drops a bit with increasing rotations, but it is much less affected than the rectangular window. At ± 60 degrees it is still larger than 82%. Moreover the detection rate drops symmetrically for clockwise and anti-clockwise rotations, which points to the fact, that the circular descriptor has a predictable and well-specified behavior for in-plane rotations. The false detection, which are caused by an increasing number of false positives, however, still remains and could not be entirely solved by using the circular detection window. Figure 6.5 shows that the number of candidates in the ear region remains constant, but similarly to the rectangular window, the number of false positives in other image regions increases with rotation. Compared to the approach using the rectangular detection window (see Figure 6.4), we nevertheless get fewer false positives.

Though the circular descriptor is robust against in plane rotations, it is not rotation invariant. Along with the increasing number of false positive detections, this behavior may be caused by interpolation artifacts. As mentioned before, we simulate variations in the pitch pose by rotating the input image by a given angle using the MATLAB function *imrotate()*. This function uses linear interpolation for generating the final rotated images, which could have affected the detection performance.

6.4.2 Experiments on UND-NDOff-2007

The series of experiments using the UND-NDOff-2007 collection is designed to measure the impact of pose variation on the rectangular and the circular HCS detector. The rectangular detector was tested on data showing only the left ear, whereas the circular detector was tested on left ear and right ear data.

As we expected, the rectangular and the circular detector are affected by pose changes.

Table 6.2: Detection rates of the circular HCS detector for different yaw and pitch poses on UND-NDOff-2007.

Pitch	**Yaw**							
	$+90°$	$+60°$	$+45°$	$+30°$	$-30°$	$-45°$	$-60°$	$-90°$
$-60°$	-	-	-	-	-	-	59.29%	-
$-45°$	-	-	57%	-	43.78%	54.42%	-	-
$-30°$	-	-	-	25.37%	30.72%	52.88%	-	-
$-15°$	-	-	-	-	32.8%	60.58%	-	-
$0°$	91.76%	71.68%	71.82%	42.61%	30.17%	64.02%	70.44%	87.43%
$+15°$	-	-	-	-	31.79%	59.51%	-	-
$+30°$	-	-	-	27.68%	30.72%	63.23%	-	-
$+45°$	-	-	70.7%	-	40.49%	61.62%	-	-

Deviations in the yaw pose and pitch pose affect the both detectors in an equal way. Because both detectors were trained to detect profile ears, they are reaching their maximum detection rate on profile view images. For the circular detector, this can be either the left or the right profile, whereas the rectangular detector only detects ears from left profile images. For large deviations in the yaw pose, the detection rate of both approaches drops in a similar manner.

Though yaw pose variations up to 45 degrees cause a significant drop in the detection rate, both approaches are still able to detect roughly 70% of the ears correctly. For yaw poses of ± 30 degrees, the detection rate is less than 55 %. This effect may be explained by the fact that the auricle protrudes from the side of the head. If the yaw pose moves towards the face, many details of the auricle are still visible. For yaw poses between -30 and + 30 degrees, important details of the auricle are occluded due to its concave shape.

When looking at the effect of pitch poses, the behavior of the rectangular and the circular detector seems to be surprising. For 30 degrees yaw poses, a pitch pose of $\pm$ 45 degrees results in a higher detection rate. In all other cases, increasing deviations in the pitch pose are lowering the detection rate of the algorithm, which is the observation that one would intuitively expect.

The increasing detection rates at poses of 30 degrees yaw and ± 45 degrees pitch can be explained by a combination of self-occlusion effects and the experimental setup during the data collection. Whereas the tragus may occlude parts of the concha or the antihelix at a pose of 30 degrees yaw and 0 degrees pitch, these details become visible again when the pitch pose is changed. Moreover, many subjects are leaning backward when looking at a point above then and forward when looking at a point below them. Hence, the images do not only show the ear from different poses, but also in different scales (see Figure 6.6 for an example). For large variations in the yaw pose, this effect appears to have a positive impact on the detection rates, whereas it has a negative effect for other yaw poses.

6.5 Conclusion

In this work we have introduced a new approach for 3D ear detection, using a circular detection window. Based on the results we obtained from UND-J2 collection, the detection performance of the circular detector could not exceed the performance of [218]. Under in-plane rotation however, the circular detector is more accurate than the original approach using a rectangular detection window. Encouraged by these results, we plan to improve the performance of the circular detector by using a longer feature vector and by taking a closer look at the causes for the slight performance drops under rotation.

$+30°, 0°$ $+30°, +45°$ $+30°, -45°$

Figure 6.6: Examples of ear images with yaw poses of 30 degrees and pitch poses of 0, +45 and -45 degrees. Note that the images show the ear in different scales and that some parts of the ear are occluded, depending on the pose.

A final remark should be made concerning the computational effort of the circular detector, which is considerably higher than the rectangular one. The reason for this is, that the integral image has to be computed separately for each detection window. In order to make the detector suitable for practical applications, the computation of the local feature vectors should be significantly accelerated.

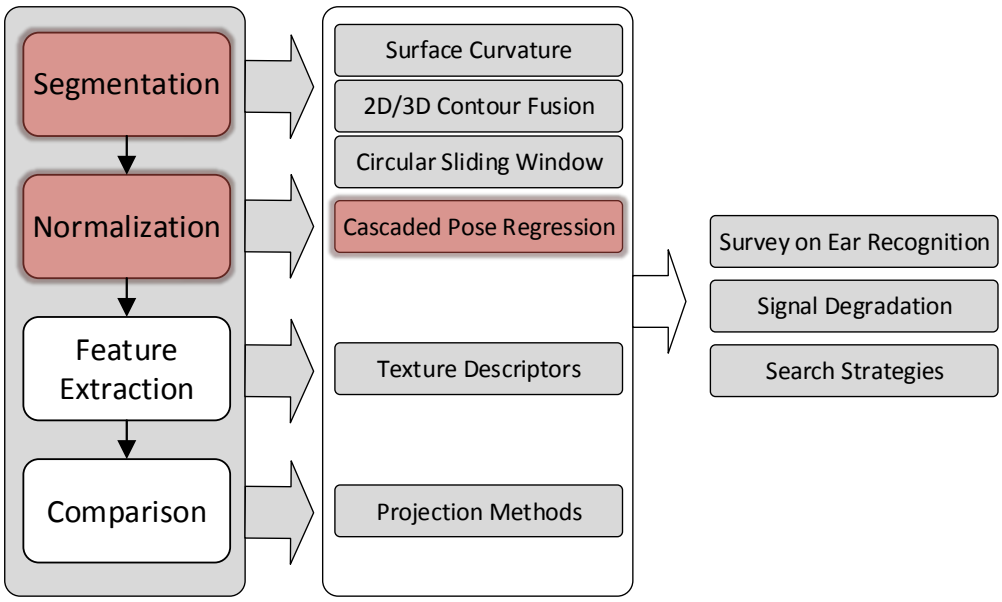
Chapter 7

Segmentation and Normalization of Human Ears using Cascaded Pose Regression

When using texture description methods for feature extraction, accurate cropping and normalization is an important step for ensuring a good system performance. In face recognition, the symmetry of the face is used to achieve an accurate normalization of the images. For ear recognition, however, we cannot rely on such a property. **Q2: How can cropped ear images be normalized with respect to rotation and scale?**

We applied Cascaded Pose Regression for segmenting and normalizing ears from face profile images. We show, that systems that were using the normalization generally performed better than system without normalization. From this we conclude that the proposed method can be used for normalizing in-plane rotation and scale of the ear region in face profile images and in pre-cropped ROIs.

The paper was published in [148]ANIKA PFLUG, CHRISTOPH BUSCH, Segmentation and Normalization of Human Ears using Cascaded Pose Regression, Nordic Conference on Secure IT Systems (NordSec), 2014



Abstract

Being an emerging biometric characteristic, automated ear recognition is making its way into forensic image analysis for law enforcement in the last decades. One of the most important challenges for this application is to deal with loosely constrained acquisition scenarios and large databases of reference samples. The research community has come up with a variety of feature extraction methods that are capable of handling occlusions and blur. However, these methods require the images to be geometrically normalized, which is mostly done manually at the moment.

In this work, we propose a segmentation and normalization method for ear images that is using cascaded pose regression (CPR). We show that CPR returns accurate rotation and scale estimates, even for full profile images, where the ear has not been segmented yet. We show that the segmentation accuracy of CPR outperforms state of the art detection methods and that CPR improves the recognition rate of an ear recognition system that uses state of the art appearance features.

7.1 Introduction

In the last decade, ear recognition has evolved towards a promising new biometric trait that can help to extract more information from half profile and profile images. Building on the work of Iannarelli [81], the shape of the outer ear is widely regarded as unique and persistent. The research community has evaluated a large number of different types of features. Traditionally, the most important application of ear recognition systems is forensic images analysis on images from surveillance cameras, where subjects are usually unaware of the image capture process. Based on ear features, several cases could be cleared, such as a series of robberies at gas stations [79].

Another anticipated application of ear recognition is automatic border control. Both scenarios, law enforcement and airport security, require a stable and reliable normalization algorithm that is capable of rotating and scale the ear image to a well-defined reference orientation. Having normalized images allows us to use all kinds of appearance features, such as LPQ and HOG for describing the ear structure. The normalization should also be tolerant to smaller pose variations and partial occlusions.

When working with CCTV footage, the ear is often small and blurred which makes landmark detection a complicated problem. Some landmarks, and sometimes even the whole ear region may be occluded when a subjects is not cooperative. We also know these problems Landmark-based approaches for face recognition in the wild. For this particular problem, feature selection methods using random forests are successfully applied. In [203] landmark candidates are selected from an image and then selected during a voting process. A major challenge in ear normalization is, that accurate landmark localization is a difficult problem as such. The extraction of landmarks is prone to errors, because many approaches, such as ASM for example, depend on proper initialization and requires a large training set. Unlike the face the ear does not have a symmetry axis to support landmark positions.

In [205], Yazdanpanah and Faez normalize previously cropped ear images by using edge information. A similar approach is also proposed by Wang [188], where the outline of the outer helix is extracted from the convex hull of a binary image that describes the outer helix and other elevated parts of the ear. In both approaches, the authors are using the axis that connect the two points with the largest distance and rotate the image, such that this axis is vertical. However, these approaches require that the ear has been located exactly and that all edges in the image actually belong to the ear. Gonzalez et al. propose to use an active shape model for locating the outer helix and then also use two farthest points on this line for normalization [70]. Both approaches require that there are no occlusions in the image and that the outer ear has already been located by a previous detection step. For a more in-depth discussion and comparison of recent advances in ear biometrics, please refer to the survey of Pflug and Busch [147].

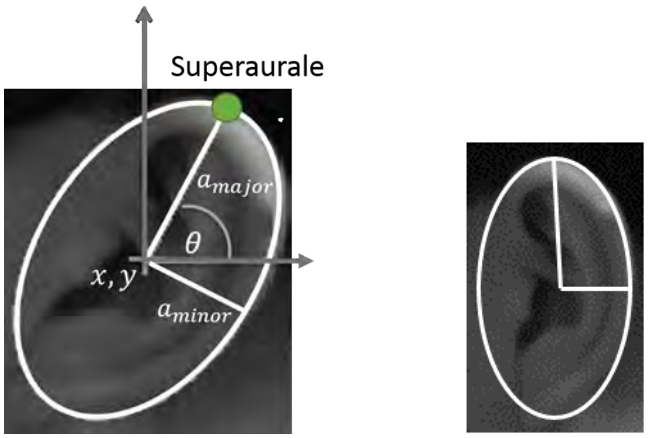


Figure 7.1: Illustration of the CPR-based geometrical normalization of ear images. We fit an ellipse that encloses the ear, and rotate the whole image such that the major axis of the ellipse is vertical.

We propose to estimate the orientation of the ear by using cascaded pose regression (CPR) [62]. This algorithm uses an abstract elliptical model for pose estimation, which fits well to the shape of the outer ear. CPR locates and normalizes the outer ear in profile images, such that a texture descriptor will have a stronger performance. CPR segments the ear region and rotates the cropped image in a way that the superaurale is oriented vertically (see Fig. 7.1). Working with normalized images enables a higher recognition performance with texture-based features and a more accurate landmark detection with Active Shape Models and Active Appearance Models.

The main contributions of this paper are (1) segmentation of the outer ear in pre-cropped full profile images and (2) rotation normalization of the ear region. The normalization experiments are conducted on three different publicly available datasets.

The remainder of this paper is organized as follows: in Section 7.2 Cascaded Pose Regression (CPR) is briefly outlined. In the subsequent section we first report on the detection performance for full profile images and compare the performance with existing state of the art approaches for object detection. The experimental results on the performance gain from normalizing the ear image with CPR are presented in Section 7.4. The paper is concluded with final remarks and future work in Section 7.5.

7.2 Cascaded Pose Regression

Cascaded Pose Regression (CPR) was proposed recently by Dollar et al. [62]. This algorithm is capable of estimating the pose of a roughly located object in an image by using a cascade of weak fern regressors with local features. In their work, Dollar et al. show that CPR converges fast and performs well with small training sets. They also show that CPR can be applied in different contexts, such as pose estimation of mice, fish and facial landmark detection [36]. What makes CPR particularly interesting for normalizing ear images is the fact that it is not relying on any symmetry constraint or other kinds of higher semantic information. It is solely using local brightness information for estimating a transformation matrix between a given state of the descriptor and the trained model. We will only briefly outline the main ideas of CPR in this selection. For more details, please refer to the original paper by Dollar et al. [62].

In the beginning of the estimation process, the model is initialized at a random position in the image. In our case, the model is described by the parameters of an ellipse, with

center coordinates x, y (see Fig. 7.1). With respect to this coordinate system, we pick a fixed number of pose indexed control point features [141]. A control point feature $h(p_1, p_2)$ is defined by the difference between two gray scale values at two locations $I(p_1)$ and $I(p_2)$ in the image I.

For each pose Φ, we can define an associated $3x3$ projection matrix H_Φ, that express p in homogeneous coordinates. Based on this, we can define

$$h_{p_1,p_2}(\Phi, I) = I(H_\Phi p_1) - I(H_\Phi p_2) \tag{7.1}$$

The algorithm iteratively optimizes the projection matrix, such that the differences between a pair of two pose indexed features is minimized.

During model training, the algorithm learns to predict the orientation of a given object by minimizing a loss function L between the current orientation Φ_i and a defined ground truth orientation Φ_i^T based on the pose indexed features. The loss function models the sum of differences of all pose indexed features between the current pose and the previous iteration. Let $d()$ be the function that computes the difference between two sets of pose indexed features. Then the loss function L can be written as

$$L = \sum_{i=1}^{N} d(\Phi_i^T, \Phi_i) \tag{7.2}$$

In this equation, i refers the i'th stage in the cascade. The training in each stage is repeated until the error drops below a target value $\epsilon \geq 1$, which reflects that the error in the previous iteration was either smaller than or as large as in the current iteration.

$$\epsilon = \sum_{i=1}^{N} d(\Phi_i^t, \Phi_i) / \sum_{i=1}^{N} d(\Phi_i^{t-1}, \Phi_i) \tag{7.3}$$

A stage in the cascade consists of a fern regressor [141] that is taking a randomly chosen subset of features and then samples random thresholds. The fern regressor is created by randomly choosing a given number of elements S from the vector of pose indexed features and them samples S thresholds randomly. We chose the best fern regressors in terms of training error from the pool of R randomly generated ferns. It may happen that CPR gets stuck in a local minimum and hence fails to estimate the correct orientation. To prevent this, the regressor in initialized K times and the solutions are clustered. CPR then returns the orientation with the highest density.

For normalizing ear images, we use a single part elliptical model that roughly encloses the outline of the ear. By doing this, CPR can make use of the rich and distinctive texture information inside the ear region. The elliptical model is defined by a set of pose indexed features that are used for optimizing the position of the model in the image as briefly described above.

A particular pose of the ear is defined by a given set of ellipse parameters x, y, a_{major}, a_{minor} and Θ, where x, y are the coordinates of the ellipse center, a_{major}, a_{minor} are the lengths of the major and minor axis of the ellipse and θ denotes the skew. When fitted to the ear, the major axis of the ellipse is cutting the lobule and pointing towards the superaurale and the minor axis of the ellipse encloses the tragus region. Fig. 7.1 shows an example of how the ellipse is placed.

Following the recommendations in [62], we choose the parameters of our trained model as follows: the number of cascades is set to $T = 512$, the number of ferns is $F = 64$, and the number of pose-indexed features is chosen to be $R = 512$. The regressor is initialized for $K = 50$ times.

Based on the ellipse parameters of a fitted model, we can segment and normalize the ear region. This is done by rotating the image about the center point of the ellipse in a way

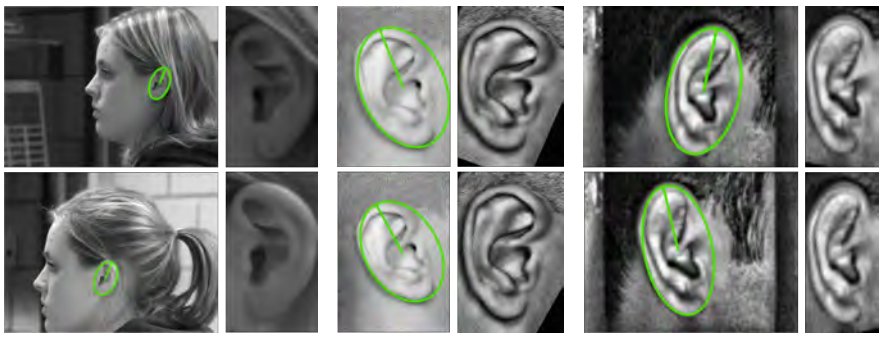


Figure 7.2: Examples for fitted ellipses in original database images from UND-J2 (a), AMI (b) and IIT-Kanpur (c) along with cropped and normalized ROIs.

that the major axis of the fitted ellipse is horizontal. The rotation angle ρ can be directly inferred from the orientation of the ellipse, denoted as θ. This is also illustrated in Fig. 7.1.

$$\rho = 90 - \theta \tag{7.4}$$

After rotating the ear into the defined pose, the ear region is cropped by using the enclosing rectangle of the fitted ellipse. When the major axis of the ellipse is rotated to be vertical, differences in scale can be removed by resizing the ROI in a way that the major axis of all ellipses have the same length. The width of the ROI is adjusted accordingly such that the aspect ratio is preserved.

In Fig. 7.2, we see pairs of example images with the fitted ellipse in the original image and the cropped region of interest (ROI). There are two pairs from the same subjects and from each database.

7.3 Detection from Profile Images

In this experiment, we determined the segmentation performance of our CPR model on full profile images. Using the UND-J2 [201] database, we compared the segmentation accuracy of cascaded pose regression with HOG features, Haar like features and LBP. We used the Implementations of OpenCV 2.4.3. Each of the four detectors was trained with 400 images of the UND-J2 database. The remaining 1377 images were used for evaluation. We are aware, that a selection of 400 images is a rather small training set for Haar-like features, LBP and HOG detectors. This is done on purpose, to highlight the fact that CPR gives exact estimations of the orientation of the outer ear, even with a small training set.

The detection accuracy was determined on the base of a manually annotated ground truth. For calculating the detection accuracy, we compare two image masks with each other and compute the overlap O of these two regions. We consider a detection as being successful if the overlap between the ground truth and the detected region is larger than 50%. This constraint is set in accordance to related work on ear detection, such as in [146], in order to be comparable. Both of these works are using left profile images from the UND collections as well. Let the ground truth region be denoted as G and the detection region be denoted as R. Then O can be written as

$$O = \frac{2|G \cap R|}{|G| + |R|} \tag{7.5}$$

The results of this experiment are summarized in table 7.1. CPR segmentation clearly outperforms HOG and Haar-like features. The detection accuracy of LBP and CPR are similar, however, the detection accuracy of CPR is still slightly better than LBP.

Table 7.1: Comparison of detection rates of different features on images from UND-J2

Algorithm	Detection Accuracy
HOG	77.05%
LBP	99.05%
HAAR	98.32%
CPR	99.63%

In contrast to cascaded detectors, CPR does not only provide information about the location of the ear, but at the same time, gives an estimate of the axis between the tragus and the superaurale for normalization. Moreover, the segmented ear regions from CPR are more accurate than the ROIs of cascaded detectors. This means that the segmented region contains only a small portion of the ear's surroundings.

7.4 Normalization

7.4.1 Experimental Setup

In the second series of experiments we evaluate the impact of CPR normalization on the recognition performance of local texture descriptors. All of these descriptors are vulnerable to pose and scale variation [1]. We hence expect an increase in recognition performance for each of these methods, when the image is normalized prior to feature extraction.

Further, we expect the performance improvement to be independent of the texture descriptor and the projection technique. In other words, we expect the EER and the rank-1 recognition rate in a configuration where CPR normalization was used to be better than in the same configuration in a reference system, where the image was not normalized. The ear recognition system for obtaining the performance indicators is outlined in Fig. 7.3. We apply 10-fold cross validation for obtaining the performance indicators in each of the configuration.

Configurations with normalized ears use the cropped and normalized ROI, that is returned from CPR. The reference system, is using manually cropped ear regions for UND-J2. For AMI and IITK, we used the complete images for computing the reference performance (see Fig. 7.2. The performance indicators for normalized images were operating on the output ROIs from CPR (see Fig. 7.2).

The recognition experiments were conducted on UND-J2, AMI and IIT-Kanpur databases with the following partitioning:

UND-J2: The UND-J2 database is an image collection with 2D and 3D left full profile images, however, only the 2D images are used in this evaluation. The ground truth image of UND-J2 have been cropped manually. Samples, where the ear detection using CPR failed, have been removed from the evaluation set in order to keep the segmentation error and the recognition error separated. We used all subjects with at least 6 samples that have not been used for CPR training. Five of these samples were used for training, and one sample was used for testing. In consequence we were using 583 images for training and 117 images for testing.

AMI: The AMI database is a collection of close-up, high resolution ear images. It consists of 8 samples per subject in total, whereas the first sample shows the right ear and all other samples show the left ear of each subject. For left ears, each sample was captured from a slightly different pose and distance. The first 30 subjects have been used for training

[1] There exist extensions for LBP that make it invariant to scale. These extensions are not covered in this work, because we would not be able to illustrate the impact of the normalization with scale-invariant features.

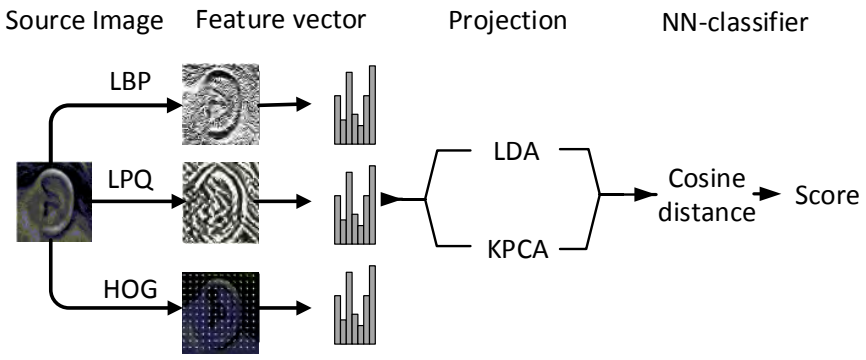


Figure 7.3: Experimental setup showing all appearance features and projection schemes.

CPR. For the remaining 67 subjects, we used sessions 2 until 6 for training and session 7 for testing. This implies that we have 335 training images and 67 testing images.

IIT-Kanpur: The IIT-Kanpur database contains ear images that have been collected on different days with a fixed distance between the subject and the camera. There are slight variations in the pitch pose. The first 50 subjects have been used for training CPR. For the remaining subjects, we used sessions 1 until 3 for training and all remaining images for testing. Hence, the training set contained 225 images and the testing set was composed of 81 images from 74 different subjects.

7.4.2 Texture Descriptors

For texture description, we apply three different state of the art texture descriptors, which are Unified Local Binary Patterns (uLBP), Local Phase Quantizaton (LPQ) and Histograms of Oriented Gradients (HOG).

Unified Local Binary Patterns (uLBP): uLBP [138] encodes local texture information on a pixel level by comparing the grey level values of a pixel to the grey level values in its neighborhood. The size of neighborhood is defined by a radius around the pixel g_j, which is at least 1 (for a neighborhood having 8 pixels). Every pixel q_i within the radius that has a larger grey level value than the center pixel is assigned the binary value 1, whereas all pixels with a smaller grey level value are assigned the binary value 0.

The binary values in the neighborhood pixels are concatenated to form a binary string corresponding to the center pixel. Only those binary strings which have at most two bit-wise transitions from 0 to 1 (or vice-versa) are considered - there are 58 such strings. This binary string is then mapped to a value between 0 and 58.

The uLBP-based ear descriptor is computed by first sliding a window of a predefined size and overlap (step size in pixels) in the horizontal and vertical direction over the LBP image. From each sub window a local histogram with 59 bins is extracted (the first 58 bins correspond to the uniform binary strings, and the 59-th bin corresponds to the rest).

The final descriptor is the concatenation of each local histogram. For a window size of 20×20 pixels and an overlap of 10 pixels, this results in a feature vector of dimension 3776.

Local Phase Quantization (LPQ): The concept behind LPQ [10] is to transform the image into the Fourier domain and to only use the phase information in the subsequent steps. Given that a blurred image can be viewed as a convolution of the image and a centrally symmetric point spread function, the phase of a transformed image becomes invariant to blur. For each pixel in the image, we compute the phase within a predefined local radius and quantize the image by observing the sign of both, the real and the imaginary part of

the local phase. Similar to uLBP, the quantized neighborhood of each pixel is reported as an eight digit binary string.

Given an image, the LPQ value is first computed for every pixel. Next, local histograms with 265 bins are computed within a sliding window. We move this sliding window, with a given overlap between two neighboring windows, in horizontal and vertical direction over the image and concatenate each local histogram. With a 20×20 window size and an overlap of 10 pixels, this results in a 16.384 dimensional feature vector.

Histogram of Oriented Gradients (HOG): Computation of the HOG [57] descriptor involves five steps, which are the gradient computation, orientation binning, histogram computation, histogram normalization and concatenation of local histograms. The algorithm starts by computing the local gradient by convolving a 3×3 region (HOG cells) with two one-dimensional filters (-101) and $(-101)^T$. The local orientation at the center of each HOG cell is the weighted sum of the filter responses of each pixel.

The local orientations within a larger sub-window, denoted as block, are then quantized into bins in the $[0, 2\pi]$ interval. Subsequently, the image is divided into blocks of equal size and a local histogram of quantized orientations is extracted. Subsequently, the local histogram from each block is normalized with the L2-norm. Finally, all local histograms are concatenated to form the HOG descriptor for the image. The HOG descriptor with block size of 8×8 pixels and 9 orientation bins has 5184 dimensions.

7.4.3 Feature Subspace Creation

The feature subspace for each of the descriptors is created with one of four different projection techniques. We apply the widely used LDA as representative for linear projection method. Additionally, we use KPCA [172] to have a non-linear technique. Finally, the most likely identity for each probe image is the reference image that has the smallest cosine distance to the projected probe image in the feature subspace. Parts of the source code for this experiment are based on the PhD Face recognition Toolbox.

7.4.4 Results

In Table 7.2 the rank-1 recognition rates and the EERs are summarized for all possible combinations between texture descriptors and the two dimensionality reduction techniques. As the numbers show, CPR normalization improves the performance all pipeline configurations using LDA and with each of the databases. Compared to the reference configuration, the EER is improved up to 3% in each of the databases. The improvement is of course dependent on the degree of pose variations in a dataset. Because the poses in UND only vary slightly, we have larger pose variations in IITK and the AMI datasets. The ear images in IITK vary in rotation and the images in AMI contain rotation and scale variations (see Fig. 7.2). Consequently, there is more potential for CPR to correct variations of scale and rotation, as well as to accurately segment the ROI. Examples for successfully corrected variations, are shown in Fig. 7.2.

When using KPCA for creating the feature subspace, we obtain high error rates However, the error rates as well as the standard deviation between the error rates from different cross-validation attempts are high for these configurations. Based on our observation, we conclude that KPCA is not a suited for the texture descriptors in our experiment. The recognition performance of all configurations using LDA yields EERs below 3,5% in all databases. On average, the lowest error rates were achieved with HOG. However, LPQ and uLBP perform similarly on all datasets.

Another factor that may also have influenced the performance of images, that have been normalized with CPR is, that CPR is capable of cropping the region of interest more accurately than other segmentation techniques, such as those used in the previous section. Hence, the feature vector contains less information for the surroundings of the ear compared to the ground truth in AMI and IITK. The region around the ear contains some in-

Table 7.2: Detection rates with normalization (CPR) and without normalization (GT) for selected recognition pipelines. EER and rank-1 recognition rate are given as percentages and are represented as tuples of the form *EER // Rank-1*.

	LDA		**KPCA**	
	CPR	GT	CPR	GT
UND-J2				
uLBP	2.58 // 93.35	3.43 // 90.94	23.36 // 38.86	29.93 // 22.28
LPQ	3.23 // 91.51	4.50 // 82.28	17.02 // 55.44	22.22 // 36.08
HOG	1.76 // 95.70	3.05 // 91.14	26.87 // 19.43	32.66 // 5.44
AMI				
uLBP	1.85 // 93.4	4.32 // 86.1	39.02 // 19.30	28.40 // 26.1
LPQ	0.40 // 97.2	5.19 // 85.2	40.53 // 16.0	25.98 // 42.1
HOG	0.68 // 98.10	5.21 // 82.2	22.99 // 43.1	28.44 // 19.9
IITK				
uLBP	0.26 // 99.72	1.67 // 97.22	11.84 // 72.64	16.88 // 70.41
LPQ	0.02 // 99.72	3.28 // 94.03	8.01 // 85.28	18.76 // 61.1
HOG	0.18 // 99.72	1.48 // 95.83	6.01 // 86.25	23.22 // 35.27

formation, which helps to increase the performance in datasets, which have been collected in a short period of time. However, we expect that these features are subject to changes in hairstyle, clothes and jewelry and hence are rather a soft biometric, than a permanent features. For real life applications, we assume that a more accurate segmentation will result in better recognition performance.

7.5 Conclusion

In this work, we have shown, that an ear image can be normalized without the need for symmetry or landmark extraction. We have provided experimental evidence showing that cascaded pose regression improves the performance in two ways. With CPR it is possible to detect and normalize the outer ear from profile images with only one processing step. The ear region can be segmented accurately to make sure that the feature vector only contains information about the ear and not about its surroundings. In addition to this, CPR also extracts information to compensate differences in rotation and scale, while being robust to minor occlusions.

We have shown that the performance improvement, which can be achieved with CPR is independent of the capture scenario by using different datasets. The performance improvement is also independent of the texture descriptor. Motivated by our results, we plan to train multi-part CPR models for landmark detection in the future.

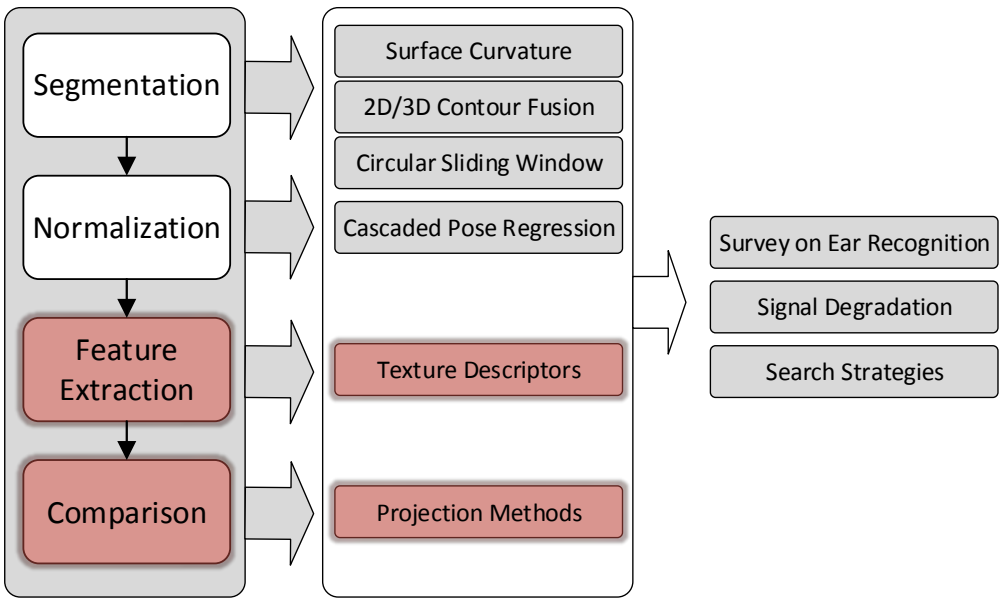
Chapter 8

A Comparative Study on Texture and Surface Descriptors for Ear Biometrics

Histogram-based texture descriptors have the advantage, that they produce a feature vector of fixed length at a high biometric performance, but are restricted to 2D images only. In this chapter we address research question **Q3: Is it possible to combine 2D and 3D data in order to obtain a better descriptor that yields a better performance than 2D or 3D alone?**

We first conduct a comparative study of different texture descriptors in 2D images and determine the optimal parameter sets for each of the datasets in our experiment. The same is also done for depth images, using the Shape Index and the Curvedness index as the input for the texture descriptors. In addition, we also propose to combine information from texture and depth images in a single histogram descriptor. We conclude that the combination of 2D and 3D information as proposed in this work, does not yield a better performance than the 2D information alone. The loss in performance is mainly due to the noisy surface data in the 3D image.

The paper was published in [150]ANIKA PFLUG, PASCAL N. PAUL AND CHRISTOPH BUSCH, A comparative Study on Texture and Surface Descriptors for Ear Biometrics, International Carnahan Conference in Security Technology (ICCST), 2014



Abstract

Recent research in texture-based ear recognition also indicates that ear detection and texture-based ear recognition are robust against signal degradation and encoding artefacts. Based on these findings, we further investigate and compare the performance of texture descriptors for ear recognition and seek to explore possibilities to complement texture descriptors with depth information. On the basis of ear images from visible light and depth maps, we extract texture and surface descriptors. We compare the recognition performance of selected methods for describing texture and surface structure, which are Local Binary Patterns, Local Phase Quantization, Histograms of Oriented Gradients, Binarized Statistical Image Features, Shape Context and Curvedness.

Secondly we propose a novel histogram-based descriptor that performs feature level fusion by combining two information channels to form a new feature vector. Our concept can either be applied for fusing two different texture or two different surface descriptors or to combine texture and depth information. Based on the results of the previous experiment, we select the best performing configuration settings for texture and surface representation and use them as an input for our fused feature vectors. We report the performance of different variations of the fused descriptor and compare the behavior of the fused feature vectors with single channel from the first series of experiments.

8.1 Introduction

As a consequence of increasing availability of high resolution cameras, the outer ear with is unique shape has moved into the focus of many forensic experts. Simultaneously, cameras that capture video and depth information have reached a state that make them applicable in semi-controlled scenarios, such as identification at ATMs, vending machines or border control. Recent research in texture-based ear recognition indicates that ear detection and texture-based ear recognition are robust against signal degradation and encoding artefacts, which implies that we can achieve a good recognition performance, even from a distance of several meters from the camera. Based on these findings, we investigate and compare the performance of texture descriptors for ear recognition and seek to explore possibilities to complement texture descriptors with depth information.

First proposed on 1964 [81], ear recognition has gained an increasing amount of attention from the biometrics and forensics community during the last years. Based on the available data, the outer ear is widely regarded as a unique, permanent and easy to capture characteristic. The rich surface structure is especially valued for forensic image analysis, where the outer ear is more frequently visible than the face. Video footage can even provide material showing both ears and the face, which pushes the interest in automated ear recognition of criminal police agencies all other the world.

Due to the increasing attentions for ear recognition, researches have shown that different techniques which have originally been proposed for face recognition can also be applied for ear recognition purposes. Among these features are approaches for landmark extraction, such as Active Shape Models [119] and SIFT [102] [220]. Key point detection and matching was also successfully applied to 3D images [211] and in combination with a global image descriptor as in [219]. Landmark and key point detection however, can be complex, time consuming and incorporates implicit assumptions about the capture scenario and object representation. Moreover the process of landmark or key point extraction adds a potential source for errors to the pipeline that stacks with the probability of a wrong segmentation. In other words, even if the ear has been successfully detected, we can still get wrong landmarks or misplaced key points from the feature extraction stage. We hence focus on appearance features in this work.

In the field of appearance feature extraction from 2D images, Damer and Fuhrer obtained good results with HOG [58] and in [72] the authors use LBP in combination with different techniques for feature subspace projection. In a survey of ear recognition algo-

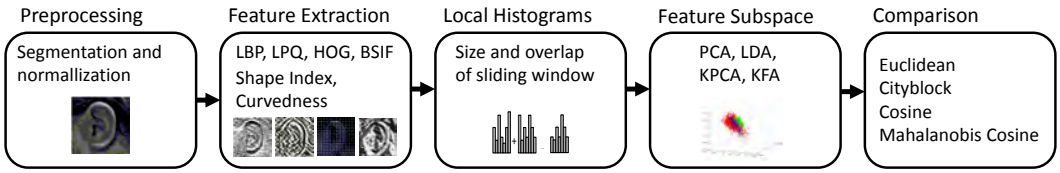


Figure 8.1: Illustration of the experimental setup for our evaluations with all possible combinations between feature extraction, local histogram concatenation, feature subspace creation and comparison.

rithms, their performance and the databases the performance metrics were obtained with can be found in [147].

The aforementioned selection of different approaches to ear recognition have been obtained from different datasets and hence are hard to compare with each other. The goal of this work is to compare different texture and surface description techniques with ear other and to give recommendations for the optimal settings under a given scenario, which is represented by a particular database. Based on the results on LBP and HOG in previous work, LPQ and BSIF are likely to give good results for ear recognition as well. However, LPQ and BSIF have not been tested and compared to previous approaches for appearance based ear recognition before. Moreover, we compare the recognition performance of the texture image and the depth image in order to see, which representation of the ear contains the most distinctive information. We also compare different projection methods for feature subspace creation. Our experiments are conducted on three different datasets, which are UND-J2 [201], AMI[70] and IITK [107]. Example images from each of the database can be found in Figure 8.3.

Our second contribution is a combined histogram descriptor, which combines 2D and 3D data. The combined histogram descriptor is based on the work of Zhou et al. [218] and [57], where two sources of data are combined to assign a given magnitude from one source to a bin that is determined by another source. The evaluation results for this descriptor are based on the UND-J2 dataset.

8.2 System Overview

As a first step, all images are transformed to gray scale and the contrast is adjusted by using CLAHE [223]. Subsequently, variations in rotation and scale are removed with cascaded pose regression (CPR)[62]. With CPR, we estimate an ellipse that encloses the ear. Based on the length of the major and minor axis and the skew of the ellipse, we rotate and scale the ear image in a way that the major axis of the ellipse is vertical and as a predefined length. This is done for the 2D images and for the depth images, such that the normalized images are still co-registered. Missing values in the 3D-Images are interpolated with a linear function for computing potential energy surfaces and a least squares approach [109]. Finally, all images are resized to 100×100 pixels. Examples images for each database after preprocessing are displayed in Fig. 8.3.

After preprocessing, we extract appearance or surface descriptors from overlapping grid cells and create fixed length histogram descriptors. The histogram descriptors are either computed from 2D images, or depth images or from both. Depending on the number of grid cells, the number of dimensions for each feature vector varies. For LBP, LPQ and BSIF we used local windows of the sizes 33×33, 20×20, 10×10 and 7×7 with different overlaps that are dependent on the window size.

The dataset is split into a training and a testing set. The training set is used to estimate a projection matrix for a feature subspace. Using the projection matrix, the remaining testing

images are projected into the feature subspace. We compare the recognition performance of feature subspaces that were created with PCA, LDA as linear projection techniques and KPCA [172] and KFA [127] as non-linear projection techniques [1] Recognition is performed with a nearest neighbor classifier and a selection of different distance metrics which are the Euclidean distance (euc), city block distance (ctb), cosine distance (cos) and mahalanobis cosine distance (mahcos). Parts of the source code for this experiment are based on the PhD Face recognition Toolbox. The whole data processing process, including all intermediate steps is summarized in Figure 8.1.

We compute performance metrics for different combinations between the feature extraction techniques, the size of the local windows, the overlap between local windows and the projection technique and the distance metric. In total we obtain more than 6000 different configurations, which were compared for this study. Each of the possible combinations is tested with 10-fold cross validation with a random selection for the training and testing set for the computation of the feature subspace.

8.3 Feature Extraction

8.3.1 Texture and Surface descriptors

Local Binary Pattern (LBP): LBP [138] encodes local texture information on a pixel level by comparing the grey level values of a pixel to the grey level values in its neighborhood. The size of neighborhood is defined by a radius around the pixel, which is at least 1 (for a neighborhood having 8 pixels). Every pixel within the radius that has a larger grey level value than the center pixel is assigned the binary value 1, whereas all pixels with a smaller grey level value are assigned the binary value 0. The binary values in the neighborhood pixels are concatenated to form a binary string corresponding to the center pixel. Only those binary strings which have at most two bit-wise transitions from 0 to 1 (or vice-versa) are considered - there are 58 such strings. This binary string is then mapped to a value between 0 and 255.

The LBP-based ear descriptor is computed by first sliding a window of a predefined size and overlap (step size in pixels) in the horizontal and vertical direction over the LBP image. From each sub window a local histogram with 256 bins is extracted. We compare the performance values of this descriptor with using radius of 1 (n-8 neighborhood).

Local Phase Quantization (LPQ): The concept behind LPQ [10] is to transform the image into the Fourier domain and to only use the phase information in the subsequent steps. Given that a blurred image can be viewed as a convolution of the image and a centrally symmetric point spread function, the phase of a transformed image becomes invariant to blur. For each pixel in the image, we compute the phase within a predefined local radius and quantize the image by observing the sign of both, the real and the imaginary part of the local phase. Similar to uLBP, the quantized neighborhood of each pixel is reported as an eight digit binary string.

Given an image, the LPQ value is first computed for every pixel. Next, local histograms with 265 bins are computed within a sliding window. We compute the concatenated histogram descriptor for varying window sizes and with different radii for the neighborhood of each pixel. We compare LPQ with radii 3, 5 and 11 in combination with different window sizes and overlap.

Binarized Statistical Images Features (BSIF):Inspired by LBP and LPQ, BSIF [98] also computes a binary string for each pixel in an image to represent the local structure of an image. The value of each bit within the BSIF descriptor is computed by quantizing the response of a linear filer. Each bit in the string is associated to a particular filter and the number of bits determines the number of filters used. As in LBP and LBP the binary code

[1] We also evaluated the system without any projection technique. The EER vary between 10 and 20 %, when comparing the feature vectors directly, without prior subspace projection.

word is then mapped to a real value between 0 and 2^x for x different filters. Finally we create a histogram from the mapped values in the BSIF image for describing the local properties of the image texture.

In our experiments, we use the standard filters, which represent eight different orientations of edges. As before, we extract a local descriptor for different window sizes, overlap between neighboring windows and different filter sizes and concatenate each local histogram to a global histogram representation. For all experiments, we use 8-bit code words and the 5×5, 11×11 and 17×17 filters

Histogram of Oriented Gradients (HOG): Computation of the HOG [57] descriptor involves five steps, which are the gradient computation, orientation binning, histogram computation, histogram normalization and concatenation of local histograms. The algorithm starts by computing the local gradient by convolving a 3×3 region (HOG cells) with two one-dimensional filters (-101) and $(-101)^T$. The local orientation at the center of each HOG cell is the weighted sum of the filter responses of each pixel.

The local orientations within a larger sub-window, denoted as block, are then quantized into bins in the $[0, 2\pi]$ interval. Subsequently, the image is divided into blocks of equal size and a local histogram of quantized orientations is extracted. Subsequently, the local histogram from each block is normalized with the L2-norm. Finally, all local histograms are concatenated to form the HOG descriptor for the image.

We evaluate the recognition performance of HOG using all possible combinations between 8×8, 16×16 and 32×32 HOG cells and 4, 9, 12 and 18 bin histograms.

Surface Descriptors: For describing three-dimensional structures in depth images, we use the shape index and curvedness. Both descriptors are based on the principal curvature of a shape and divide different surface structures into discrete categories, which are represented by values between 0 and 1. According to [28], the shape index for the maximum principal curvature k_{max} at a given position p in the image and the minimum principal curvature k_{min}, respectively, is defined as

$$S(p) = \frac{1}{2} - \frac{1}{\pi} \arctan\left(\frac{k_{max}(p) + k_{min}(p)}{k_{max}(p) - k_{min}(p)}\right) \tag{8.1}$$

Accordingly, the curvedness for a given image position can be written as

$$C(p) = \sqrt{\frac{k_{max}^2(p) + k_{min}^2(p)}{2}}. \tag{8.2}$$

8.3.2 Combined Descriptor

We compute a descriptor that combines 3-dimensional and 2-dimensional data; we extract local histograms from an image region that is defined by a sliding window with a given size and a given overlap between neighboring windows. For each local window, we extract the code word images from both, the texture images (2D) and the depth image (3D). The number of bins can either be the total number of values that can occur in the code word image, or any other number that divides the value range of the descriptor into n equally sized bins between the minimum and the maximum possible value code word image.

Subsequently, we create a local histogram from these features by using the code word (feature) from position $I_{2D}(p)$ from the 2D-image for determining the bin. The code word at position $I_{3D}(p)$ from the 3D-image determines the value that is added to the bin size. All local histograms are normalized using the L_2 norm and then concatenated to a global histogram descriptor.

An example of the computation of a combined feature vector using LPQ and the Shape Index is depicted in Fig 8.2. In steps 1 and 2 we compute a code word image using LPQ for the texture image and the Shape Index for the depth image. In the upper right corner of the Shape Index image, we can see an interpolated region that clearly differs from the noisy signal of the depth sensor. Information from both channels is combined in step 3, where

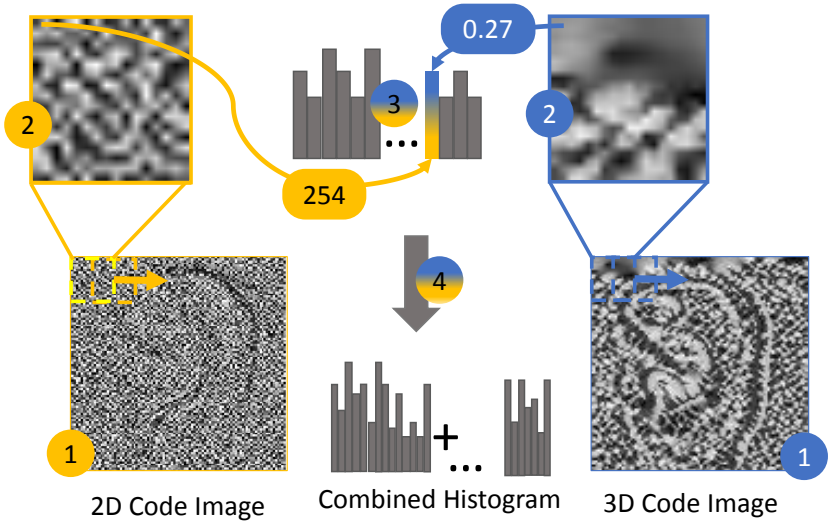


Figure 8.2: Example for feature level fusion for the creation of the combined histogram descriptor.

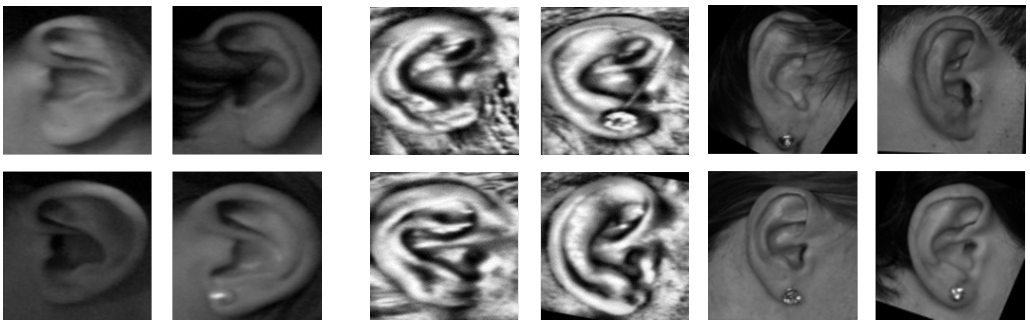


Figure 8.3: Example images from UND-J2, AMI and IITK (from left to right) after geometrical normalization with CPR and resized to 100×100 pixels.

the code word from the texture image determined the bin and the shape index from the depth is used for adding a weight to this bin. Finally, the combined histogram from LPQ and Shape Index data is normalized and then concatenated.

Obviously, this fusion scheme offers many different possibilities for combining texture images and co-registered depth images. We can either use the texture images for determining the bin and the depth images for adding a weight to this bin or vice versa. We evaluate the performance of both options. Moreover, we can vary the number of bins to get longer of shorter descriptors. Finally, this scheme also allows us to combine two different feature vectors from texture images or depth images. In our experiments we explore different combinations between the best performing configurations in texture- and depth images, as well as different combinations between texture descriptors in texture images and as the shape index or curvedness for depth images.

8.4 Experimental Results

For our experiments, we select all subjects with at least 5 samples from each of the databases. For each selected subject, we randomly pick 4 samples for training the projection matrix and for serving as reference images. The remaining samples are used for testing. For UND-J2 [201] we select 158 different subjects with 790 images in total. For AMI [70] the experiments are based on 500 images from 100 subjects and for IITK [107], we use 288 images from 72 different subjects. The settings for a particular configuration depend on the algorithm. Apart from the feature extraction method itself, a configuration also indicates the projection technique and the distance metric. Configurations are encoded as follows:

LBP: LBP - <radius; number of pixels> - <window size> - <overlap> - <projection technique> - <distance metric>

LPQ: LPQ - <radius> - <window size> - <overlap> - <projection technique> - <distance metric>

BSIF: BSIF - <filter size> - <window size> - <overlap> - <projection technique> - <distance metric>

HOG: HOG - <block size> - <number of bins> - <projection technique> - <distance metric>

Surface: < SI | C> (SI = Shape Index; C = Curvedness)

The configuration settings for combined descriptors are encoded as follows:
<bin configuration> - <number of bins> + <magnitude configuration> + <window size> - <overlap> - <projection technique> - <distance metric>

8.4.1 2D and 3D Appearance Features

In Table 8.1 we have collected a selection of results from our evaluations. In general, we observe that the overlap between neighboring windows does not have a major influence on the performance on any of the databases. The general expectation is, that smaller sizes of the local window yield a better performance than configurations with a larger local window. However, smaller local windows also imply that the information in the local histogram is bound to a smaller portion of the image. Moreover, it implies that the number of dimensions in the feature increases. The relation between these two factors should be balanced carefully. The local window size should be chosen in a way that the texture descriptor is still able to extract a sufficient amount of information. For instance, LPQ with a radius of 7 only extract 9 values from a 10×10 local window, which is not sufficient for a good recognition performance. In general, we recommend to observe the sparsity of a local histogram for balancing the feature extraction parameters and the local window size. In our experiments, the best performance could be achieved, if at least 25% of the bins in each local histogram are different from zero.

Concerning the selection and parameters for the texture descriptors, we achieved excellent performance indicators with BSIF and LPQ. Some configurations for HOG and LBP result in good performance values too, but are in general inferior to the performance of BSIF and LPQ. The best configuration for the grid size is, in many of the cases, the 20×20 window with an overlap of 5 pixels between neighboring windows.

Concerning the selection of a good window size, HOG plays a special role, compared to BSIF, LPQ and LBP. The local and non-overlapping windows in HOG (also referred to as HOG cells) are used directly for encoding local information, whereas all other algorithms extract local information from a sliding window. It is obvious that the window size has a strong impact on the performance of HOG. Based on our evaluations, we recommend window sizes between 8×8 and 16×16 pixels. Configurations with window sizes that are

Table 8.1: Equal Error Rate (EER) and Rank-1 performance (Rank-1) (in %) for selected configurations on datasets UND-J2, AMI and IITK. The best configuration for each dataset is marked in bold.

Algorithm	**UND-J2**		**AMI**		**IITK**	
	EER	Rank-1	EER	Rank-1	EER	Rank-1
LPQ-3-20-10-LDA-cos	1.35	96.27	0.75	96.50	0.21	99.31
LPQ-5-33-15-LDA-cos	1.14	97.47	**0.00**	**100.00**	0.71	97.78
LPQ-11-20-10-LDA-cos	**0.67**	**98.73**	0.01	99.00	0.50	98.89
LPQ-11-33-15-LDA-cos	0.72	98.54	**0.00**	**100.0**	**0.17**	**99.03**
BSIF-5-20-15-LDA-cos	**0.39**	**98.67**	2.00	97.00	0,59	98,33
BSIF-5-20-15-PCA-cos	12,13	82,53	30.26	51.00	16.57	72.92
BSIF-11-20-5-LDA-mahcos	0.83	97.91	0.01	100.0	**0.19**	**99.44**
BSIF-11-20-10-LDA-cos	0.64	97.85	**0.00**	**100.0**	0.36	99.44
BSIF-11-20-15-LDA-cos	0.72	97.97	0.75	95.00	0,31	98.75
BSIF-17-20-15-LDA-cos	0.89	97.78	0.26	97.5	0.54	99.17
LBP-18-10-5-LDA-cos	0.98	97.34	0.68	97.50	0.69	98.47
LBP-18-20-5-LDA-cos	1.04	96.77	**0.55**	**97.20**	1.53	98.47
LBP-18-20-10-LDA-cos	1.73	96.77	0.66	97.50	1.94	97.92
LBP-18-20-5-LDA-cos	**0.94**	**97.22**	1.04	96.20	**0.48**	**98.89**
LBP-18-33-15-LDA-cos	7.88	69.24	5.14	82.70	0.68	97.87
HOG-8-18-LDA-cos	**1.27**	**97.85**	2.20	95.00	1.11	98.33
HOG-8-12-LDA-cos	1.44	96.77	2.61	93.50	1.66	97.64
HOG-16-18-LDA-cos	3.11	93.23	**0.13**	**100.0**	1.67	96.81
HOG-16-4-LDA-cos	12.11,	71.15	3.00	93.00	**0.52**	**98.61**
HOG-32-18-LDA-mahcos	14.72	66,52	16.25	48.50	3.30	93.47

larger than 16×16 pixels did not perform well. To our surprise, the number of histogram bins plays a minor role for the effectiveness of the HOG descriptor.

The best combination between feature subspace projection technique and the distance measure is LDA in conjunction with cosine distance. The performance of LDA with cosine distance is close and in many cases within the confidence interval. We recommend the cosine distance measure together with LDA for ear recognition using texture features. Other distance measures resulted in significantly less recognition accuracy than cosine and mahalanobis distance and can hence be discarded. Kernel methods for feature subspace projection (KFA and KPCA) take longer time for computation and yield less accuracy for recognition. The performance of configurations using PCA and BSIF, LPQ or LBP is between two and 10 times worse than LDA, depending on the database and feature extraction algorithm. For configurations using HOG, the performance of PCA is similar to LDA.

In Table 8.2 we compare the performance of 2D and 3D images using selected configurations from the previously used texture recognition configurations. The performance of depth images is always lower than for texture images. Even though depth images do not contain artefacts from lighting, they seem to lack a sufficiently large amount of detail. Moreover, the surface structures in our test data are noisy.

Since we obtain a recognition performance that is significantly different from zero, we conclude that texture descriptors can represent surface structures in depth images. We also observe that a good performance for a particular configuration of a texture descriptor also indicates a good performance for the same configuration in the depth image. We conclude from this that texture descriptors can be used to describe surface structure in depth images

Table 8.2: Equal Error Rate (EER) and Rank-1 performance (Rank-1) (in %) for texture and depth images in dataset UND-J2 and selected configurations. The best configuration for each dataset is marked in bold.

Algorithm	**texture image (2D)**		**depth image (3D)**	
	EER	Rank-1	EER	Rank-1
BSIF-5-33-5-LDA-cos	1.01	96.96	2.09	95.31
BSIF-5-20-15-LDA-cos	0.39	98.67	1.81	95.25
BSIF-11-20-15-LDA-cos	0.72	97.97	2.58	93.23
BSIF-17-20-15-LDA-cos	0.89	97.78	5.43	85.39
LPQ-3-20-10-LDA-cos	1.35	96.27	2.81	95.24
LPQ-5-33-15-LDA-cos	1.14	97.47	3.59	90.32
LPQ-11-20-10-LDA-cos	0.67	97.97	2.34	93.80
LPQ-11-33-15-LDA-cos	0.72	97.59	0.61	98.86
LBP-18-20-15-LDA-cos	0.94	97.22	2.50	95.25
LBP-18-10-15-LDA-cos	0.98	97.34	3.51	91.51
LBP-18-33-15-LDA-cos	7.88	69.24	5.76	79.56
HOG-8-18-LDA-cos	1,27	97.85	3,58	91.96
HOG-16-18-LDA-cos	3.12	93.23	3.29	91.46

as well and that the behavior of the feature vectors for texture and depth images is similar. We also expect that the performance of depth image could exceed the performance of texture images, if the texture images contain shadows or have a low contrast.

8.4.2 Combined Histogram Features

A selection of evaluation results from the second series of experiments is summarized in Table 8.3. Compared to the results on texture images in the previous experiment, the recognition performance could not be improved by fusion texture and depth images.

The best combination in our experiments was to use the texture information for defining the bin and the depth information for computing the magnitude. In many cases the results of these configurations were comparable to that of the texture based pipelines. LDA with cosine distance still appeared to be the best combination for most of the combined feature descriptors and the overlap between neighboring 20×20 windows should be 15 pixels.

An exception to this is LBP, where we observed low performance rates for all combined descriptors. Moreover we also found that the number of bins should be the same as for the texture descriptors only (namely 256). All configurations with a smaller number of bins performed significantly worse, which is due to a loss of information when several bin descriptors are mapped to the same bin. The fact that the combined descriptors generally perform worse than the texture descriptors may be caused by large amounts of noise in the depth channel, which can be seen in the example image on the right hand side in Fig. 8.2. Smoothing however, does not improve the performance, because it also destroys valuable details. Artefacts from the interpolation process may also affect the recognition performance.

Combining texture descriptors in the texture and the depth channel appears to be infeasible. The large amount of noise in the depth channel causes many errors during the assignment of bins, which is responsible for the low recognition rates. The fact that the performance for a small number of bins in the combined histogram does not degrade implies, that the combined histograms are sparse. Apparently, the loss of information from merging neighboring bins does not affect the performance.

Table 8.3: Equal Error Rate (EER) and Rank-1 performance (Rank-1) (in %) for selected combined descriptors.

Algorithm	**Performance**	
	EER	Rank-1
2D bins, 3D magnitude		
BSIF-5-256-SI-20-15-LDA-cos	**0.95**	**97.53**
BSIF-5-256-C-20-15-LDA-cos	1.32	96.08
LPQ-3-256-SI-20-10-LDA-cos	1.96	93.67
LPQ-3-256-C-20-10-LDA-cos	3.47	88.16
LBP-3-256-SI-20-15-LDA-cos	18.53	36.65
LBP-3-256-C-20-15-LDA-cos	19.60	30.76
3D bins, 2D magnitude		
SI-64-LBP-18-20-15-LDA-cos	26.06	19.24
SI-256-LBP-18-20-15-LDA-cos	27.35	18.54
SI-256-BSIF-5-20-15-LDA-cos	10.98	60.89
SI-8-BSIF-5-20-15-LDA-cos	12.27	61.58
SI-64-BSIF-5-20-15-LDA-cos	9.15	69.94
SI-256-LPQ-3-20-15-LDA-cos	19.40	37.66
SI-32-LPQ-3-20-15-LDA-cos	**8.89**	**74.74**
C-256-LBP-18-20-15-LDA-cos	33.43	9.11
C-64-LBP-18-20-15-LDA-cos	44.46	2.91
C-256-BSIF-5-20-15-LDA-cos	8.89	74.75
C-8-BSIF-5-20-15-LDA-cos	31.14	17.48
C-64-BSIF-5-20-15-LDA-cos	10.47	64.57
C-256-LPQ-3-20-15-LDA-cos	16.43	46.14
C-32-LPQ-3-20-15-LDA-cos	32.48	15.76
Texture descriptors (2D, 3D)		
BSIF-5-256-LPQ-3-20-15-LDA-cos	**4.21**	**85.25**
BSIF-5-256-LBP-18-20-15-LDA-cos	15.18	53.54
LPQ-3-256-LBP-18-20-15-LDA-cos	26.61	18.04
LPQ-3-256-BSIF-5-20-15-LDA-cos	2.67	94.75
LBP-18-256-LPQ-3-20-15-LDA-cos	14.13	56.14
LBP-18-256-BSIF-5-20-15-LDA-cos	14.15	54.62

8.5 Conclusion

Texture descriptors that were originally proposed for face recognition or other computer vision tasks can also be applied for ear recognition in texture and depth images. Our extensive evaluations show that the best performance for three different datasets is achieved with the LPQ and the BSIF descriptor in conjunction with LDA as dimensionality reduction methods and the cosine distance for a nearest neighbor classifier. The size and overlap of the local window should be balanced with the parameters for the feature extraction approach. In our experiments, we found that smaller local windows with more spatially bound descriptors do not improve the performance, because smaller radii for the local descriptors are more vulnerable to noise and the number of dimensions in the concatenated histogram becomes overly long.

We also proposed an approach, where texture and depth data is fused on the feature level and evaluated the performance. However, the performance of the combined descriptor turned out to be inferior to comparable configurations, where only the texture data is

used. The usage of surface descriptors does not allow for a linear assignment of histogram bin, because the resulting histograms are sparsely populated.

Even though the use of histogram-based texture descriptors provides a method for generating compact feature vectors, we observe that the local histograms are sparsely populated. We plan to work towards a binary representation of the local feature vectors with the goal of providing a fast and accurate ear recognition system.

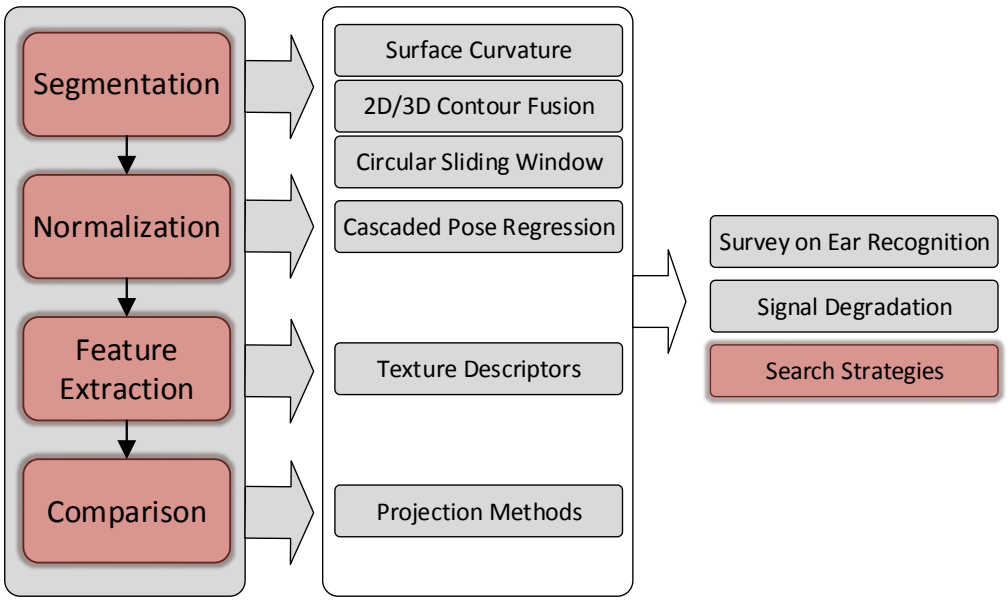
Chapter 9

Binarization of Histogram Models: An Application to Efficient Biometric Identification

In this chapter we address the research question **Q4: How can ear templates be represented in order to enable fast search operations?** Binary representation of histogram descriptors can be used for retrieving a short list with the most likely candidates from the database. The fast comparison speed of binary data accelerates the comparison by an order of magnitude and reduces the chance for a phase positive at rank 1 of the re-sorted candidate list.

This work makes use of the observation that we obtain high recognition performance rates with sparsely populated histograms. This brought us to the conclusion that the fact that a given bin is different from zero may already be an important information that would allow us to pre-screen the dataset for the most likely candidates. We propose a generic method for deriving binary feature vectors from fixed-length histograms and show that these binary feature vectors can, for example, be used for a sequential search in ear recognition and for palm print recognition.

This work is published in [151] ANIKA PFLUG, CHRISTIAN RATHGEB, ULRICH SCHERHAG, CHRISTOPH BUSCH, Binarization of Histogram Models: An Application to Efficient Biometric Identification, International Conference on Cybernetics(CYBCONF), 2015



Abstract

FFeature extraction techniques such as local binary patterns (LBP) or binarized statistical image features (BSIF) are crucial components in a biometric recognition system. The vast majority of relevant approaches employs spectral histograms as feature representation, *i.e.* extracted biometric reference data consists of sequences of histograms. Transforming these histogram sequences to a binary representation in an accuracy-preserving manner would offer major advantages *w.r.t.* data storage and efficient comparison.

We propose a generic binarization for spectral histogram models in conjunction with a Hamming distance-based comparator. The proposed binarization and comparison technique enables a compact storage and a fast comparison of biometric features at a negligible cost of biometric performance (accuracy). Further, we investigate a serial combination of the binary comparator and histogram model-based comparator in a biometric identification system. Experiments are carried out for two emerging biometric characteristics, *i.e.* palmprint and ear, confirming the soundness of the presented technique.

9.1 Introduction

According to the ISO standard, biometrics is referred to as "automated recognition of individuals based on their behavioural and biological characteristics" [93]. For different biometric characteristics, *e.g.* face or palmprint [217, 104], the task of recognizing individuals can be reduced to texture modelling/ recognition tasks. In the past years, numerous holistic image analysis techniques, which are designed to construct local image descriptors which efficiently encode texture information of image regions, have been found to perform well in biometric recognition. However, biometric applications have to fulfil additional requirements which may not be met by texture recognition approaches. In particular, the representation of extracted features as vectors of spectral histograms requires a complex comparator and, thus, hampers biometric systems to be operated in identification mode which requires an exhaustive $1 : N$ comparison, where N represents the number of subjects registered with the system. In addition, the storage of biometric reference data (templates) on smart card chips, magnetic stripes, or 2D bar codes requires a highly compact representation of feature vectors.

In order to address the aforementioned issues we propose a generic binarization scheme for spectral histogram models. We evaluate this approach for palmprint and ear images, however our method is likely to work for any other N-class separation problem where spectral histogram models can be applied. In particular, this means that the input images should be segmented and normalized w.r.t. rotation and scale. In the presented approach coefficients of histogram-based feature vectors are binarized according to a statistical analysis, reducing the template size by an order of magnitude. Moreover, the most reliable bits are detected for each subject. We design a Hamming distance-based comparator for efficient pair-wise comparison, employing a bit-mask in order to only compare the most reliable bits of subjects within relevant regions. Due to a potential loss of information during binarization we also suggest to employ a serial combination of comparison subsystems based on binary feature vectors and spectral histogram models, in order to accelerate biometric identification. For this purpose binary templates are used for pre-screening the entire database, *i.e.* the original comparison subsystem is only applied for a short-list of top-ranked candidates. The effectiveness of this approach has already been demonstrated for other biometric characteristics, cf. [69, 30]. Experiments, which are conducted on the publicly available PolyU palmprint database and the UND-J2 ear database, show that the proposed system based on binarized feature vectors maintains accuracy comparable to systems where exhaustive search using a real-valued feature vector is applied for identity retrieval [1].

[1] We also conducted experiments on iris and face data and obtained similar results as fr ear and palm. We believe that our approach can also be applied to other modalities than ear and palm.

9.2 Related Work

Spectral Histogram Models have been widely used in the field of biometrics. Prominent representatives include LBP [9], BSIF [98], local phase quantisation (LPQ) [10] or histograms of oriented gradients (HOG) [56]. Based on a sub-window-wise binary encoding of texture features, spectral histograms are extracted for a sub-window in order to retain a certain amount of spatial information. During recognition, sequences of histograms from probe feature vectors are compared to reference histograms. A variety of approaches using edges, statistical models and holististic descriptors in combination with subspace projection were evaluated for both biometric characteristics [104, 147]. According to previous work, spectral histogram descriptors were found to yield good recognition performance for both, ear and palm print images [150, 160].

However, spectral histograms represent a rather in-efficient representation of biometric reference data, requiring large template sizes (histogram bins have to be stored as non-binary values, e.g. integers, where in general a reasonable number of bins is expected to remain empty) and comparators require complex calculations in order to estimated comparison scores, e.g. χ^2 distance. In order to provide a compact representation and fast comparison of biometric feature vectors different approaches have been proposed to obtain binary feature vectors from biometric characteristics, e.g. minutiae cylinder codes [44] or iris-codes [59].

When referring to workload reduction w.r.t. biometric identification, we coarsely distinguish three key approaches: (1) classification, (2) indexing, and (3) a serial combination of a computationally efficient and an accurate (but more complex) algorithm. Let N denote the number of subjects registered with a biometric system and w be the workload for a pair-wise comparison of two templates. Then the overall workload W for biometric identification is defined as $W = wN + \delta$, where δ summarizes any additional one-time costs, e.g. sorting of candidates. In case the entire feature space is divided into c classes (i.e. subsets), W can be reduced to $wN/c + \delta$, given that the registered subjects are equally distributed among all classes. For instance, in [94, 166] and [152] fingerprint and ear images are assigned to $c = 5$ and $c = 4$ classes, respectively. It is generally conceded that small intra-class and large inter-class variations as well as sufficient image quality represent essential preliminaries in order to achieve acceptable preselection error rates.

Biometric indexing aims at reducing the overall workload in terms of $\mathcal{O}$-notation. While an optimal indexing scheme would require a workload in $\mathcal{O}(1)$, existing approaches focus on reducing the workload to at least $\mathcal{O}(\log N)$, yielding $W = w\log(N)$. In the majority of cases this is achieved by introducing hierarchical search structures which tolerate a distinct amount of biometric variance. Most noticeable indexing schemes for iris and fingerprints have been presented in [74] and [45], respectively.

Within serial combinations computationally efficient algorithms are used to extract a short-list of $\mathcal{L}N$ most likely candidates, with $\mathcal{L} \ll 1$. Therefore, W is reduced to $\hat{w}N + w\mathcal{L}N$, where $\hat{w}$ is the workload of a pair-wise comparison of the computationally efficient algorithm, $\hat{w} \ll w$. In other words, identification is accelerated if $w(1 - \mathcal{L}) > \hat{w}$ holds. In [69] and [30] $\mathcal{L}$ was reduced to ∼10% for iris and voice, respectively, significantly accelerating biometric identification. Compared to indexing and classification a serial combination of algorithms enables a more accurate operation of the resulting trade-off between computational effort and accuracy by setting an adequate threshold for $\mathcal{L}$.

9.3 Binarization of Spectral Histogram Models

In the following subsections we define the terminology used w.r.t. spectral histogram models and provide a detailed description of the proposed binarization, the corresponding comparator, and the resulting serial identification scheme.

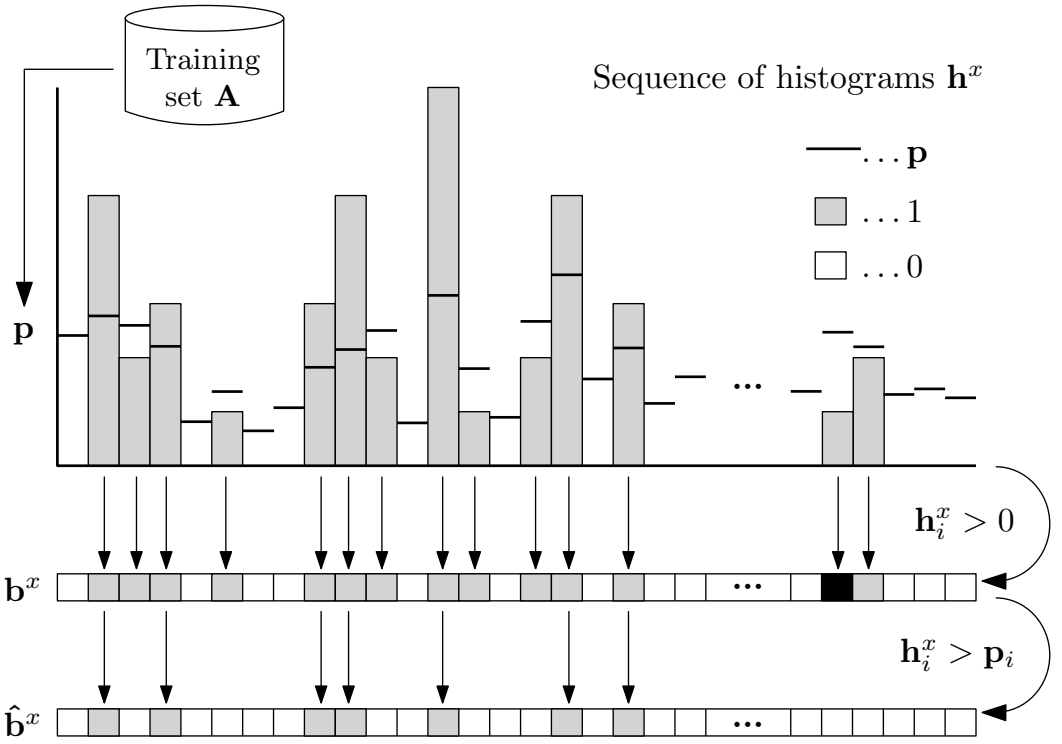


Figure 9.1: Proposed binarization: two binary feature vectors are extracted out of a sequence of histogram coefficients.

9.3.1 Binarization

Given a preprocessed input image of subject x, we extract a sequence of n histogram coefficients $\mathbf{h}^x \in \mathbb{N}_0^n$. According to the employed sub-window dimension this sequence may consist of numerous fixed length spectral histograms. As mentioned earlier, the purpose of sub-windows in conjunction with spectral histogram features is to preserve spatial information within spectral histograms.

In the first step, $\mathbf{h}^x$ is binarized in order to obtain $\mathbf{b}^x \in \{0,1\}^n$ which points at all non-zero coefficients of $\mathbf{h}^x$,

$$\mathbf{b}_i^x = \begin{cases} 0, & \text{if } \mathbf{h}_i^x = 0 \\ 1, & \text{if } \mathbf{h}_i^x > 0. \end{cases} \tag{9.1}$$

As we will show in our experiments, the distribution of non-empty bins is discriminative for each subject. Depending on the particular settings, e.g. size of the local sub window or number of feature descriptor values, a significant number of histogram bins can be expected to remain empty.

Using a training set $\mathbf{A}$ the mean-vector $\mathbf{p} \in \mathbb{R}^{n+}$ of all non-zero coefficients is estimated,

$$\mathbf{p}_i = \frac{\sum_{a \in \mathbf{A}} \mathbf{h}_i^a}{\sum_{a \in \mathbf{A}} \mathbf{b}_i^a}. \tag{9.2}$$

In the second step $\mathbf{h}^x$ is, again, binarized obtaining $\hat{\mathbf{b}}^x \in \{0,1\}^n$ in relation to the previously estimated $\mathbf{p}$,

$$\hat{\mathbf{b}}_i^x = \begin{cases} 0, & \text{if } \mathbf{h}_i^x \leq \mathbf{p}_i \\ 1, & \text{if } \mathbf{h}_i^x > \mathbf{p}_i. \end{cases} \tag{9.3}$$

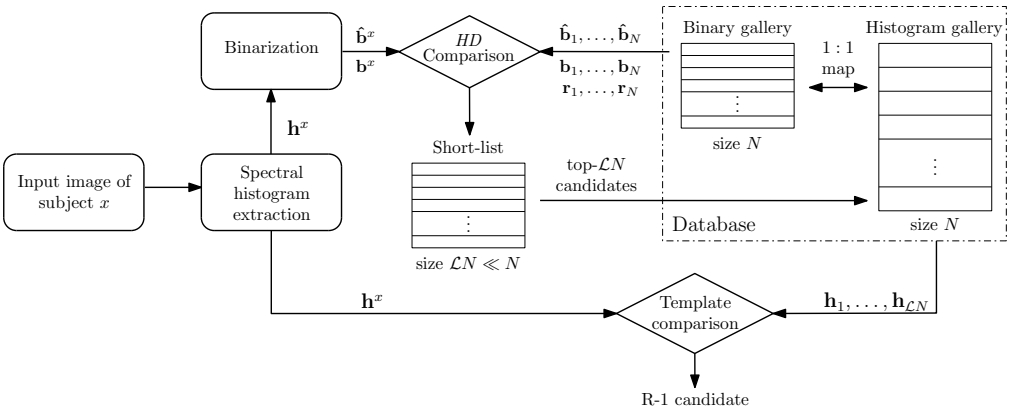


Figure 9.2: Serial combination of computationally efficient and original comparator: the Hamming distance-based comparator is employed to perform an $1 : N$ comparison, returning a list of $\mathcal{L}N$ candidates on which the original comparator is applied.

The process of generating both binary vectors is schematically depicted in Fig. 9.1. We assume that for each subject $e > 1$ sample images are acquired during enrolment. Based on e acquisitions we calculate $\mathbf{p}^x$, $\boldsymbol{\sigma}^x \in \mathbb{R}^{n+}$, representing the subject-specific mean and variance vector, with $\boldsymbol{\sigma}_i^x = Var(\mathbf{p}_i^x)$. Then vector $\mathbf{r}^x \in \mathbb{R}_0^{n+}$, which defines the reliability of each coefficient, is defined as,

$$\mathbf{r}_i^x = \frac{\parallel \mathbf{h}_i^x - \mathbf{p}_i^x \parallel}{\boldsymbol{\sigma}_i^x}. \tag{9.4}$$

In other words, we identify those coefficients as reliable which are far from the mean (discriminative) and exhibit low variance (constant). Based on $\mathbf{r}^x$ we can obtain a binary vector $\mathbf{r}_k^x \in \{0,1\}^n$, pointing at the k most reliable bits of subject x.

9.3.2 Comparator

The (dis-)similarity $d(x,y)$ of a given probe sample $\mathbf{h}^y$ to a gallery template of subject x is finally defined as the fractional Hamming distance between binary feature vectors $\hat{\mathbf{b}}^x$ and $\hat{\mathbf{b}}^y$ deemed to the k most reliable bits of the gallery vector intersected with sets of non-zero coefficients,

$$d(x,y) = \frac{\parallel (\hat{\mathbf{b}}^x \oplus \hat{\mathbf{b}}^y) \cap \mathbf{r}_k^x \cap \mathbf{b}^x \cap \mathbf{b}^y \parallel}{\parallel \mathbf{r}_k^x \cap \mathbf{b}^x \cap \mathbf{b}^y \parallel}. \tag{9.5}$$

The XOR operator $\oplus$ detects disagreements between any corresponding pair of bits between the binary vectors $\hat{\mathbf{b}}^x$ and $\hat{\mathbf{b}}^y$. The result is intersected with the reliability mask $\mathbf{r}_k^x$ of the claimed identity x, i.e. the AND operator $\cap$ ensures that only the k most discriminative bits are used for comparison. Further, the resulting bit vector is intersected with $\mathbf{b}^x$ and $\mathbf{b}^y$ such that only relevant (non-zero) areas are considered at the time of comparison. Subsequently, the final score is normalized accordingly. The overlap mask is necessary, because we need to make sure that a probe and a reference feature vector have a sufficiently large and overlapping number of non-zero bins. The coefficient in these bins should have a small distance to the gallery feature vector.

It is important to note that the proposed comparator is highly efficient as it is only based on those two logical operators. It has been shown that Hamming distance-based comparators are capable of performing millions of comparisons per second [59].

Table 9.1: Properties of the Poly-U palmprint datbase and the UND-J2 ear database and the number of resulting identification attempts.

Dataset Dataset	Number of Subjects	Number of Images	Image Resolution	Number of Identifications
Poly-U	250	6000	128×128	500
UND-J2	312	1536	100×100	312

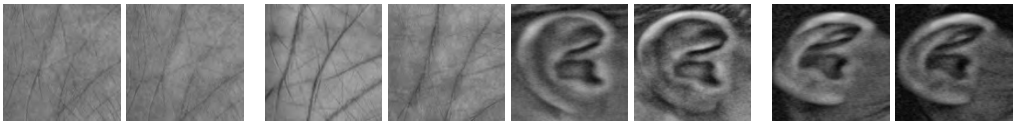


Figure 9.3: Sample images of two subjects of the Poly-U palmprint datbase (top row) and two subjects of the UND-J2 ear database (bottom row).

9.3.3 Application to Serial Identification

Since the proposed binarization may cause a significant loss of biometric information and, thus, biometric performance, binarized templates can be alternatively used in a serial combination in order to accelerate identification. In this case the computational efficient Hamming distance-based comparator is utilized to pre-screen the entire database. For this purpose N pair-wise comparisons are performed, resulting in a vector $\mathbf{D}$ of dissimilarity scores, sorted in descending order $\mathbf{D} = (d_1 \leq d_2 \leq \cdots \leq d_N)$ where d_i denotes the score between the query and the i-th enrolment data record. Finally, the top-$\mathcal{L}N$ candidates, i.e. the candidates which the first $\mathcal{L}N$ scores in $\mathbf{D}$ point at, are returned.

Based on the short-list returned in the pre-screening stage the probe is compared against a fraction of $\mathcal{L}N$ original gallery templates applying a more complex comparator, e.g. Euclidean distance or χ^2 distance. The entire process is depicted in Fig. 9.2.

Based on the terminology defined in Sect. 9.2, we assume that $\hat{w} \ll w$. By defining a speed-up factor $\alpha = w/\hat{w}$, we can estimate the maximum list size $\mathcal{L}$ such that $\hat{w}N + w\mathcal{L}N < wN$ still holds, yielding $\mathcal{L} \leq 1 - 1/\alpha$ (assuming that additional one-time costs are comparable for both systems). For example, if the pre-screener is 5 times faster than the original comparator the $\mathcal{L}$ should be significantly smaller than 80% of the entire number of registered subjects in order to obtain a substantial speed-up.

9.4 Experimental Evaluations

In the following subsections we describe the setup of conducted experiments and report the performance of the proposed approach as well as a serial combination of comparators with respect to accuracy and identification speed.

9.4.1 Experimental Setup

Experiments are carried out on two different databases, the Poly-U palmprint database[2] [214] and the UND-J2 ear database[3] [202]. Properties of these datasets are summarized in Table 9.1. We use four images per subjects for enrolment and the remaining images for performance evaluation. ROIs of the palm print images from PolyU are already segmented

[2]Publicly available at http://www4.comp.polyu.edu.hk/~biometrics/MultispectralPalmprint/MSP.htm

[3]Publicly available at http://www3.nd.edu/~cvrl/CVRL/Data_Sets.html

Table 9.2: Identification rates and hit rates for various values of $\mathcal{L}$ (in %) for PolyU-MS (top) and UND-J2 (bottom) and feature extraction algorithms using k most reliable bits during comparison.

k	**IR**	$\mathcal{L}$=**5**	$\mathcal{L}$=**10**	$\mathcal{L}$=**30**	$\mathcal{L}$=**40**	$\mathcal{L}$=**50**	k	**IR**	$\mathcal{L}$=**5**	$\mathcal{L}$=**10**	$\mathcal{L}$=**20**	$\mathcal{L}$=**30**	$\mathcal{L}$=**50**
PolyU-MS with LPQ							**PolyU-MS with BSIF**						
1%	69.11	97.33	99.33	100.0	100.0	100.0	1%	52.44	93.78	97.56	99.11	100.0	100.0
2%	88.67	99.33	99.78	100.0	100.0	100.0	2%	75.33	98.67	99.56	100.0	100.0	100.0
3%	91.33	99.56	100.0	100.0	100.0	100.0	3%	84.00	99.11	100.0	100.0	100.0	100.0
5%	96.22	99.78	100.0	100.0	100.0	100.0	5%	88.00	99.78	99.78	99.78	100.0	100.0
7%	97.77	99.78	100.0	100.0	100.0	100.0	7%	93.78	100.0	100.0	100.0	100.0	100.0
10%	98.22	100.0	100.0	100.0	100.0	100.0	10%	97.11	100.0	100.0	100.0	100.0	100.0
30%	99.78	100.0	100.0	100.0	100.0	100.0	30%	99.33	100.0	100.0	100.0	100.0	100.0
100%	100.0	100.0	100.0	100.0	100.0	100.0	100%	100.0	100.0	100.0	100.0	100.0	100.0
UND-J2 with LPQ							**UND-J2 with BSIF**						
1%	36.03	66.67	78.38	91.89	94.59	97.29	1%	52.68	71.43	78.57	89.29	95.54	98.21
2%	47.37	64.91	80.70	94.74	98.25	99.12	2%	56.64	72.57	84.07	94.69	96.46	99.11
3%	61.40	78.95	83.33	92.11	96.49	100.0	3%	46.85	69.34	81.08	90.99	93.69	99.10
5%	64.66	81.90	91.38	98.28	100.0	100.0	5%	62.39	76.15	86.24	96.33	98.17	98.17
7%	61.54	77.78	86.32	98.29	100.0	100.0	7%	63.79	78.44	87.93	96.55	97.41	99.14
10%	52.99	78.63	90.60	96.58	99.14	100.0	10%	65.52	76.72	81.03	93.10	95.67	99.14
30%	63.48	78.26	89.57	99.13	100.0	100.0	30%	65.79	79.82	86.84	94.74	98.25	100.0
100%	73.15	84.26	90.74	95.37	98.15	99.07	100%	78.38	84.68	90.99	96.39	99.10	99.10

and normalized, such that we can extract the feature vectors without any additional preprocessing. For the UND-J2 dataset, which is acquired from a specified distance and contains slight variations in pose, we perform an automated normalisation process using Cascaded pose regression. The normalization algorithm was trained using images of 92 subjects. For more details on the preprocessing toolchain for ear images we refer the reader to [148]. Fig. 9.3 shows some examples for the palmprint images from PolyU and the segmented and normalized ear images from UND-J2.

Performance is estimated in terms of (true-positive) identification rate (IR). In accordance to the ISO/IEC IS 19795-1 [90] the IR is the proportion of identification transactions by subjects enrolled in the system in which the subject's correct identifier is the one returned. In experiments identification is performed in the closed-set scenario returning the rank-1 candidate as identified subject (without applying a decision threshold). Further, focusing on the serial combination of the binary system and the original one we report the penetration rate as the pre-chosen value $\mathcal{L}$ and the pre-selection error denoted by $\mathcal{P}$, defined as the error that occurs when the corresponding enrolment template is not in the preselected subset of candidates when a sample from the same biometric characteristic on the same user is given. We define the hit-rate of a system as $1 - \mathcal{P}$ for a chosen value of $\mathcal{L}$.

In the feature extraction stage we use LPQ and BSIF as representatives for spectral histogram features, for further details on these algorithms the reader is referred to [10, 98]. Both feature extractors use 20×20 pixel sub windows with an overlap of 15 pixels between neighbouring windows. LPQ is computed with a radius of 3 pixels and BSIF is using a 5×5 pixel filter. This results in a fixed length histogram feature vectors comprising 20736 and 9216 histogram coefficients for both databases for the LPQ and the BSIF feature vectors, respectively. At the time of comparison of spectral histogram features we employ the χ^2-distance which estimates the normalized quadratic distance between histogram coefficients of fixed length vectors.

Finally, we choose a suitable size for the training set $\mathbf{A}$ used to generate the mean vector $\mathbf{p}$. For this purpose we estimate the IR for different training set sizes using only the vectors $\hat{\mathbf{b}}$s which are binarized according to the resulting mean vectors. As shown in Fig. 9.4, for

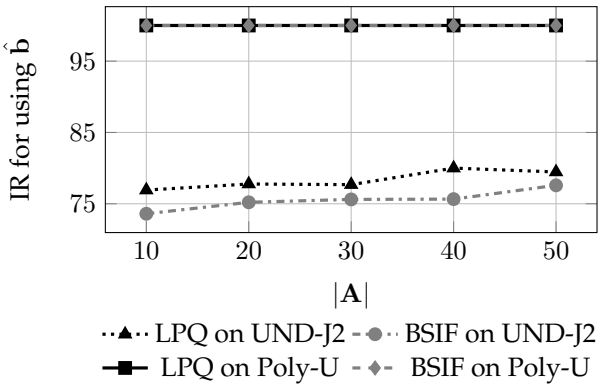


Figure 9.4: IR for different numbers of training subjects $|\mathbf{A}|$ using only the vectors $\hat{\mathbf{b}}$s which are binarized according to the obtained mean vector $\mathbf{p}$.

the favourable Poly-U dataset a training set size of $|\mathbf{A}| = 10$ already reveals a perfect system (IR=100%). However, for the more unconstrained UND-J2 IRs improve with an increased $|\mathbf{A}|$, i.e. we identify $|\mathbf{A}| = 50$ as a suitable choice for both datasets.

9.4.2 Histograms vs. Binarized Features

For the LPQ feature extraction the original systems based on histogram features and the χ^2-distance yield a baseline performance of IRs of 100% and 83.05% for the Poly-U and the UND-J2 dataset, respectively. By analogy for the BSIF feature extraction we obtain IRs of 100% and 81.6%. We also tested the Euclidean distance as an alternative comparator, however, it was outperformed by χ^2-distance in all experiments. We observed that binary comparison is α =10.61 times faster than χ^2 and α =7.61 times faster than the Euclidean distance in our C++ implementation of the system. Table 9.2 summarizes the biometric performance in terms of IR and rank-$\mathcal{L}$ identification rate which corresponds to $1 - \mathcal{P}$ for binarized vectors obtained from both datasets using both feature extractors for various $\mathcal{L}$s and numbers of reliable bits, ks. For the Poly-U dataset (upper half of Table 9.2) the proposed binarization technique maintains the biometric performance of IR=100% for both feature extraction algorithms. Further, we observe that the amount of employed bit comparison can be reduced to k =30% of most reliable bits maintaining comparable biometric performance. At higher ranks, e.g. $\mathcal{L}$ =30%, hit-rates of 100% are obtained even for comparing k =1% of most reliable bits.

On the more challenging UND-J2 dataset (lower half of Table 9.2) the proposed binarization technique suffers from a significant loss of biometric information. However, at higher ranks, reasonable hit-rates are obtained for both feature extractors as shown in the cumulative match score distribution (CMC) curves plotted in Fig. 9.5 for k =30%. The binarized feature vectors achieve a comparable performance w.r.t. the original system at ranks of approximately 10-15%. That is, for the challenging UND-J2 dataset we identify the proposed binarization technique as a suitable candidate for an efficient pre-screener in a serial combination of algorithms. Further, the need for employing both types of binarized feature vectors is underlined, as the sole use of $\hat{\mathbf{b}}$ reveals significantly inferior results for both feature extractors compared to the proposed combination of $\mathbf{b}$ and $\hat{\mathbf{b}}$.

9.4.3 Serial Identification

In our second experiment, we apply the proposed binarization technique in a serial combination on the UND-J2 dataset (see Sect. 9.3.3). We use the best configurations from the previous experiments and use them for generating the short list of $\mathcal{L}N$ candidates. We then

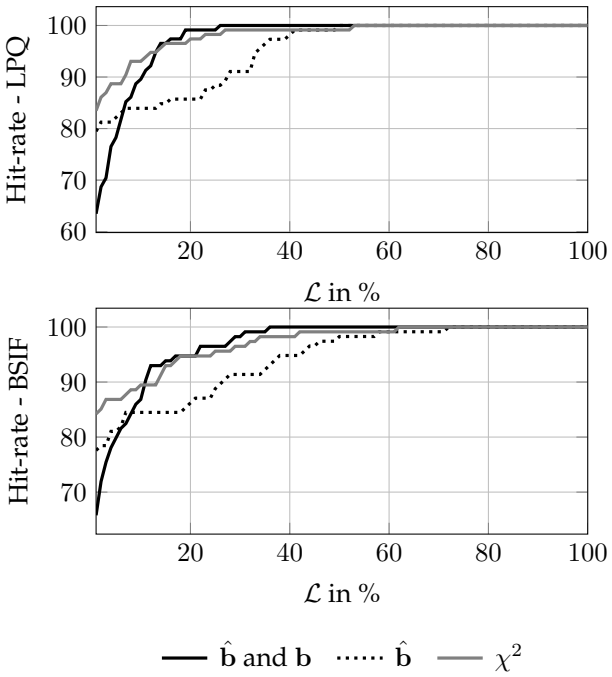


Figure 9.5: CMC curves of the proposed binarizations compared to the original systems.

re-order this short-list using the χ^2-distance on the original histogram-based templates and return the IR of the re-ordered list. The performance of the serial identification system is hence reported as a function of $\mathcal{L}$ and the IR of the re-ordered list which is depicted in Fig. 9.6 for both feature extraction algorithms. The IR of the serial identification system reaches it's maximum for list sizes between $\mathcal{L}$ =10-30% of the dataset. Based on the previously estimated speed-up factor of α =10.61 our short list has to be significantly smaller than 90% of the entire dataset in order to achieve a substantial speed-up compared to an full $1:N$ search using histogram-based feature vectors. Given this, we can conclude that the proposed serial identification system can perform an exhaustive search in 30% of the time compared to an exhaustive search using the χ^2-distance.

When comparing the IR of the serial identification system with the reference IR in Fig. 9.5, we can also see that the serial identification system outperforms the reference system that does an exhaustive $1:N$ search using the χ^2-distance. This increase in performance is achieved due to the fact that the proposed system already discards candidates in the pre-screening stage which would have been falsely identified by the original system within an exhaustive search.

9.5 Conclusions

In this paper we have presented a generic binarization scheme for holistic spectral histogram descriptors and provided empirical results on two datasets and two different texture descriptors. The binary feature vectors are directly computed from the histogram representation and hence do not require an additional feature extraction. The proposed binarization technique can be used to reduce the size of biometric reference data and to perform efficient identification. We do not require an additional processing step for acquiring a binary representation of the feature vectors. Instead, we compute a binary representation directly from the spectral histogram descriptors.

The proposed method can be applied to all images, where a spectral histogram feature

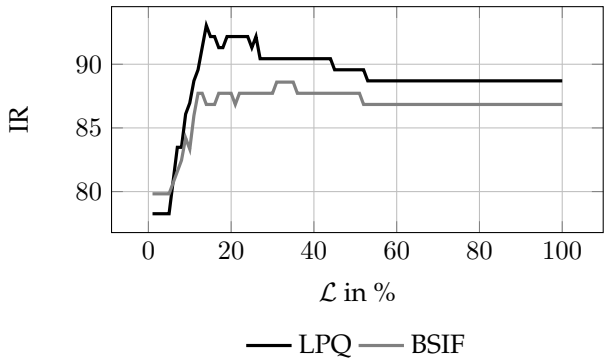


Figure 9.6: IR of the serial identification system for different values of $\mathcal{L}$.

vector is used. In cases, where we have accurately segmented and registered images, the binary feature vector can be directly applied without a loss in biometric performance. On datasets acquired in a less constrained environment we suggest to employ the proposed system as pre-screener in a serial identification system in order to accelerate search operations. We found that the serial identification system even outperforms an exhaustive search of the original system.

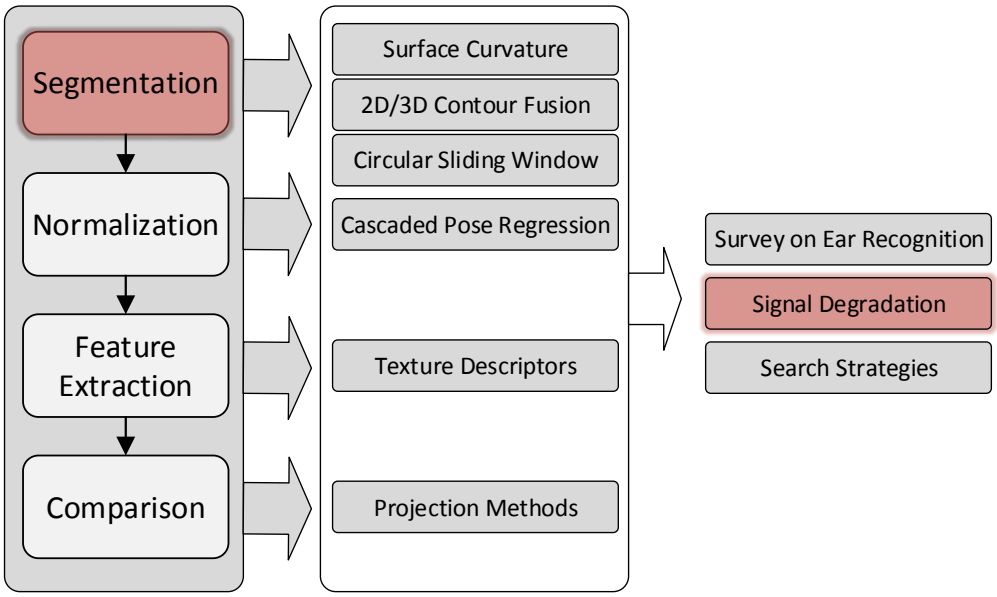
Chapter 10

Effects of Severe Signal Degradation on Ear Detection

Having an ear recognition system that works well under laboratory conditions, the next step is to ask about the performance of system under harder conditions. This brings us to research question **Q5: Which impact does signal degradation have on the performance of ear recognition systems?** In this chapter, we will focus on the ear segmentation step.

Such conditions could for instance be signal degradations, such as noise and blur. In first attempt to evaluate the vulnerability of an ear recognition system to signal degradations, we compared the impact of different blur and noise level on the accuracy of segmentation. As a performance indicator, we report the failure to extract (also referred to as the detection rate n this work).

The paper was published in [186]JOHANNES WAGNER, ANIKA PFLUG, CHRISTIAN RATHGEB, CHRISTOPH BUSCH, Effects of Severe Signal Degradation on Ear Detection, 2nd International Workshop on Biometrics and Forensics (IWBF), 2014



Abstract

Ear recognition has recently gained much attention, as for surveillance scenarios identification remains feasible, in case the facial characteristic is partly or fully covered. However video footage stemming from surveillance cameras is often of low quality. In this work we investigate the impact of signal degradation, i.e. out-of-focus blur and thermal noise, on the segmentation accuracy of automated ear detection. Realistic acquisition scenarios are constructed and various intensities of signal degradation are simulated on a comprehensive dataset. In experiments different ear detection algorithms are employed, pointing out the effects of severe signal degradation on ear segmentation performance.

10.1 Introduction

In April 2013, NYPD detectives were able to track down an arsonist who was accused of burning mezuzahs in Brooklyn, USA. Police got onto the suspect's trail through a clear picture of his ear from a surveillance camera. The image was run through the police facial recognition database, which contained a profile image of the person where the outer ear was visible, returning the suspect's name. Thus investigators reported, *"We had a good angle on his ear that helped to identify him"* [135]. Success stories like this and a constantly increasing number of surveillance cameras underline the potential of automated ear recognition for forensic identification and confirm that the intricate structure of the outer ear represents a reliable and stable biometric characteristic. However, the ability to determine a person's identity based on automated ear recognition highly depends on the image quality and resolution [79, 170]. In addition, images captured by surveillance cameras may suffer from signal degradations particularly, in case of outdoor installations.

Automated ear biometric recognition systems hold tremendous promise for the future, especially in the forensic area [3]. While the long standing success story of ear recognition goes back to the 19th century [26] nowadays forensic applications have only recently started to pay attention to automated ear recognition. In past years numerous approaches focusing on ear detection, feature extraction, and feature comparison have been proposed, achieving promising biometric performance (for a detailed survey on detection and recognition techniques for ears see [147]). However, the vast majority of experimental evaluations are performed on datasets acquired under rather favorable conditions, which in most cases does not reflect image data acquired in forensic scenarios. So far, no studies have been conducted on the impact of signal degradation on automated ear detection, which represents a considerable significant point of failure for any automated ear recognition system.

The contribution of this work is the investigation of the effects of severe signal degradation on automated ear detection. Considering different reasonable scenarios of data acquisition (according to surveillance scenarios), profile images of a comprehensive dataset are systematically degraded, simulating frequent distortions, i.e. out-of-focus blur and thermal noise. In our experiments well-established ear detection algorithms are evaluated, which clearly illustrate the impact of signal degradation on ear detection. Furthermore, a detailed discussion of consequential issues is given.

The remainder of this paper is organized as follows: in Sect. 10.2 considered scenarios and applied signal degradations are described in detail. The effects of signal degradation on ear detection algorithms are investigated in Sect. 10.3. Finally, conclusions are drawn in Sect. 10.4.

10.2 Acquisition and Signal Degradation

10.2.1 Acquisition Scenarios

Table 10.1 summarizes different state-of-the-art surveillance cameras made available by major vendors and relevant characteristics, i.e. focal length, resolution, and sensor type

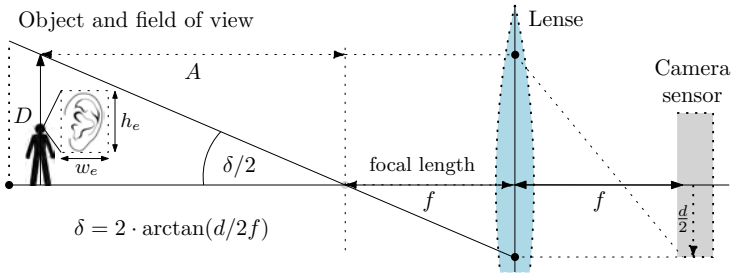


Figure 10.1: Simulated data acquisition scenario.

(characteristics refer to currently best products). Based on this comparison we simulate a camera providing (1) a focal length of 8mm, (2) a resolution of 1920×1080, and (3) a sensor diagonal of 1/2.5 inch. We examine two different acquisition scenarios $\mathcal{S}_1$, $\mathcal{S}_2$ with respect to the distance of the subject to the camera considering distances of 2m and 4m, respectively. Fig. 10.1 schematically depicts the considered acquisition scenario.

We assume that we are able to detect the presence of a subject in a video by one of the state-of-the art detection techniques, that are summarized in [137]. After successfully detecting a subject, the head region can be roughly segmented (cf. Fig. 10.2). These pre-segmented images are the basis of further processing, such as ear detection.

Table 10.1: State-of-the-art camera models and characteristics.

Vendor	**Product**	**Focal length**	**Resolution**	**Sensor**
ACTi[1]	D82	2.8-12mm	1920×1080	1/3.2″
AXIS[2]	P3367V	3-9mm	1920×1080	1/3.2″
GeoVision[3]	GV-FD220G	3-9mm	1920×1080	1/2.5″
Veilux[4]	VVIP-2L2812	2.8-12mm	1920×1080	1/2.5″

[1] http://www.acti.com/
[2] http://www.axis.com/
[3] http://www.geovision.com.tw/
[4] http://www.veilux.net/

Let $C(f, d, w, h)$ be a camera with focal length f, sensor diagonal d, and resolution $w \times h$. Then the diagonal D of the field of view at a distinct distance A is estimated as,

$$\begin{aligned} D = &\ A \cdot \tan\big(2 \cdot \arctan((d/2f)/2)\big) \\ = &\ A \cdot d/2f. \end{aligned} \tag{10.1}$$

In our scenario the aspect ratio is 16:9, i.e. the field of view in object space corresponds to

$$16 \cdot \sqrt{D^2/(16^2+9^2)}\ \text{m} \times 9 \cdot \sqrt{D^2/(16^2+9^2)}\ \text{m}. \tag{10.2}$$

In [173] the average size of the outer ear of males and females across different age group is measured as 61.7 mm ×37.0 mm and 57.8 mm ×34.5 mm, respectively. For an average angle of auricle of 32.5 degrees across age groups and sex we approximate the bounding box of an ear of any subject as 70 mm × 60 mm. For both scenarios $\mathcal{S}_1$, $\mathcal{S}_2$ the considered camera $C(8\text{mm}, 1/2.5'', 1920\text{px}, 1080\text{px})$ would yield images where ear regions comprise approximately $w_e \times h_e = 110{\times}90$ and $55{\times}45$ pixels, respectively.

Table 10.2: Blur and noise conditions considered for signal degradation (denotations of σ are defined in 10.2.2.1 and 10.2.2.2).

Blur condition		Noise condition		Intensity
Abbrev.	Description	Abbrev.	Description	
B-0	–	N-0	–	none
B-1	$\sigma = 2$	N-1	$\sigma = 20$	low
B-2	$\sigma = 3$	N-2	$\sigma = 25$	medium
B-3	$\sigma = 4$	N-3	$\sigma = 30$	high

10.2.2 Signal Degradation

Signal degradation in this work is simulated by means of blur and noise where blur is applied prior to noise (out-of-focus blur is caused before noise occurs). Four different intensities (including absence) of blur and noise and combinations of these are considered and summarized in Table 10.2.

10.2.2.1 Blur Conditions

Out-of-focus blur represents a frequent distortion in image acquisition mainly caused by an inappropriate distance of the camera to the eye (another type of blur is motion blur caused by rapid movement which is not considered in this work). We simulate the point spread function of the blur as a Gaussian

$$f(x,y) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\pi\sigma^2}} \tag{10.3}$$

which is then convolved with the specific image, where the image is devided into 16×16 pixel blocks.

10.2.2.2 Noise Conditions

Amplifier noise is primarily caused by thermal noise. Due to signal amplification in dark (or underexposed) areas of an image, thermal noise has a high impact on these areas. Additional sources contribute to the noise in a digital image such as shot noise, quantization noise and others. These additional noise sources however, only make up a negligible part of the noise and are therefore ignored during this work.

Let P be the set of all pixels in image $I \in \mathbb{N}^2$, $w = (w_p)_{p \in P}$, be a collection of independent identically distributed real-valued random variables following a Gaussian distribution with mean m and variance σ^2. We simulate thermal noise as additive Gaussian noise with $m = 0$, variance σ^2 for pixel p at x, y as

$$N(x,y) = I(x,y) + w_p, p \in P, \tag{10.4}$$

with N being the noisy image, for an original image I. Examples of results of simulated signal degradation are depicted in Fig. 10.2 for images considered in both scenarios.

10.3 Experimental Evaluations

10.3.1 Experimental Setup

For our evaluation, we have composed a dataset of mutually different images of the UND-G [202], UND-J2 [201] and UND-NDOff-2007 [64] database. The merged dataset contains 2369 left profile images from 510 subjects with yaw poses between -60 and -90 degrees. Right profile views from UND-G were mirrored horizontally. The manually annotated

$\mathcal{S}_1$ B-0 | N-0 $\mathcal{S}_1$ B-3 | N-0 $\mathcal{S}_1$ B-0 | N-3 $\mathcal{S}_1$ B-3 | N-3

$\mathcal{S}_2$ B-0 | N-0 $\mathcal{S}_2$ B-3 | N-0 $\mathcal{S}_2$ B-0 | N-3 $\mathcal{S}_2$ B-3 | N-3

Figure 10.2: Maximum intensities of blur and/ or noise for each of the two scenarios. The upper row shows examples from $\mathcal{S}_1$ and the lower row from $\mathcal{S}_2$ respectively (`id_02463d677`).

vspace-0.3cm

ground truth in form of ear bounding boxes yields an average size of 125×95 pixels for the entire data set, i.e. original images are employed in scenario $\mathcal{S}_1$. For the second scenarios $\mathcal{S}_2$ images are scaled with factor 0.5 prior to applying blur and noise.

We evaluate the performance of cascaded object detectors with a fixed-size sliding window. The object detectors are trained with (1) Haar-like features [182], (2) local binary patterns [138] (LBP) and (3) histograms of oriented gradients [57](HOG). The detectors were trained with images from WPUTEDB [66] and negative samples from the INIRA person detection dataset.

10.3.2 Performance Evaluation

We calculate the detection error E from the segmentation result $S(I)$ for an image I with corresponding ground truth mask G_I (both of dimension $m \times n$ pixels), such that for all positions x, y, $G_I[x, y] = 1$ labels "ear" pixels (otherwise $G_I[x, y] = 0$), as

$$E = \frac{1}{m \cdot n} \sum_{x=0}^{m-1} \sum_{y=0}^{n-1} G_I[x, y] \oplus S(I)[x, y]. \tag{10.5}$$

Table 10.3 summarizes the detection errors for different detection algorithms for intensities of blur, noise and combination of these for both considered scenarios. The quality of generated images is estimated in terms of average peak signal to noise ratio (PSNR). Errors E_1, E_2 and E_3 correspond to the detection results employing Haar-like, LBP, and HOG features, respectively. In Fig. 10.3 detection errors are plotted for all detection algorithms and scenarios for all combinations of signal degradation. See Table 10.4 for a collection of examples for ground truth and detection under different conditions.

10.3.3 Discussion

As can be seen in Fig. 10.3 and Table 10.3, Haar-like features turn out to be most robust against noise, followed by LBP where we observe slightly higher error rate. Haar-like features rely on the position of edges and are using the ratio of dark and light pixels in an

Table 10.3: Error rates for different detection algorithms for both scenarios (errors have been multiplied by 10^2). Results are visualized in Fig 10.3.

		Scenario $\mathcal{S}_1$				Scenario $\mathcal{S}_2$			
Blur	**Noise**	**PSNR**	E_1	E_2	E_3	**PSNR**	E_1	E_2	E_3
B-0	N-0	∞	2.12	2.86	1.94	∞	2.64	2.75	1.97
B-1	N-0	33.69 db	2.62	2.91	2.10	31.66 db	3.77	3.50	5.56
B-2	N-0	32.18 db	2.89	3.12	2.88	29.58 db	4.18	3.90	6.01
B-3	N-0	31.20 db	3.22	3.24	3.42	28.38 db	4.30	3.87	6.08
B-0	N-1	22.56 db	2.08	2.77	3.45	22.56 db	2.09	3.23	3.18
B-1	N-1	22.21 db	2.09	2.82	3.76	22.01 db	2.34	3.37	5.51
B-2	N-1	22.07 db	2.09	2.59	4.02	21.71 db	2.53	3.41	5.94
B-3	N-1	21.96 db	2.16	2.77	4.00	21.48 db	2.61	3.58	6.12
B-0	N-2	20.74 db	2.19	2.70	3.75	20.74 db	2.18	3.20	3.73
B-1	N-2	20.50 db	2.17	2.80	4.00	20.35 db	2.39	3.40	5.30
B-2	N-2	20.40 db	2.17	2.59	4.14	20.14 db	2.56	3.52	5.60
B-3	N-2	20.32 db	2.22	2.77	4.14	19.97 db	2.66	3.59	5.29
B-0	N-3	19.27 db	2.22	2.82	3.95	19.26 db	2.27	3.40	4.04
B-1	N-3	19.09 db	2.25	2.68	4.13	18.98 db	2.40	3.52	4.86
B-2	N-3	19.02 db	2.27	2.57	4.18	18.82 db	2.59	3.50	4.87
B-3	N-3	18.96 db	2.33	2.84	4.18	18.69 db	2.77	3.74	4.82

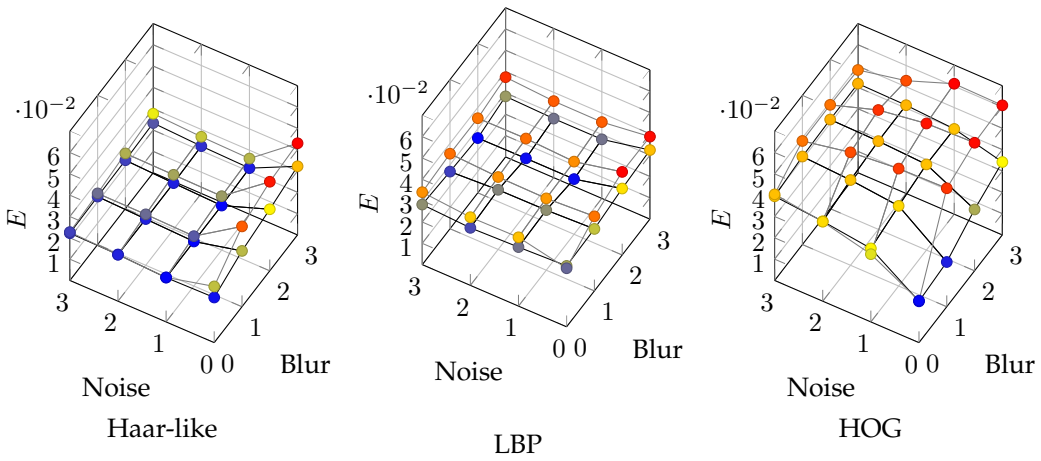


Figure 10.3: Errors for different segmentation algorithms for intensities of blur, noise and combination of these for both scenarios. The black mesh shows the performance for $\mathcal{S}_1$ and the grey mesh for $\mathcal{S}_2$, respectively.

image patch. This makes these features robust to noise but vulnerable to blur. Combinations between blur and noise perform better, because adding noise after blur results in images with more intense edges.

For LBP, we obtain a mostly stable, although higher, error rate than for Haar-like features. Generally, degradations have the least impact in LBP, because it encodes local texture information as a histogram without the need of particular local features, such as edges. Alterations on a pixel level add noise uniformly for all local features to a mostly skin colored texture, which can be compensated by the detector.

HOG performs well under ideal conditions, but with increasing noise and blur, the accuracy degenerates quickly. Blur and noise alter the local gradient orientation, length and direction of the image, which makes it difficult for HOG to match the trained pattern with local texture information. Although noise and blur are causing this effect, blur has a significantly larger impact on local gradients than noise.

10.4 Conclusion

We have quantified the impact of signal degradation in particular, noise and blur, on ear detection systems for surveillance purposes. Experiments were carried out for three well-established detection algorithms, Haar-like features, LBPs and HOG. With respect to the simulated conditions, the tested algorithms turn out to be vulnerable to both, to blur and noise. Our future work will comprise research on other forms of signal degradations as well as the impact of signal degradation on feature extraction and recognition performance.

Table 10.4: Detection accuracy of different feature sets with strong noise and blur. The left column shows the ground truth, the middle column shows the detection result in the original image and the right column shows the result after signal degradation.

		
(a) Ground Truth	(b) B-0 N-0 — Haar Detection	(c) B-0 N-3 — Haar Detection
		
(a) Ground Truth	(b) B-0 N-0 — Haar Detection	(c) B-0 N-1 — Haar Detection
		
(a) Ground Truth	(b) B-0 N-0 — HOG Detection	(c) B-0 N-3 — HOG Detection
		
(a) Ground Truth	(b) B-0 N-0 — HOG Detection	(c) B-3 N-0 — HOG Detection
		
(a) Ground Truth	(b) B-0 N-0 — LBP Detection	(c) B-3 N-0 — LBP Detection
		
(a) Ground Truth	(b) B-0 N-0 — LBP Detection	(c) B-3 N-3 — LBP Detection

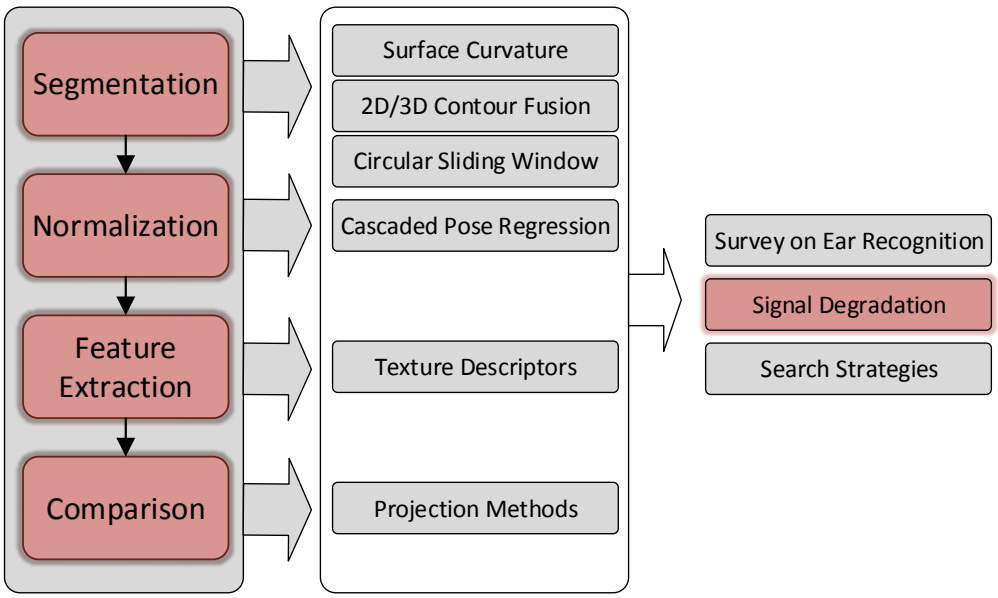
Chapter 11

Impact of Severe Signal Degradation on Ear Recognition Performance

This chapter represents the second part of the answer to research question **Q5: Which impact does signal degradation have on the performance of ear recognition systems?**

Directly based on the previous work on ear detection, we also evaluate the impact of different blur and noise levels on the biometric performance of an ear recognition system that is using histogram-based texture features. In our experiments on the recognition performance, we use manually cropped ROIs, such that we can assume that the failure to extract is zero. We measure the Equal error rate and the identification rate (rank-1 recognition rate) as correlate the recognition performance with the image quality.

The paper was published in [153]ANIKA PFLUG, JOHANNES WAGNER, CHRISTIAN RATHGEB AND CHRISTOPH BUSCH, Impact of Severe Signal Degradation on Ear Recognition Performance, Biometrics, Forensics, De-identification and Privacy Protection (BiForD), 2014



Abstract

We investigate ear recognition systems for severe signal degradation of ear images in order to assess the impact on biometric performance of diverse well-established feature extraction algorithms. Various intensities of signal degradation, i.e. out-of-focus blur and thermal noise, are simulated in order to construct realistic acquisition scenarios. Experimental evaluations, which are carried out on a comprehensive database comprising more than 2,000 ear images, point out the effects of severe signal degradation on ear recognition performance using appearance features.

11.1 Introduction

Following the first studies on forensic evidence of ear images of A. Iannarelli in 1989 [81], Hoogstrate *et al.* presented a study on the evidential value of ear images from CCTV footage [79]. This work was motivated by a series of gas station robberies in Utrecht, Netherlands. During the incidents, the perpetrators appeared in several CCTV videos, however their faces were occluded by baseball hats in all of the videos. In all of the cases, the CCTV videos contained several frames with profile views, where the outer ear of one of the perpetrators was clearly visible. Hoogstrate *et al.* showed that these ear images can be employed for identification by a forensic expert, if the quality of the videos is sufficient. Such impairing factors for the image quality in surveillance videos can be, for instance, blur and thermal sensor noise. Image quality plays an even more important role when potential candidates should be pre-selected automatically by a biometric system, prior to manual inspection by a forensic expert.

Automated ear biometric recognition systems hold tremendous promise for the future, especially in the forensic area [3]. While the long standing success story of ear recognition goes back to the 19th century [26] nowadays forensic applications have only recently started to pay attention to automated ear recognition. In past years numerous approaches focusing on ear detection, feature extraction, and feature comparison have been proposed, achieving promising biometric performance (for a detailed survey see [147]). However, the vast majority of experimental evaluations are performed on datasets acquired under rather favorable conditions, which in most cases does not reflect image data acquired in forensic scenarios. So far, no studies have been conducted on the impact of signal degradation on automated ear recognition, which represents a considerable significant point of failure for any automated ear recognition system.

The contribution of this work is the investigation of the effects of severe signal degradation on automated ear recognition using appearance features. Considering different reasonable scenarios of data acquisition (according to surveillance scenarios), ear images of a comprehensive dataset are systematically degraded, simulating frequent distortions, i.e. out-of-focus blur and thermal noise. On the one hand, a synthetic degradation of ear images allows a comprehensive experimental evaluation of existing dataset and, on the other hand, it is feasible to measure and reproduce the source of image degradation. In previous work [186] we have shown that state-of-the-art ear detection algorithms are capable of overcoming simulated signal degradations caused by out-of-focus blur and thermal noise. In this work emphasis is put on recognition performance, i.e. the impact of signal degradation on biometric performance is analyzed and a detailed discussion of consequential issues is given.

The remainder of this paper is organized as follows: Sect. 11.2 considered scenarios and applied signal degradations are described in detail. The effects of signal degradation on ear recognition algorithms are investigated in Sect. 11.3. Finally, conclusions are drawn in Sect. 11.4.

Table 11.1: State-of-the-art camera models and characteristics.

Vendor	Product	Focal length	Resolution	Sensor
ACTi	D82	2.8-12mm	1920×1080	1/3.2″
AXIS	P3367V	3-9mm	1920×1080	1/3.2″
GeoVision	GV-FD220G	3-9mm	1920×1080	1/2.5″
Veliux	VVIP-2L2812	2.8-12mm	1920×1080	1/2.5″

http://www.acti.com/
http://www.axis.com/
http://www.geovision.com.tw/
http://www.veliux.net/

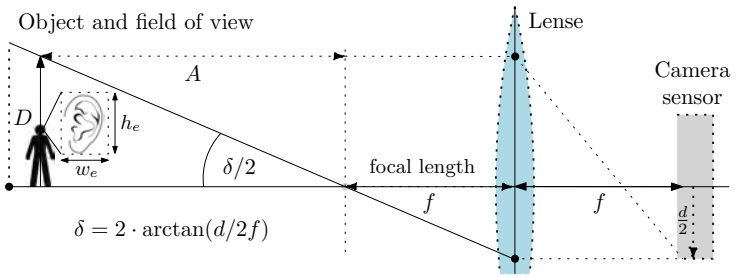


Figure 11.1: Simulated data acquisition scenario.

11.2 Acquisition and Signal Degradation

11.2.1 Acquisition Scenarios

Table 11.1 summarizes diverse state-of-the-art surveillance cameras made available by major vendors and relevant characteristics, i.e. focal length, resolution, and sensor type (characteristics refer to most developed products of according vendors). Based on this comparison we consider a camera providing (1) a focal length of 8mm, (2) a resolution of 1920×1080, and (3) a sensor diagonal of 1/2.5 inch. We examine two different acquisition scenarios $\mathcal{S}_1$, $\mathcal{S}_2$ with respect to the distance of the subject to the camera considering distances of approximately 2m and 4m, respectively. Fig. 11.1 schematically depicts the considered acquisition scenario.

We assume that we are able to detect the presence of a subject in a video by one of the state-of-the art detection techniques, that are summarized in [137]. After successfully detecting a capture subject, the head region can be roughly segmented. In [186] we have demonstrated that cascaded ear detectors, e.g. based on Haar-like features, output stable detection results even in presence of severe signal degradation. In order to estimate the mere effect of signal degradation on feature extraction and classification modules we restrict ourselves to the analysis of cropped images of size 165×92 and 83×46 pixels. These cropped images are generated on the basis of a manually segmented ground-truth, in both scenarios, respectively (cf. Fig.11.2).

Let $C(f, d, w, h)$ be a camera with focal length f, sensor diagonal d, and resolution $w \times h$. Then the diagonal D of the field of view at a distinct distance A is estimated as,

$$\begin{aligned} D &= A \cdot \tan\big(2 \cdot \arctan((d/2f)/2)\big) \\ &= A \cdot d/2f. \end{aligned} \tag{11.1}$$

In our scenario the aspect ratio is 16:9, i.e. the field of view in object space corresponds to

$$16 \cdot \sqrt{D^2/(16^2+9^2)}\ \text{m} \times 9 \cdot \sqrt{D^2/(16^2+9^2)}\ \text{m}. \tag{11.2}$$

Table 11.2: Blur and noise conditions considered for signal degradation (denotations of σ are defined in 11.2.2.1 and 11.2.2.2).

Blur condition		Noise condition		Degradation
Abbrev.	**Description**	**Abbrev.**	**Description**	**Intensity**
B-0	–	N-0	–	none
B-1	$\sigma = 2$	N-1	$\sigma = 20$	low
B-2	$\sigma = 3$	N-2	$\sigma = 25$	medium
B-3	$\sigma = 4$	N-3	$\sigma = 30$	high

In [173] the average size of the outer ear of males and females across different age group is measured as 61.7 mm ×37.0 mm and 57.8 mm ×34.5 mm, respectively. For an average angle of auricle of 32.5 degrees across age groups and sex we approximate the bounding box of an ear of any subject as 70 mm × 60 mm. For both scenarios $\mathcal{S}_1$, $\mathcal{S}_2$ the considered camera $C(8\text{mm}, 1/2.5'', 1920\text{px}, 1080\text{px})$ would yield images where ear regions comprise approximately $w_e \times h_e = 110{\times}90$ and $55{\times}45$ pixels, respectively.

11.2.2 Signal Degradation

Signal degradation in this work is simulated by means of blur and noise where blur is applied prior to noise (out-of-focus blur is caused before noise occurs). Four different intensities (including absence) of blur and noise and combinations of these are considered and summarized in Table 11.2.

11.2.2.1 Blur Conditions

Out-of-focus blur represents a frequent distortion in image acquisition mainly caused by an inappropriate distance of the camera to the eye (another type of blur is motion blur caused by rapid movement which is not considered in this work). We simulate the point spread function of the blur as a Gaussian

$$f(x,y) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\pi\sigma^2}} \tag{11.3}$$

which is then convoluted with the specific image, where the image is devided into 16×16 pixel blocks.

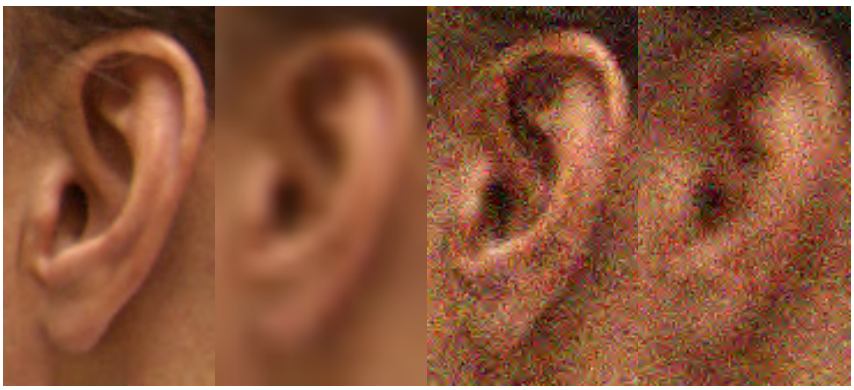
11.2.2.2 Noise Conditions

Amplifier noise is primarily caused by thermal noise. Due to signal amplification in dark (or underexposed) areas of an image, thermal noise has a high impact on these areas. Additional sources contribute to the noise in a digital image such as shot noise, quantization noise and others. These additional noise sources however, only make up a negligible part of the noise and are therefore ignored during this work.

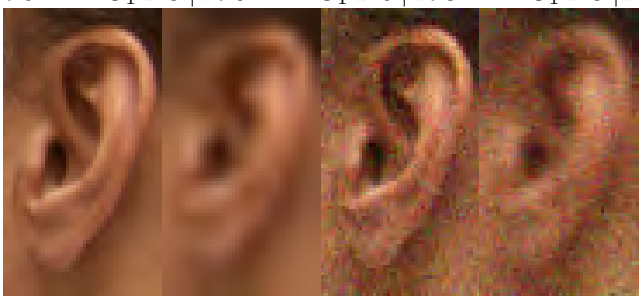
Let P be the set of all pixels in image $I \in \mathbb{N}^2$, $w = (w_p)_{p \in P}$, be a collection of independent identically distributed real-valued random variables following a Gaussian distribution with mean m and variance σ^2. We simulate thermal noise as additive Gaussian noise with $m = 0$, variance σ^2 for pixel p at x, y as

$$N(x,y) = I(x,y) + w_p, p \in P, \tag{11.4}$$

with N being the noisy image, for an original image I. Examples of results of simulated signal degradation are depicted in Fig. 11.2 for a single image considered in both scenarios.



$\mathcal{S}_1$ B-0 | N-0 $\mathcal{S}_1$ B-3 | N-0 $\mathcal{S}_1$ B-0 | N-3 $\mathcal{S}_1$ B-3 | N-3



$\mathcal{S}_2$ B-0 | N-0 $\mathcal{S}_2$ B-3 | N-0 $\mathcal{S}_2$ B-0 | N-3 $\mathcal{S}_2$ B-3 | N-3

Figure 11.2: Maximum intensities of blur and/ or noise for the two scenarios. The upper row shows examples from $\mathcal{S}_1$ and the lower row from $\mathcal{S}_2$ respectively (`id_02463d677`).

11.3 Experimental Evaluations

11.3.1 Experimental Setup

For our evaluation, we have composed a dataset of mutually different images of the UND-G [202], UND-J2 [201] and UND-NDOff-2007 [64] database. The merged dataset contains 2111 left profile images from 510 subjects with yaw poses between -60 and -90 degrees. The manually annotated ground truth in form of ear bounding boxes yields an average size of 125 $\times$ 95 pixels for the entire data set. Based on these ear bounding boxes images are cropped (based on the center of boxes) to images of 165$\times$92 pixels which are employed in scenario $\mathcal{S}_1$. For the second scenario $\mathcal{S}_2$ cropped images are scaled with factor 0.5 prior to adding blur and noise. Prior to extracting features, we apply CLAHE [223] to normalize the image contrast. In order to find the optimal settings for each of the feature extraction methods, we compared different parameter settings for each of the feature extraction techniques. We only give the results for the best performing parameter settings.

In our experiments, we randomly select four images of each subject for training purposes and one image for testing. Hence, our setup requires that we have at least five samples per subject, which, however, is not the case for all the subjects in the database. Our total test set consists of 132 probes from 132 different subjects. The training set contains 528 images of the same 132 subjects. All performance indicators reported in this work are median values based on a five-fold cross validation.

Performance is estimated in terms of equal error rate (EER) and (true-positive) identification rate (IR). In accordance to the ISO/IEC IS 19795-1 [92] the false non-match rate (FNMR) of a biometric system defines the proportion of genuine attempt samples falsely

declared not to match the template of the same characteristic from the same user supplying the sample. By analogy, the false match rate (FMR) defines the proportion of zero-effort impostor attempt samples falsely declared to match the compared non-self template. As score distributions overlap EERs are obtained, i.e. the system error rate where FNMR = FMR. The IR is the proportion of identification transactions by subjects enrolled in the system in which the subject's correct identifier is the one returned. In experiments identification is performed in the closed-set scenario returning the rank-1 candidate as identified subject (without applying a decision threshold).

11.3.2 Feature Extraction and Classification

11.3.2.1 Local Binary Patterns

Local Binary Patterns (LBP) represent a widely used texture descriptor that has been applied for various biometric characteristics, and recently was also used in an ear recognition system [3]. LBP encodes local texture information on a pixel level by comparing the grey level values of a pixel to the grey level values in its n-8 neighborhood. Every pixel with a grey level value exceeding the grey level of the central pixel is assigned the binary value 1, whereas all pixels with a smaller grey level value are assigned the binary value 0. Subsequently, the binary image information is extracted by concatenating these binary values according to a certain predefined scheme. This results in a binary-valued local descriptor for a particular image patch. This concept can also be extended to any other definitions of a local neighborhood, in particular to different radii around the center pixel.

We extract LBP features from the n-8 neighborhood of each pixel in the image. We divide the image into a regular grid of 10×10 pixels and concatenate the local LBP histogram within each grid cell.

11.3.2.2 Local Phase Quantization

Local Phase Quantization (LPQ) is designed to be robust against Gaussian blur, by exploiting the blur invariance property of the Fourier phase spectrum [139]. It could be shown in [10], that LPQ is superior to LBP for face recognition, if the image is degraded with Gaussian blur.

Within LPQ the image is transformed into the Fourier domain, where the signal can be split into the magnitude and the phase part. Then the phase angles are estimated and transformed into a 2-bits code word by using a quantization function. This procedure is repeated for all points within a specified radius. All code words within the given radius are then put into a histogram, which represents the local phase information on an image patch. In this paper we extract local LPQ histograms from a regular grid with 11×11 pixels cells and concatenate each of the local histograms to obtain the global feature vector.

11.3.2.3 Histograms of Oriented Gradients

Originally introduced as a descriptor for person detection in 2005, Histograms of Oriented Gradients (HOG) soon became a popular texture feature in other fields of computer vision, too [57]. HOG uses the local gradient direction in a particular image patch and then concatenates this information to local histograms that reflect the distribution of gradient directions of a particular object in the image. Each of the local histograms is normalized before all of the histograms are concatenated to form the complete descriptor. The HOG descriptor in our experiment is using a patch size of 8×8 pixels with 9 different orientations.

11.3.2.4 Projection and Classification

Employed appearance features described above usually have a large number of dimensions. For creating our feature space, we compute a projection matrix based on our train-

Table 11.3: Equal error rates (EER) and true-positive identification rates (IR) for different algorithms with different blur and noise settings in $\mathcal{S}_1$.

Blur	Noise	Scenario $\mathcal{S}_1$	LBP		LPQ		HOG	
		PSNR	EER	IR	EER	IR	EER	IR
B-0	N-0	∞	6.37	82.57	1.64	92.42	5.91	87.87
B-1	N-0	32.51 db	6.06	87.12	6.51	83.33	5.90	81.81
B-2	N-0	30.45 db	2.88	90.15	1.32	93.93	4.72	87.12
B-3	N-0	29.19 db	5.00	84.09	4.24	84.84	8.12	75.75
B-0	N-1	22.81 db	9.40	68.18	6.66	75.75	16.21	53.78
B-1	N-1	22.31 db	15.29	50.00	8.64	64.39	19.47	35.60
B-2	N-1	22.03 db	16.53	43.18	12.39	46.96	25.42	21.96
B-3	N-1	21.80 db	18.04	38.63	16.94	44.69	26.20	18.18
B-0	N-2	21.01 db	12.58	56.06	10.13	63.63	19.55	41.66
B-1	N-2	20.66 db	17.37	40.90	14.26	44.69	26.20	26.51
B-2	N-2	20.46 db	19.54	37.87	15.48	41.66	30.15	12.87
B-3	N-2	20.30 db	21.05	28.78	17.85	39.39	30.79	10.60
B-0	N-3	19.56 db	14.68	50.00	10.45	70.45	24.85	26.51
B-1	N-3	19.30 db	20.76	33.33	15.70	37.87	28.81	12.87
B-2	N-3	19.15 db	25.90	25.00	21.97	28.03	31.37	11.36
B-3	N-3	19.02 db	23.93	21.96	22.53	28.78	37.57	6.81

Table 11.4: Equal error rates (EER) and true-positive identification rates (IR) for different algorithms with different blur and noise settings in $\mathcal{S}_2$.

Blur	Noise	Scenario $\mathcal{S}_2$	LBP		LPQ		HOG	
		PSNR	EER	IR	EER	IR	EER	IR
B-0	N-0	∞	2.29	92.98	4.85	83.33	5.91	81.81
B-1	N-0	34.77 db	6.36	77.27	1.25	95.45	7.49	76.69
B-2	N-0	31.87 db	6.06	78.78	2.95	89.23	8.78	67.42
B-3	N-0	30.27 db	6.96	75.75	5.48	81.81	9.73	65.15
B-0	N-1	30.09 db	12.44	58.33	3.78	87.82	16.21	44.69
B-1	N-1	28.68 db	15.34	40.90	5.75	78.03	21.39	28.03
B-2	N-1	27.69 db	18.79	33.33	9.71	62.87	25.42	17.99
B-3	N-1	26.97 db	23.31	31.81	8.48	65.15	26.20	18.18
B-0	N-2	28.21 db	13.55	45.45	3.79	82.57	19.55	33.33
B-1	N-2	27.22 db	21.22	33.33	8.05	65.90	26.20	24.24
B-2	N-2	26.48 db	23.93	20.45	10.16	54.54	29.23	12.87
B-3	N-2	25.91 db	25.65	17.42	12.85	51.51	30.79	14.39
B-0	N-3	26.68 db	17.85	37.12	6.36	80.30	24.85	21.96
B-1	N-3	25.94 db	24.56	17.42	12.24	53.78	28.81	13.63
B-2	N-3	25.36 db	27.42	18.18	16.06	44.69	31.37	10.60
B-3	N-3	24.91 db	29.72	15.90	14.35	41.66	31.80	9.09

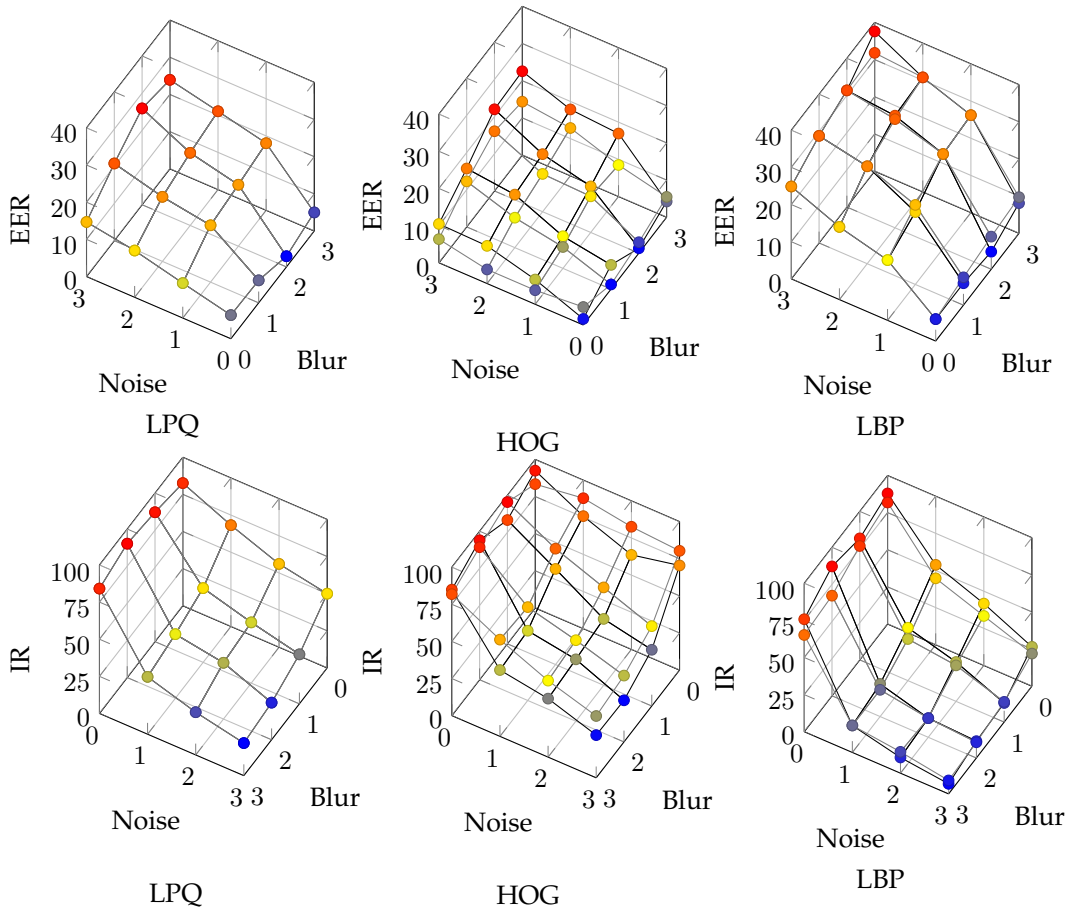


Figure 11.3: EER (upper row) and IR lower row) rates for different algorithms for intensities of blur, noise and combination of these. The black mesh shows the performance for $\mathcal{S}_1$ and the grey mesh for $\mathcal{S}_2$, respectively.

ing data by using LDA. After computing the feature space from the training images, we project the test images into the same feature space. Subsequently, we assign an identity based on a NN-classifier and cosine distance. The source code for feature space projection and classification is based on the PhD face recognition toolbox [184].

11.3.3 Performance Evaluation

Tables 11.3 and 11.4 summarize the biometric performance with respect to EERs and IRs for different feature extraction algorithms for intensities of blur, noise and combination of these for both considered scenarios. The quality of generated images is estimated in terms of average peak signal to noise ratio (PSNR). Fig. 11.3.3 illustrated the change of biometric performance according to the simulated intensities of blur and noise.

11.3.4 Discussion

The general expectation in this experimental setup is, that the appearance of all images converges towards an average ear shape, the more noise and blur are added to the image. Blur is removing details, whereas noise is virtually adding random information to the image signal.

The recognition performance of Scenario $\mathcal{S}_1$ and $\mathcal{S}_2$ only differs significantly at some points. In general, the pipelines in $\mathcal{S}_2$ perform slightly better than in $\mathcal{S}_1$. From this we may conclude that automatic ear recognition is yielding good results in scenarios with large distances to the camera and with low resolution.

For all tested pipelines, the sole presence of blur only slightly degrades the recognition performance. Thermal noise however, has a significant impact on the recognition accuracy of all of the features. However, when blur is combined with noise, the two types of degradation amplify each other, which results in low recognition performance for all of the features.

The best performing algorithm in our experiments is LPQ. It turns out to be relatively resilient against slight presence of noise and blur, as well as against combinations of these. However, even though LPQ was designed to be a blur invariant descriptor, in practice it is not entirely robust against blur. This can be explained by the fact the descriptor is only invariant to blur, if the window size of the descriptor is unlimited [10]. This means that the larger the window for feature extraction, the more robust LPQ becomes against blur. However, with increasing window size we also lose the locality of information and become more vulnerable to occlusions. Smaller ROIs slightly improve the recognition performance of LPQ, which is due to the fact that the window size was constant in $\mathcal{S}_1$ and $\mathcal{S}_2$. Hence, the local window covers a larger portion of the ROI, which means that the resilience against blur is higher.

As pointed out earlier, LBP is exclusively relying on small image patches around given pixels. Hence, this descriptor is vulnerable against both types of signal degradation, noise and blur. Whereas blur removes high frequencies from the image, it retains the relative grey level value in homogeneous regions of the image. This is why LBP still performs reasonably well on blurred images. Noise, however changes the grey level values randomly at different spots in the image, which has a direct impact on the local LBP histogram values and, hence, results into a more severe performance decrease. Combinations of noise and blur destroy the local pixel information by introducing false patterns in homogeneous patches and dithers patches that were formerly containing edges, which lets the performance of LBP drop significantly.

In order to create a distinctive feature vector, the HOG descriptor needs a sufficient number of edges that are representing the object. As edge information is gradually removed by blur and dithered by noise, HOG is affected by both types of degradations. Blur alone, however only changes the length of the local gradients, but not the directions, which is why blur can be handled well by HOG. Adding additional noise changes the local gradient direction and hence alters the feature vector, which is reflected by the performance drop at the maximum amount of blur and noise. This behavior is observed for both scenarios.

11.4 Conclusion

In this work, we have investigated the impact of two different types of signal degradation on the recognition performance of well-established appearance features. Based on publicly available data, we have added noise and blur to the images to create a controlled environment, such that we can draw conclusions about which factor has the largest impact on the recognition performance.

Experiments show that LBP, HOG and LPQ are relatively robust, although not invariant to blur. Noise has a larger effect on the recognition performance compared to blur. Combinations of noise and blur amplify each other, such that the performance drops significantly, when they occur together. The size of the ROI only has a minor effect on the recognition performance, which lets us conclude that the outer ear can still be captured with sufficient resolution from large distances.

In future work, we will focus on the impact of other kinds of signal degradation on the detection accuracy as well as on the recognition performance.

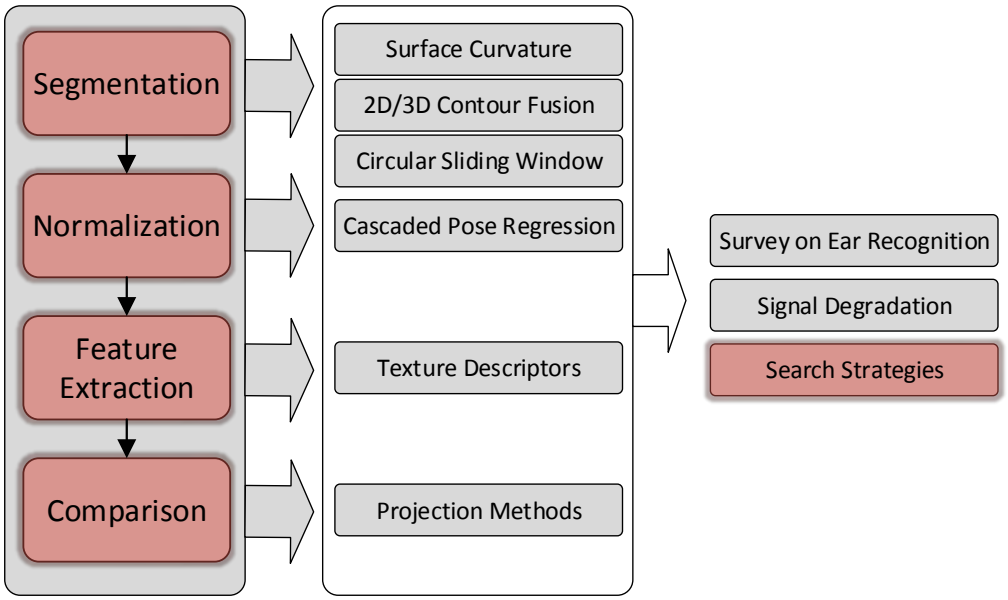
Chapter 12

2D Ear Classification Based on Unsupervised Clustering

Knowing that we can achieve a high performance with histogram-based texture descriptors, we address the question whether we can find clusters within these feature spaces that represent certain categories of ears. Assuming that the assignment of shape classes as proposed by Iannarelli [81] can be arbitrary, we try to the answer of research question **Q6: Is it possible to automatically find categories of ear images?**

Classification can be useful for performing search operations in large databases with limited bandwidth or for selecting similar candidates form a number of suspects for further manual analysis. As opposed to the shape-based categorization of Iannarelli, we automatically find and assign categories within the feature space without a prior shape extraction step.

The paper was published in [152]ANIKA PFLUG, ARUN ROSS, CHRISTOPH BUSCH, 2D Ear Classification Based on Unsupervised Clustering, In Proceedings of International Joint Conference on Biometrics (IJCB), 2014



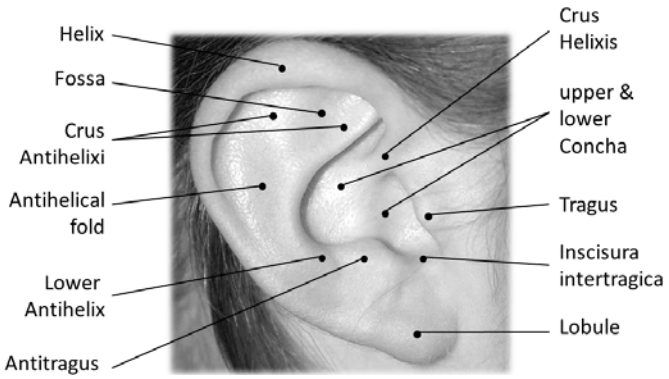


Figure 12.1: Morphology of the outer ear.

Abstract

Ear classification refers to the process by which an input ear image is assigned to one of several pre-defined classes based on a set of features extracted from the image. In the context of large-scale ear identification, where the input probe image has to be compared against a large set of gallery images in order to locate a matching identity, classification can be used to restrict the matching process to only those images in the gallery that belong to the same class as the probe. In this work, we utilize an unsupervised clustering scheme to partition ear images into multiple classes (i.e., clusters), with each class being denoted by a prototype or a centroid. A given ear image is assigned class labels (i.e., cluster indices) that correspond to the clusters whose centroids are closest to it. We compare the classification performance of three different texture descriptors, viz. Histograms of Oriented Gradients, uniform Local Binary Patterns and Local Phase Quantization. Extensive experiments using three different ear datasets suggest that the Local Phase Quantization texture descriptor scheme along with PCA for dimensionality reduction results in a 96.89% hit rate (i.e., 3.11% pre-selection error rate) with a penetration rate of 32.08%. Further, we demonstrate that the hit rate improves to 99.01% with a penetration rate of 47.10% when a multi-cluster search strategy is employed.

12.1 Introduction

Classification involves assigning a class label to a subject based on features extracted from the subject's biometric data. The number of classes is usually much smaller than the number of subjects in the gallery database and each subject is typically assigned to exactly one class. Class labels can either be based on anatomical properties of the observed biometric characteristic or on inherent structural or geometric properties of the biometric sample.While classification and/or indexing techniques have been developed for fingerprints [121, 167], iris [130, 74] and face [181, 145], the possibility of classifying ear images has received limited attention in the biometrics literature. To the best of our knowledge, this is the first work on automated unsupervised classification of ear images.

In this work, we explore the possibility of clustering 2D ear patterns into multiple categories based on their texture and structure. The texture is captured by the use of a texture descriptor, while local histograms capture the structure of the ear. We used *texture-based* features rather than explicit *shape-based* features because (a) extracting *shape* information from 2D ear images is a challenging problem, that often requires a highly constrained capture scenario [6], and (b) the discriminability of shape-based features is limited in low-quality 2D images.

One of the earliest work on ear classification was done by Ianerelli [81], where he classified ear images into 4 categories - round, oval, triangular, rectangular - based on a *visual*

assessment of the ear. However, this classification process is difficult to automate due to the subjective nature of the assessment process. Further, as noted in [68], the number of members in each class is unevenly distributed.

In previous work, Khorsandi and Abdel-Mottaleb [100] categorized ear images into two groups - male and female - based on Gabor Filters and Sparse Representation. Motivated by their work, we aim to further explore the capability of texture descriptors for ear classification. In this regard we analyze commonly used texture descriptors, viz. Histograms of Oriented Gradients (HOG) [57], unified Local Binary Patterns (uLBP)[138] and Local Phase Quantization (LPQ)[10]. LBP and HOG have already been successfully used in the context of ear *recognition* [32, 72, 58]. LPQ has recently been used in face recognition, where it is shown to be more robust to blur than LBP [10]. For an elaborate survey of features used for ear recognition, we refer the reader to [6].

In this work, we use unsupervised clustering in conjunction with texture-based local histograms to discover classes of ear patterns. Instead of using pre-defined labels such as triangular, oval, etc. [81], we deduce clusters based on the distribution of texture features in a high-dimensional space. Although this approach may not result in classes that can be trivially interpreted by a human, it allows us to circumvent ambiguities in class label assignment and results in classes with more evenly distributed numbers of members. The extraction of shape features can be complex and time consuming. Our goal is to use simple features that can be generated quickly and that do not bear the risk of error in the feature extraction process.

The primary contributions of this work are (a) an analysis of the clustering tendencies of feature spaces corresponding to 3 different texture descriptors; (b) a detailed experimental evaluation demonstrating the benefits of the proposed clustering approach for ear classification; and (c) a method for fusing the outputs of multiple classification schemes.

12.2 Clustering 2D Ear Patterns

The proposed approach has two distinct phases: the training phase and the classification phase. The *training phase* has two stages: (a) feature subspace creation, where texture-based feature vectors are extracted from a set of training images in order to define a feature subspace; and (b) cluster generation, where unsupervised k-means clustering is used to discover clusters in this feature space, with each cluster being denoted by its centroid. It must be noted that the subjects whose images were used in the training phase are not used in the testing/classification phase.

The *classification phase* has two stages: (a) gallery classification, where each gallery image in the database is projected onto the feature subspace created in the training phase and assigned a class label (i.e., a cluster index) based on the minimum distance criteria; and (b) identity retrieval, where the class label of an input probe image is first computed and a matching identity is obtained from the database by restricting the search to only those gallery images associated with this class label. Below we describe each phase in more detail. An overview of the complete system, including feature space creation, cluster assignment for gallery images and identity retrieval for probe images is given in Fig. 12.2.

All training and test ear images are pre-processed, in order to remove the influence of rotation, scale and illumination. We first adjust the contrast by using CLAHE [223]. Subsequently, we crop and geometrically normalize the images. This is done by applying cascaded pose regression (CPR)[62]. Using CPR, we train a classifier that fits an ellipse to the ear region, such that the ear is fully enclosed by the ellipse (see Fig 12.3). We then normalize rotate the cropped ear image, such that the major axis of the ellipse is vertical.

Finally, all images are resized to 100×100 pixels in order to compensate for different resolutions and to facilitate the extraction of local histograms in the subsequent step. We experiment with three different texture descriptors, namely LBP, LPQ and HOG.

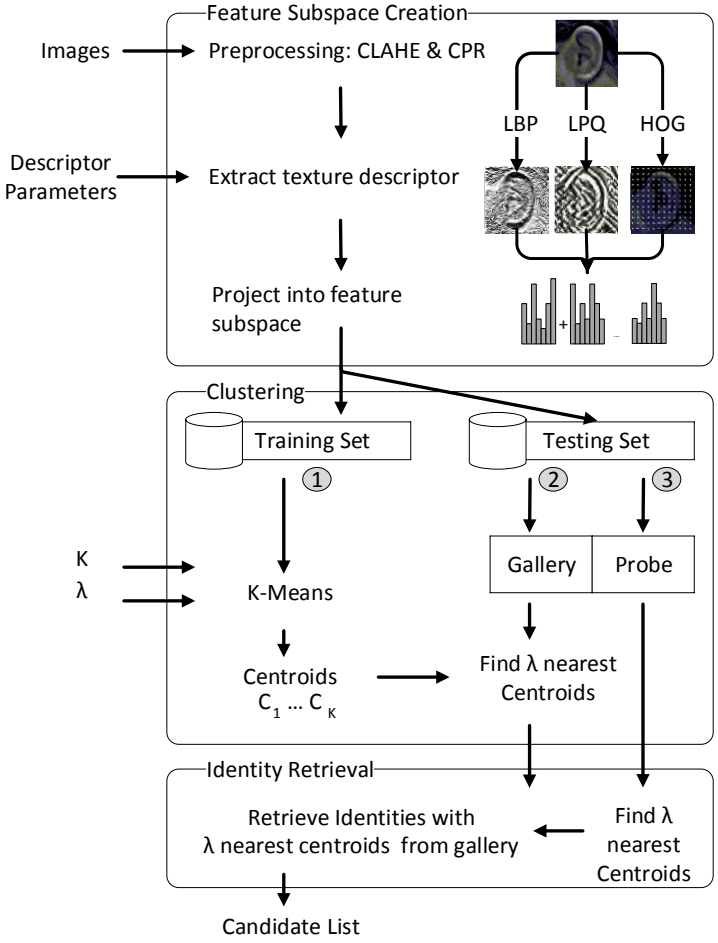


Figure 12.2: Illustration of the clustering scheme with preprocessing, cluster assignment and identity retrieval. The input parameters for each stage are shown on the left. The output is a list of possible gallery candidates to search.

12.3 Training Phase

12.3.1 Feature Subspace Creation

Uniform Local Binary Pattern (uLBP): uLBP [138] encodes local texture information on a pixel level by comparing the grey level values of a pixel to the grey level values in its neighborhood. The size of the neighborhood is defined by a radius around the pixel g_j, which is at least 1 (for a neighborhood having 8 pixels). Every pixel q_i within the radius that has a larger grey level value than the center pixel is assigned the binary value 1, whereas all pixels with a smaller grey level value are assigned the binary value 0.

The binary values of the neighborhood pixels are concatenated to form a binary string corresponding to the center pixel. Only those binary strings which have at most two bit-wise transitions from 0 to 1 (or vice-versa) are considered - there are 58 such strings. Each binary string is then mapped to a value between 0 and 58 (the first 58 bins correspond to the uniform binary strings, and the 59-th bin corresponds to the rest). The uLBP-based ear descriptor is computed by sliding a window of a predefined size and overlap (step size in pixels) in the horizontal and vertical direction over the LBP image. From each sub window

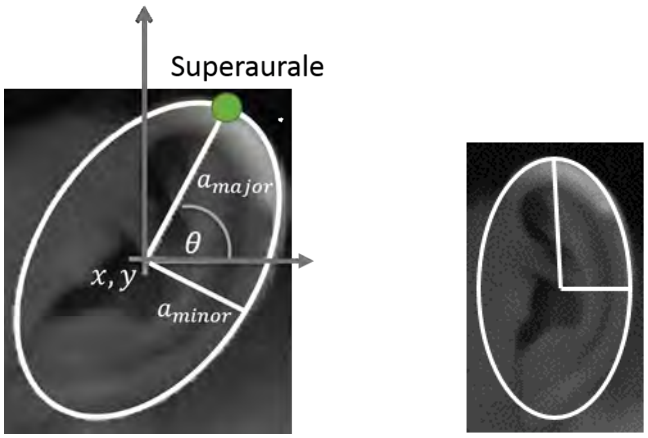


Figure 12.3: Illustration of the CPR-based geometrical normalization of ear images. We fit an ellipse that encloses the ear (left), and rotate the whole image such that the major axis of the ellipse is vertical (right).

a local histogram with 59 bins is extracted.

The final descriptor is the concatenation of each local histogram. For a window size of 20×20 pixels and an overlap of 10 pixels, this results in a feature vector of dimension 3776.

Local Phase Quantization (LPQ): The concept behind LPQ [10] is to transform the image into the fourier domain and to only use the phase information in the subsequent steps. Given that a blurred image can be viewed as a convolution of the image and a centrally symmetric point spread function, the phase of a transformed image becomes invariant to blur. For each pixel in the image, we compute the phase within a predefined local radius and quantize the phase by observing the sign of both the real and the imaginary part of the local phase. Similarly to uLBP, the quantized neighborhood of each pixel is reported as an 8-bit binary string.

Given an image, the LPQ value is computed for every pixel. Next, local histograms with 256 bins are computed within a sliding window. We move this window, with a certain overlap between two neighbouring windows, in the horizontal and vertical directions over the image and concatenate the resulting local histograms. For a 20×20 window size and an overlap of 10 pixels, this results in a 16,384 dimensional feature vector.

Histogram of Oriented Gradients (HOG): Computation of the HOG [57] descriptor involves five steps: gradient computation, orientation binning, histogram computation, histogram normalization and concatenation of local histograms. The algorithm starts with computing the local gradient by convolving a 3×3 region (HOG cells) with two one-dimensional filters (-101) and $(-101)^T$. The local orientation associated with the center of each HOG cell is the weighted sum of all filter responses within the cell. The local orientation is quantized into a bin value in the $[0, 2\pi]$ interval. Subsequently, the image is divided into blocks of equal size and a local histogram of quantized orientations is computed for each block. This histogram is normalized with the L2-norm. Finally, all local histograms are concatenated to form the HOG descriptor for the image. The HOG descriptor for a block size of 8×8 pixels and 9 orientation bins has 5184 dimensions.

Subspace projection: Once the texture descriptors are computed, they are projected onto a lower dimensional subspace using a projection matrix (one for each descriptor). The projection matrix is computed using PCA on the training set. The optimal number of dimensions for the target feature subspace is estimated using Maximum Likelihood Estimation (MLE) [113]. Depending on the choice of training data and the texture descriptor used, the resulting feature space has at least 18 and in some cases up to 150 dimensions.

12.3.2 Cluster Generation

Once the feature subspace corresponding to a texture descriptor is derived, the next step is to cluster the training data in this subspace (see step ① in Fig. 12.2). The K-means algorithm [1] is used to accomplish this. The input to the K-means algorithm is the projected feature vectors from the training data. The output consists of K cluster centroids, $\{C_1 \dots C_K\}$.

12.4 Testing Phase

12.4.1 Gallery Classification

In step ②, we divide the test set into two distinct parts, the gallery and the probe set. The subjects in the test set are different from the ones in the training set. The gallery set contains exactly one image for each identity. The probe set may contain any number of images for each identity. The images in the gallery and probe sets do not overlap. We use the feature extraction and projection matrix that were computed in the training stage to project the gallery images into the feature space. Let I_g be a gallery image and F_g be the projected feature vector (corresponding to one of the texture descriptors). Then, the distances between F_g and the cluster centroids is computed as $d_i = ||F_g - C_i||$, $i = 1 \dots K$. These distances are sorted in ascending order and the gallery image (identity) is labelled with the cluster indices corresponding to the $\lambda < K$ smallest distances.

12.4.2 Probe Identity Retrieval

In the retrieval step ③, the given probe image, I_p, is projected into the feature subspace (corresponding to a texture descriptor), resulting in a feature vector F_p. The distance between F_p and the K centroids is next computed, and the probe is assigned the cluster indices corresponding to the λ smallest distances. Thus the search is confined to the gallery images (i.e., identities) in λ target clusters. Note that $\lambda = 1$ denotes a single-cluster search, while $\lambda > 1$ denotes a multi-cluster search. The output of the retrieval process is the list of gallery identities, $L_{gallery}$, corresponding to the λ target clusters.

It is also possible to generate two different feature subspaces (e.g., corresponding to two different texture descriptors) and generate clusters independently in these individual subspaces. Let there be two sets of clusters corresponding to two subspaces S^1 and S^2 with centroids $\{C_1^1 ... C_n^1\}$ and $\{C_1^2 ... C_m^2\}$. The classification process will now result in two sets of cluster indices, one corresponding to S^1 and the other corresponding to S^2. Thus the output of the retrieval process will be two sets of gallery identities, $L^1_{gallery}$ and $L^2_{gallery}$. Subsequently, we can combine the two lists of identities using simple set operations such as union and intersection. The final list of gallery identities to be matched will be $(L^1_{gallery} \cup L^2_{gallery})$ or $(L^1_{gallery} \cap L^2_{gallery})$, respectively.

12.5 Experimental Analysis

The classification performance is defined in terms of a tradeoff between the pre-selection error and the penetration rate as defined in [92]. The pre-selection error rate (PSE) computes the probability that an image I_p from the probe set is not assigned to the same cluster as the corresponding identity I_g in the gallery set [2]. The penetration rate (PEN) is defined as the average fraction of the gallery database that has to be searched based on the list of

[1] We also evaluated Hierarchical Clustering and Gaussian Mixture Models, but neither of them returned satisfactory results. Hierarchical Clustering does not converge well and produces inconsistent solutions, whereas GMM returns one large cluster that covers nearly all of the identities and $K - 1$ small clusters that contain outliers.

[2] The pre-selection error rate is (1 - hit rate)

Table 12.1: $PSE//PEN$ (in %) for different configurations of the three texture descriptors for a single-cluster search. The performance is reported for different values of K. The best tradeoff between PSE and PEN for each configuration is denoted in bold. The best performance was achieved with LPQ-3-20-15.

	Number of Clusters (K)				
Algorithm	**K=2**	**K=3**	**K=4**	**K=5**	**K=10**
LPQ-3-20-0	1.37 // 65.46	7.6 // 35.37	**7.06 // 30.72**	9.13 // 29.30	30.00 // 14.25
LPQ-3-30-10	1.97 // 63.94	4.71 // 37.90	**5.10 // 32.49**	12.57 // 25.94	31.50 // 14.03 1
LPQ-3-20-15	0.77 // 65.81	4.75 // 38.65	**3.11 // 32.08**	3.77 // 32.64	30.43 // 13.4
LPQ-3-12-7	1.07 // 65.52	4.28 // 42.76	**5.09 // 32.94**	6.53 // 31.13	27.01 // 16.22
LPQ-5-20-15	0.61 // 57.25	7.66 // 35.34	**5.9 // 31.70**	10.23 // 27.85	27.21 // 14.31
LPQ-10-20-10	**4.16 // 53.06**	9.66 // 35.10	12.46 // 28.06	18.85 // 20.47	33.02 // 12.07
HOG-8-9	**6.46 // 49.88**	16.87 // 37.81	19.75 // 29.55	25.33 // 21.41	33.63 // 11.35
HOG-16-32	**11.14 // 50.08**	24.64 // 36.05	26.61 // 29.09	31.58 // 21.44	43.27 // 11.96
uLBP-1-20-0	5.60 // 50.69	**7.96 // 35.88**	9.12 // 30.24	19.25 // 21.70	33.17 // 12.30
uLBP-1-20-10	5.07 // 50.09	**6.64 // 35.54**	9.24 // 29.54	16.52 // 21.81	31.94 // 12.20
uLBP-1-20-15	4.61 // 50.85	**5.15 // 36.00**	5.13 // 20.83	16.43 // 18.83	31.17 // 12.63
uLBP-2-20-10	**5.64 // 50.04**	8.61 // 36.18	10.17 // 20.33	21.6 // 17.16	33.90 // 11.93

retrieved gallery identities. The ultimate goal in classification is to reduce both the penetration rate as well as the pre-selection error rate. In an ideal clustering and classification scheme, the pre-selection error rate would be zero and the penetration rate would be $1/n$ where n is the number of images in the gallery set.

Let K be the total number of clusters in a feature subspace. Further, let n be the number of images in the gallery set and m the total number of images in the probe set. In our experiments, each identity in the gallery has exactly one image. Let $\xi_{p_i} \subset \{C_1 ... C_K\}$ be the cluster labels of I_{p_i}, the i-th probe, and L_{p_i} be the corresponding list of gallery images (identities) retrieved from the database. Moreover let $\xi_{g_i} \subset \{C_1 ... C_K\}$ be the cluster labels of the corresponding gallery image I_{g_i} with the same identity as I_{p_i}. Note that $|\xi_{p_i}| = |\xi_{g_i}| = \lambda$.

$$Hit(p_i) = \begin{cases} 1, & \text{if} \quad \xi_{p_i} \subseteq \xi_{g_i} \\ 0, & \text{otherwise} \end{cases} \tag{12.1}$$

$$PSE = 1 - \frac{1}{m} \sum_{i=1}^{m} Hit(p_i) \tag{12.2}$$

Accordingly, the penetration rate can be written as

$$PEN = \frac{1}{m} \sum_{i=i}^{m} \frac{|L_{p_i}|}{n}. \tag{12.3}$$

12.6 Evaluation and Results

All results in this section are based on a heterogeneous dataset that has been composed of images from the UND-J2 (1800 images from 415 subjects) [201], AMI (700 images from 100 subjects) [70] and IITK (494 images from 125 subjects) [107] databases. The dataset used in our classification experiments consists of 2432 images from 555 subjects: 363 subjects from UND-J2, 67 subjects from AMI and 125 subjects from IITK. There are at least two samples per subject. (Images of 52 subjects from UND J2 and 33 subjects from AMI were used to train the CPR model for ear normalization, and were not used in subsequent experiments).

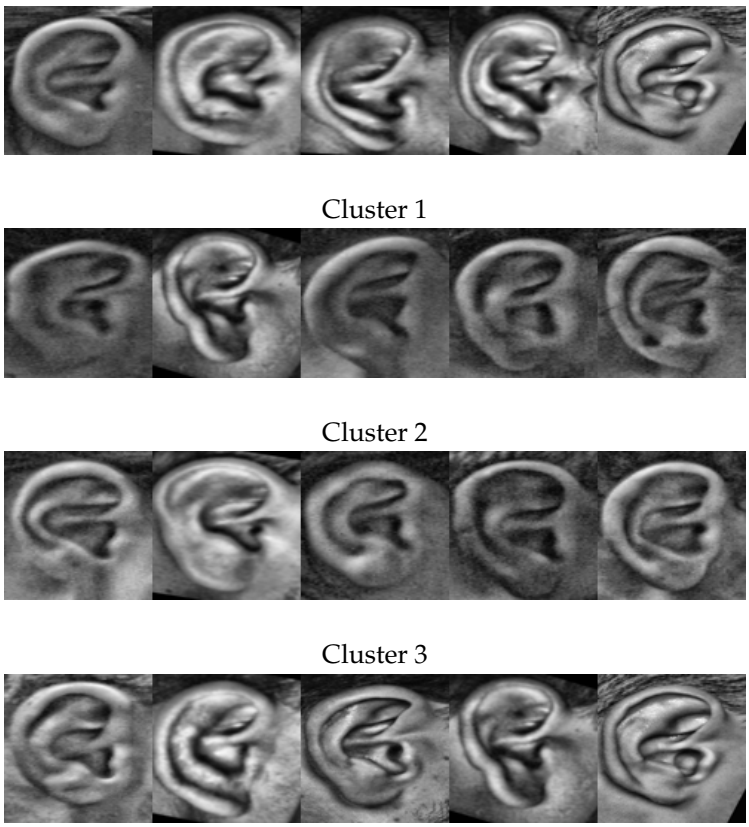


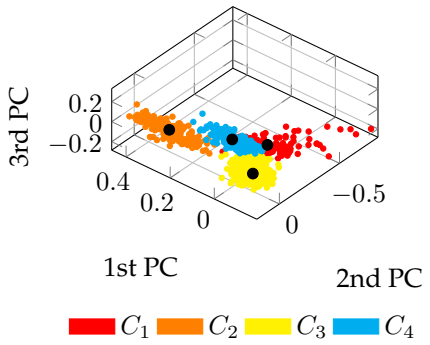
Figure 12.4: Example images in each cluster for LPQ-3-20-15 with K=4. The images show the closest ears to each cluster centroid in ascending order (from left to right)

Training is accomplished by randomly selecting 253 subjects; all images of the 253 subjects are used in the training phase. On average, the training set contains approximately 1100 images. The remaining 302 subjects are used for testing. For each test subject, 1 image is added to the gallery database while the remaining images are used as probes. All experiments were conducted with 10-fold cross-validation.

In order to generate the clusters for a specific feature subspace, the K-means algorithm is used. The input consists of the projected feature vectors for a set of training images and the output is a set of cluster centroids. Since the output relies on the initialization process, the K-means technique is run 1000 times with a different set of initial centroids each time. From these 1000 solutions, we pick the one with the smallest sum of distances between all feature vectors and their respective cluster centroids. An analysis of the best solution using the silhouette measure [168] indicated that small values of K result in more coherent clusters than large values of K.

Given that we have a solution with a fixed number of centroids, we evaluated the performance of the proposed classification and retrieval scheme in three different steps. In the first, step we focus on the pre-selection error when the search process is confined to a single cluster for each probe. In the second experiment, we allow the search to expand to multiple clusters corresponding to the nearest centroids. Finally we evaluate the classification performance when candidate lists corresponding to multiple feature subspaces are combined [3].

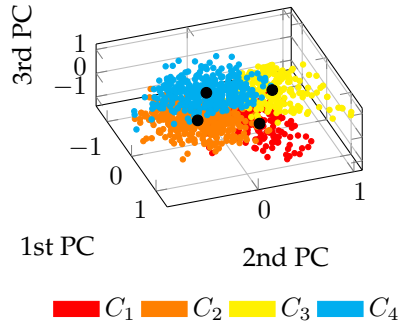
[3]We found that fusing the identity lists of more than two cluster spaces does not improve the performance.



Probe Cluster Label

Gallery Cluster Label	1	2	3	4
1	37	1	0	0
2	0	69	0	0
3	0	1	137	1
4	0	0	2	47

Figure 12.5: Cluster analysis for LPQ-3-20-15 with $K = 4$. The left figure shows the first three principal components (PC) of the feature space showing 4 different clusters (colors) and their respective centroids (black) and to the right, the confusion matrix.



Probe Cluster Label

Gallery Cluster Label	1	2	3	4
1	20	7	3	11
2	4	69	3	11
3	4	2	53	5
4	3	22	7	78

Figure 12.6: Cluster analysis for HOG-8-9 with $K = 4$. The left figure shows the first three principal components (PC) of the feature space showing 4 different clusters (colors) and their respective centroids (black) and to the right, the confusion matrix.

12.6.1 Single Cluster Search

In the first experiment, we compare the classification performance due to feature subspaces generated from uLBP, LPQ and HOG. Each texture descriptor was tested with different parameter sets and with different window sizes. However, we found that many of the configurations result in similar performance. In Table 12.1, we report the pre-selection error (denoted as PSE) and the penetration rate (denoted as PEN) of specific configurations. The configurations for LPQ and uLBP-techniques are defined as follows: `<algorithm> - <radius> - <windowSize> - <overlap>`. The configuration for HOG is defined as `<algorithm> - <block size> - <number of bins>`.

For all texture descriptors in Table 12.1, we see that the PSE declines monotonically with an increasing number of clusters. As expected, the penetration rate decreases with an increasing number of clusters. This implies that there is no optimal number of clusters that can be automatically determined for each of the feature subspaces.

When comparing the performance of LBP, LPQ and HOG feature spaces, we observe that LPQ with a small radius and large overlap between the local windows has the best classification performance (also see Fig. 12.4). In our experiments, HOG yields the largest pre-selection error rates. The performance of uLBP lies between HOG and LPQ when the number of clusters is smaller than 6. For $K \geq 10$ the classification performance of HOG and uLBP becomes similar. When performing single cluster search with LPQ, solutions

Table 12.2: PSE and PEN (in %) for different configurations of the three texture descriptors when multi-cluster search is used. Here K = 10. The best tradeoff between PSE and PEN for each configuration is denoted in bold. The best results were achieved with LPQ-3-20-15.

	Number of clusters searched (λ) for K = 10.				
Algorithm	**λ=2**	**λ=3**	**λ=4**	**λ=5**	**λ=6**
LPQ-3-20-0	10.58 // 28.01	2.77 // 40.99	**1.30 // 53.48**	0.27 // 65.93	0.10 // 74.98
LPQ-3-20-10	8.77 // 27.90	2.00 // 40.80	**0.83 // 53.56**	0.27 // 66.55	0.03 // 75.24
LPQ-3-20-15	6.20 // 32.74	**0.99 // 47.10**	0.40 // 61.69	0.00 // 73.58	0.00 // 81.68
LPQ-3-12-7	8.07 // 31.41	**1.70 // 46.08**	0.60 // 59.99	0.30 // 72.49	0.20 // 81.06
LPQ-5-20-15	8.30 // 28.12	1.47 // 41.65	**0.53 // 53.17**	0.23 // 64.48	0.03 // 75.64
LPQ-10-20-10	14.45 // 23.08	6.89 // 34.23	3.41 // 45.67	1.97 // 56.52	**0.57 // 67.31**
HOG-8-9	17.43 // 22.34	8.83 // 33.58	4.27 // 44.31	2.17 // 54.71	**1.30 // 65.11**
uLBP-1-20-10	11.32 // 23.62	4.30 // 34.66	2.40 // 44.42	**1.73 // 53.93**	0.97 // 63.21

with K=4 appear to be a good choice, whereas for uLBP, $K \leq 3$ appears to be good. The HOG descriptor does not seem to lend itself to clustering since the pre-selection error rate is larger than 5% for K=2.

The penetration rate for HOG and uLBP is roughly $1/K$, whereas the penetration rate for LPQ tends to be larger than $1/K$. We can conclude from this, that the points in all examined feature subspaces are not uniformly distributed and that the number of identities per cluster is different. This is further illustrated in Fig. 12.5, where an example solution for LPQ-3-20-15 is shown. Fig. 12.6 shows an example solution in the HOG feature space, where the preselection error is significantly larger than for LPQ. As shown in Fig. 12.5, clusters 3 and 4 mainly contribute to the overall pre-selection error, because these two clusters are located next to each other in Fig. 12.5. [4] As shown in Fig. 12.6.3 the number of identities per cluster varies across the clusters. These variations can partly be explained by the fact that the input images come from three different datasets that contain a different number of subjects. C_2 mainly contains images from IITK, whereas C_1 contains many images from AMI. Fig. 12.4 shows examples of the five closest ear images to each centroid for LPQ-3-20-15. Images that originate from a particular database are overrepresented in some of the clusters; however, each cluster contains images from all of the three original databases. This implies that the classification not only reflects the identity of a person, but also contains information about skin tone (IITK contains ear images from Asian Indians, while AMI and UND-J2 mainly contain images from Caucasians). This is confirmed by evaluations using only a single database, where the capture settings and the skin tone of most subjects are the same. The performance of these individual databases is lower than that of the combined dataset. On UND-J2, for instance, we obtain a penetration rate of 81.45% for a pre-selection error rate of 1.11%.

Cross-database evaluations show that texture descriptors contain information that captures the demographic attributes of the subjects and the acquisition settings. The cluster centroids obtained from one database do not properly reflect those from the other databases. We plan on incorporating additional features related to the shape of the ear to mitigate this concern.

12.6.2 Multi Cluster Search

In the second experiment, we explore the impact of multi cluster search. Based on the probe feature vector, the clusters corresponding to λ nearest centroids are searched. This

[4] The reader has to exercise caution when interpreting these figures. These are projected features - the original dimensionality is 73

potentially decreases the pre-selection error, but will also increase the penetration rate. The results for this experiment are summarized in Table 12.2. Here, the best configurations from Table 1 corresponding to K = 10 were used. We found that, for solutions with a larger K, the number of clusters does not influence the rate in which the penetration rate converges. This means that a graph depicting the ratio between pre-selection error rate and penetration rate will have the same shape, regardless of K (see Fig. 12.8). For higher values of K, we have more possibilities to select λ in a way that meets the requirements of a particular application.

As the results show, multi cluster search quickly reduces the pre-selection error at the cost of increased penetration rate. Due to the fact that the number of identities in each cluster varies, the penetration rate increases much faster than $1/\lambda$ with $\lambda < K$. However, searching through the closest two clusters significantly improves the performance by keeping the penetration rate below 50% while reaching a pre-selection error that is as small as 0.1% for LPQ-3-20-15. For other LPQ-based configurations with a radius of 3, the pre-selection error falls below 1% when four clusters are included in the search process at the cost of a higher penetration rate between 53.5 and 60%. All the other configurations result in searching at least 75% of the gallery images in order to obtain a pre-selection error below 1%.

Upon evaluating the identification performance, we obtain a rank-1 recognition rate of 93.06% when searching through three neighboring clusters using LPQ-3-20-15. As opposed to an exhaustive search, where we would obtain a similar performance, we only have to compare against 47.10% of the images, on an average, in the database.

12.6.3 Feature Space Fusion

As pointed out in Section 12.4.2, the identity lists corresponding to multiple feature subspaces can be combined to facilitate the retrieval process. Fusion is carried out by either using the union or the intersection of these two lists. Additionally, the previously mentioned multi-cluster search can be used in each of these feature spaces. Our results are based on different numbers of clusters searched (λ) and K = 10.

As expected, the penetration rate when using the union operator on the identity lists is much higher than when using the intersection operator. The intersection operator results in a pre-selection error rate of 1.98% at a penetration rate of 55.53% when a single cluster search strategy is used in each subspace. Searching through 5 clusters only slightly improves the performance and yields a pre-selection error rate of 1.65% and a penetration rate of 66.69%. Using the union operator results in a pre-selection error rate of 0.99% at a penetration rate of 47.65% when searching through 3 clusters in each feature space. This implies that the classification performance was not necessarily improved when fusing two identity lists.

12.7 Summary and Future Work

Using a single cluster search strategy the best results were obtained using LPQ with a radius of 3 and a 20×20 window size with 15 pixels overlap (pre-selection error rate was 3.11% with a penetration rate of 31.7%). A multi cluster search strategy further reduces the pre-selection error to 0.99% with a penetration rate of 47.1%. In summary, we have the following observations.

- Unsupervised classification of 2D ear images using texture descriptors is possible.
- Solutions with four clusters are a good choice for single cluster search when using the LPQ texture descriptor.
- A multi cluster search strategy further improves the classification performance.

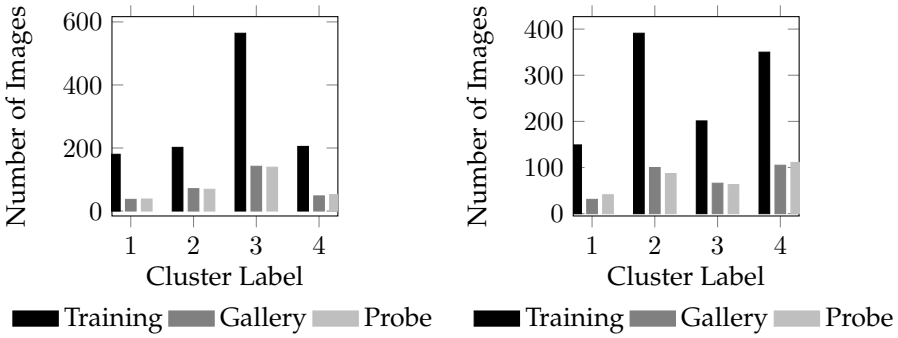


Figure 12.7: Number of images per cluster in the training set, gallery set and probe set for LPQ-3-20-15 (left) and HOG-8-9 (right)

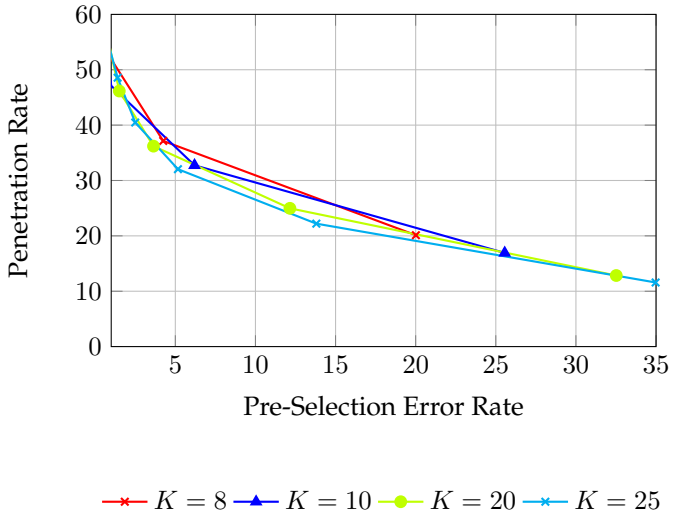


Figure 12.8: Impact of K on the convergence rate of the pre-selection error rate and penetration rate trade-off with LPQ-3-20-15. The number of clusters to be searched increases from right to left.

- Fusion of candidate lists corresponding to two different feature subspaces using the union or intersection operator does not improve the classification performance.

This work can be extended in many ways: (a) We will investigate if all three texture descriptors cluster the subjects similarly. (b) In this work, only left ear images were used. We plan to investigate if the left and right ears of subjects are clustered similarly. (c) We will study the classification error when the quality of the input image is degraded. (d) We plan on incorporating shape features to improve classification accuracy, especially in homogeneous datasets.

Part III

Conclusions

Chapter 13

Summary of Results

13.1 Segmentation

We have proposed three different approaches for segmenting the ear region from face profile images and coarsely segmented images. In addition, we have evaluated the impact of signal degradation and image compression on a selected of sliding window detectors using state of the art features.

13.1.1 Segmentation from Face Profile Images using Surface Curvature and Texture

In this work we have proposed two schemes for segmenting ears from depth maps (3D) and 2D images (see Chapters 4 and 5). We saw that the surface structure of the outer ear is so unique, that a small set of simple rules is already sufficient for segmenting the ear region from the rest of the face profile image. This set of rules describes simple geometric properties of the ear that can be extracted from a profile depth map by observing clusters of high surface curvature. The set of rules is deigned in a way that the algorithm is invariant to in-plane rotations and robust to pose variations and partial occlusion. We also showed that this approach can be further refined by combining edge and surface information.

Similar approaches, where clusters of high surface curvature are analyzed in order to locate the ear region from face profile images were proposed by Chen and Bhanu [48] and Prakash and Gupta [157]. One should be cautious when comparing these approaches, because they were evaluated on different datasets, with a different ground truth and under different constraints for a successful detection. Our segmentation approach has a similar detection performance as the approach of Prakash and Gupta.

13.1.2 Segmentation from Face Profile Images using a Sliding Window Approach

Motivated by the excellent performance in Zhou et al. [218], we made an attempt to modify their approach in a way that the algorithm becomes invariant to in-plane rotations (see Chapter 6). Instead of a rectangular detection window, we used a circular window and created a fixed length histogram descriptor from the resulting polar coordinate system. The goal of rotation invariance was, however only party achieved at high computational costs. We recommend to stick with rectangular detection windows and apply a geometric normalization instead.

In addition to the high computational costs, the detection performance of the circular descriptor was not satisfactory. Smoothing effects and data loss during the projection from the Cartesian into the polar coordinate space result in a loss of information and ultimately in a lower detection rate. In a later analysis, we learned that the inner and outer radii of the projected space have a tendency towards an average feature vector. In polar representation, pixels from the inner radii are over-represented. Pixels on the outer radii are merged by calculating the average of several neighboring pixels. Hence, we lose information during the projection for pixels on the outer radii. As a result, the circular feature vector contains less distinctive information than the rectangular descriptor and hence performs worse.

Table 13.1: Summary of results on ear segmentation from our work

Name	Summary	Database	Perf.
Surface Curvature (Chapter 4)	Reconstruct ear outline mean curvature points based on a fixed set of rules. Invariant to rotations.	UND-J2	95.65%
Surface and Texture (Chapter 5)	Combine mean curvature with edges from texture and apply extended rule set for outline reconstruction.	UND-J2, UND-NDOff	99%
Circular HCS (Chapter 6)	Train SVN for ear and non-ear decision using a histogram descriptor and select ear region from density.	UND-J2	94.17%
CPR (Chapter 7)	Train cascaded detector for fitting an ellipse that encloses the ear region. Returns exact ear region orientation.	UND-J2, AMI, IITK	99.63%

This issue could be solved by modifying the circular descriptor in a way that each row in the projected space contains as many pixels as there were in the original image. The resulting descriptor would still be of fixed length, but the influence of each row would scale with the amount of contained information. A proof of concept for this idea is left for future work.

13.1.3 Normalization of Ears in Cropped Images

In Chapter 7 we have applied an existing approach for the estimation of orientation of fish and mice to ear detection [62]. We have trained a single-part elliptical model that detects the ear region in face profile images (UND-J2) and refines the segmentation in pre-cropped images (IITK and AMI). The detector returns the parameters of an ellipse and hence also contains information about the orientation of the ear. To our best knowledge, this is the first attempt for estimating the orientation of the outer ear.

We have shown that the segmentation accuracy of Cascaded Pose Regression (CPR) is competitive with existing sliding window approaches. We also showed that the estimated orientation is accurate enough to improve the recognition performance of an ear recognition system using texture descriptors. The same authors have shown in a later publication that CPR is also suitable for the detection of prominent landmarks, such as the nose tip and the mouth in face images [36].

13.1.4 Impact of Signal Degradation and Compression

Our research on ear segmentation is concluded with a study on the robustness of sliding window detectors to signal degradation. We trained Adaboost ear detectors using Haar-like features, LBP and HOG (using images that were compressed with PNG). The detectors were then applied to copies of the database, where we added noise and blur to each image. To our best knowledge, this is the first study on the impact of signal quality on ear detection systems. Our conclusion is that LBP is the best choice if the quality of the probe image is degraded by noise and blur.

The same experiment (same partitioning, same detectors, same error metrics) was repeated on a copy of the database where the ear images were with JPEG and JPEG 2000. The latter series of experiments was, however, not part of the original publication in Chap-

Figure 13.1: Close-up of the ear in an uncompressed image (left) and the compressed versions of the same images for JPEG (Middle) and J2K (right) with 0.2 bpp.

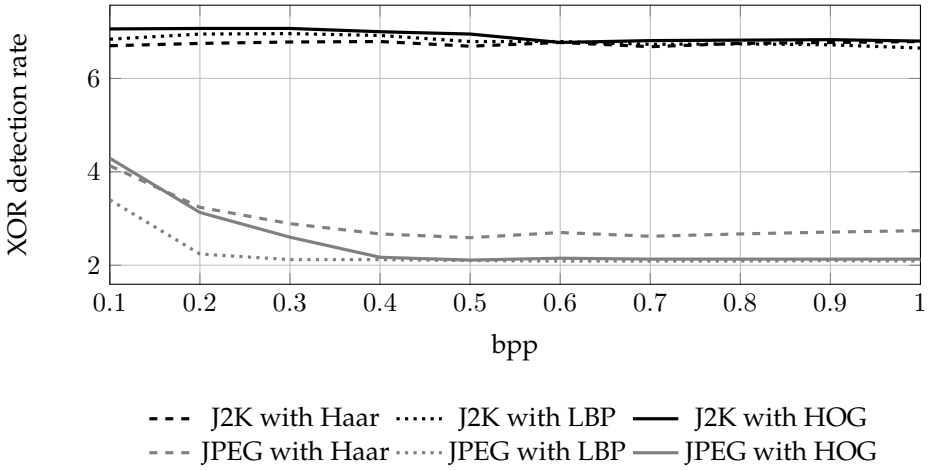


Figure 13.2: XOR-detection errors for JPEG and J2K under different compression rates, measured in bits per pixel (bpp).

ter 10. The results in Figure 13.2 and the example images in Figure 13.1 are taken from the bachelor thesis of Johannes Wagner [185].

For compressed images, we obtain a higher average error rate than for the images where only blur and noise was added. Unlike in the experiments with noise and blur, the detection performance of the scenario with half of the resolution is sometimes slightly better compared to the experiment where the full resolution of the images is used. Again, LBP proves to be a good choice for a satisfactory detection rate, even under strong compression. To our surprise, the average detection rates for all feature vector using JPEG images are higher than the detection rates of J2K. We assume that this is due to the fact that JPEG preserves edges at low compression rates, whereas J2K introduces blur. However, further and more in-depth analysis is needed to confirm this observation.

13.2 Feature Extraction

13.2.1 Texture and Surface Descriptors for Ear Recognition

During the work on this thesis and on GES-3D, we have evaluated different approaches for landmark detection in ear images, including force field descriptors, steerable filters and Active Shape Models (ASM). None of these methods was able to give us stable and repeatable landmarks, even though some authors reported good performance rates on their datasets. Our approaches with steerable filters (see Section 13.4.3) and force fields were vulnerable to occlusion and pose variations. Active Shape Models and Active Appearance Models require a sufficiently large amount of training data, which we do not have at hand. Even after carefully initializing the ASM in a normalized ear image, the results were unstable

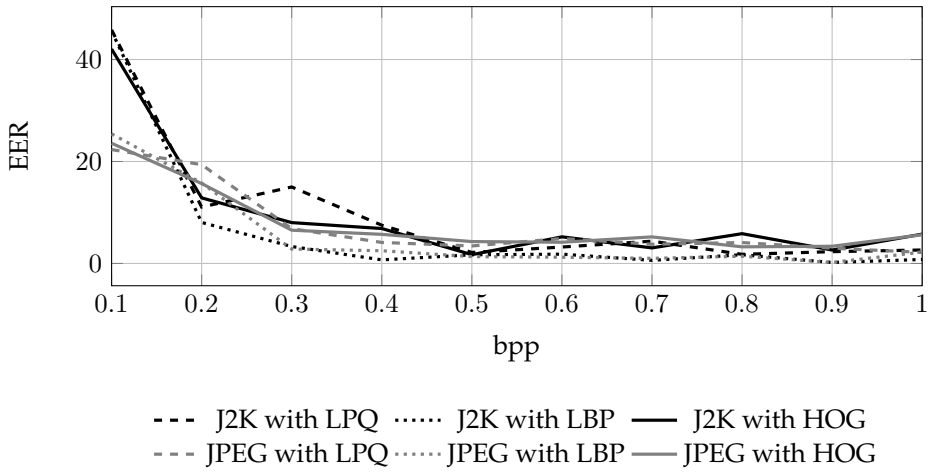


Figure 13.3: EER for JPEG and J2K under different compression rates.

and the extracted edges have a strong tendency to rather resemble the average ear than the ear in the current image.

We tried to use the surface clusters from the 3D ear detection step for extraction prominent points, such as the tragus and the helix. This approach partly worked, but the number of available landmarks is too small and their position is not stable enough, especially under pose variations (also see Section 13.4.2)).

After a number of unsuccessful attempts, we decided to omit the landmark detection step and to use texture and surface features instead. In Chapter 8, we provide a comparative study on the performance of fixed length texture descriptors on different datasets and propose a combined texture descriptor that uses 2D and depth information. Our descriptor, however, was not able to outperform the 2D reference configurations. On average, the LPQ and the BSIF descriptor showed the best performance on all of the three datasets. Our work confirms that the parameters of a method can be engineered towards a certain dataset in order to obtain perfect results on that particular database.

The same configuration on another dataset, however, may yield different results. From this series of experiments, we learned to interpret raw performance numbers with caution, especially if they were generated with the same dataset. Gradual differences between single EER or IR are a very limited indication for one or the other approach.

13.2.2 Impact of Signal Degradation and Compression

We conclude our research on ear recognition by a study on the robustness to blur, noise in Chapter 11. The scenario and the parameters of this study are the same as for the study on ear detection from the previous chapter. In conformance with our results from Chapter 10, the most reliable feature extraction approach under signal degradation through noise and blur is LBP.

A second series of experiments was conducted on compressed images. The settings for this experiment are the same as described before in Section 13.1.4. An excerpt from the results is shown in Figure 13.3. Please refer to [185] for the further results. Similarly to the results from ear detection, the performance for all texture descriptors drops for compression rates that are smaller than 0.4 bits per pixel. These results indicate that ear images can be compressed to about a quarter of their original size with a minimal loss of recognition performance. Hence, ear recognition systems can work with highly compressed data (under the absence of illumination, scale and pose variations).

13.3 Fast Comparison Techniques

As the literature survey from Chapter 3 already indicated, many known approaches from computer vision in general and approaches that work for other characteristics have been applied to ear recognition. The next step towards a competitive recognition system ware search and indexing techniques. We have presented into two different approaches for accelerating the search and minimizing errors in ear databases. Another approach for indexing of 3D ears using KD-trees was proposed by Maity and Abdel-Mottaleb in December 2014 [120].

13.3.1 Two-stage search with Binary Representation

During the statistical evaluations on the behavior of texture descriptors, we observed that many descriptors with a good performance are sparsely populated. Based on this observation, we converted different LPQ and BSIF texture descriptors into a binary representation and uses the binary feature vectors for creating a short list prior to comparing the original, real-valued histograms. We were able to perform this two-stage search in 30% of the time compared to the time would be required when we would perform an exhaustive search with real-valued feature vectors (see Chapter 9).

These experiments confirm, that the bin value of a histogram is only of secondary importance. According to our experiments, this observation is not restricted to ear images. Our binarization method can be applied to all characteristics, where fixed length histogram descriptors are suitable feature representations. The distribution of non-zero bins in the histogram allows us to narrow the search space to only 10% of the dataset. Additionally, the two-stage search minimizes the chance for a false positive identification, at the cost of a possible pre-selection error(*i.e.* if the subject we are searching for is not in the short list).

The concept of two-stage search has been proposed and successfully applied to other biometric characteristics before in [126] and a similar approach was recently proposed by Billeb et al. [30] for speaker recognition. The recognition performance that can be achieved with our approach is comparable to the results from Maity and Abdel-Mottaleb [120].

13.3.2 Unsupervised Classification

An alternative to an exhaustive search is unsupervised clustering. We know that the distribution of shape classes as defined by Iannarelli [81] is dominated by the oval ear shape, which makes shape-based classification unsuitable for labelling ear images in order to reduce the number of comparisons. We also know that the assignment to one to these shape classes may be arbitrary, because there are no fixed criteria for any of Iannarelli's shape classes. Our goal was to define classes directly from the feature space.

We applied different analysis techniques and clustering methods and finally came to the conclusion that K-means provides us with a stable and reliable classes in the feature space. Although there are no naturally separated convex or concave clouds in the feature space, we found that there are feature spaces where subjects are projected into stable positions. We also found that these feature spaces are not necessarily well-suited for recognition. We recommend to create a large number (K=10) of clusters and to extend the search to multiple adjacent clusters. With this approach, we were able to reduce the search space to less than 50% of the dataset with a 99% chance that the correct identity is included in the preselected data.

Unsupervised classification also reveals additional information about the ethnicity and the gender of the subject, which may be useful for forensics. Compared to the previously introduced two-stage binary search, classification is the better solution when the database exceeds a certain size and if many search operations have to be handled. Our classification approach reduces the number of comparisons and may save some bandwidth. However,

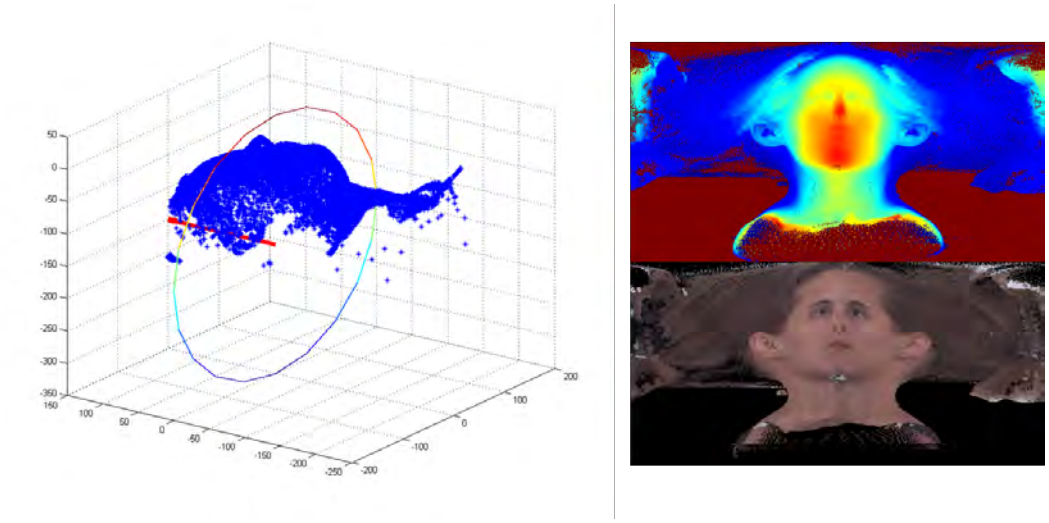


Figure 13.4: Left: Example for the positioning of a cylinder in 3D space. The central axis of the cylinder is drawn in red. Right: The projected depth and texture image on the cylinder surface.

for identity retrieval systems where all binary feature vectors fit into the RAM, the two-stage search should be the preferred solution.

13.4 Further Results

During the project work, several other approaches that have not been mentioned yet were proposed. These approaches were intended to be part of the processing pipeline for the multi-model recognition system in GES-3D, but were not further pursued. In this section we briefly discuss these approaches and reasons for not further pursuing them.

13.4.1 Cylinder Projection

As pointed out in the introduction, the GES-3D system is intended to be robust to pose variations through using 3D models as references. The probe images, however are 2D images from video streams. The idea was to create a representation of the 3D model, that would fit to any view in the probe video and that would enable the usage of fast image processing algorithms. The solution is inspired by the work of Spreeuwers [178], who proposed to project the surface of the 3D model onto a cylinder surface.

A series projected 3D surfaces (a 2.5D image) on a cylinder surface contains all possible views of the camera to the subject, such that we would have the correct part of the face for any azimuth angle between the probe image and the subject. Figure 13.4 shows an example of a cylinder enclosing the 3D data and the projected texture and surface with an azimuth angle equal to zero. We also developed several concepts of how to use this representation for creating a descriptor for the face and both ears in the projected texture image.

During our analysis, we observed several drawbacks of this method in the forensic context, which finally brought us to the decision to not including it into the GES-3D System.

- **Distortion:** Depending on how the cylinder surface is initialized, the projected image is slightly distorted (see Figure 13.4). These distortions alter the appearance of the reference model, which is undesirable when the system is used in the context of forensics.

- **Positioning of the central axis:** The correct positioning of the central axis of the cylinder is an important step for computing the projected surface. In our implementations, we encountered problems with finding the same central axis for different 3D models of the same subject. The repeatability of the positioning is considered to be important, because the distortions of the surface depend on the central axis of the cylinder.

- **Pose estimation in probe video:** The projected reference images need to be matched with the right angle of the probe video stream. The estimation of this pose, however, turned out to be a challenging problem. Without a good pose estimation, we are unable to map the correct part of the projected texture to the probe image.

13.4.2 Features from Ear outlines

Based on the encouraging results in Chapter 4 and 5, we evaluated different possibilities of using the ear outlines from the detection algorithms for describing the proportions of the outer ear. The goal was to have a method that does not only segment the ear from depth images, but also delivers the basis for the subsequent feature extraction step. Figure 13.5 gives an overview of the features we considered. The approach was to extract a descriptor for the edges from the reference image and to find the best matches for these edges in the probe image. We also experimented with global shape descriptors, such as Shape Context [24] and Chain Codes [65].

As a part of this research, we made an attempt to locate the tragus region, the uppermost part of the outer helix and the lobule using the ear outlines returned by the segmentation. We were able to locate the uppermost part of the outer helix in many cases, however localization of the lobule and the tragus turned out to be a challenging task when the pose of the input images varies. The accuracy of the localization of these landmarks was also limited by occlusions. Even an accurate localization of these three landmark would most likely be insufficient for distinguishing between subjects in large datasets.

In addition to this, the edges from the segmentation have a large intra-class variation. This has several reasons. Firstly, the selection of edges for an ear candidate depends on the order in which these edges are selected. As such, two ear candidates from the same subject may be composed of different shapes. They often look similar, but not alike. Secondly, the shape candidates (edges) from the 2D image vary with illumination and occlusion. We are unable to distinguish between spurious edges and edges that actually represent a part of the ear. For the segmentation, this is not an issue, because we usually find enough edges for creating reliable ear candidates.

We finally decided to stop the work in this approach. Our conclusion is, that more sophisticated approaches are needed for reliably extracting the ear outline, especially if no depth information is available.

13.4.3 Fourier-CHOG Features

The Fourier-CHOG detector by Skibbe and Reisert [177] was originally intended for detecting self-similar objects, such as pollen in microscope images. Fourier-CHOG is a steerable filer that is trained with a set of annotated images. During the training phase, the filter derives a number of coefficients from the training images, such that the filter responds with a maximal value for a specific type of objects. The filter is capable of detecting objects, with smoothly bent edges. This constraint can also be applied to several parts of the outer ear, such as the helix, lobule or the tragus. The goal was to use Fourier-CHOG for locating the ear regions and then use the relative positions and the shape of the extracted regions as a feature.

We trained five different Fourier-CHOG filters and used them for the detection of five different regions in an ear image (see Figure 13.6). The accuracy of the filter depended on the image region. Whereas the upper and lower helix, could be detected reliably, the

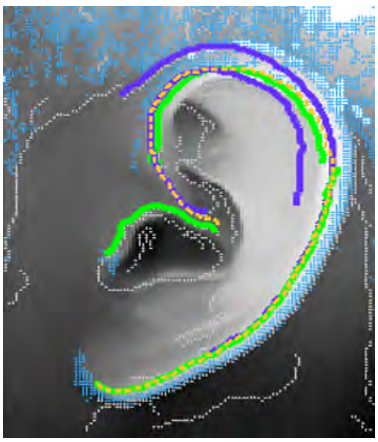


Figure 13.5: Possible candidates for edge features from the segmentation method in Chap. 5. The light blue dots denote local maxima of the surface curvature, the blue bold lines are edges from the 2D image and the green lines are edges that were generated from the 3D surface curvature. The dotted yellow line represents the ear candidate, which is generated by the segmentation method.

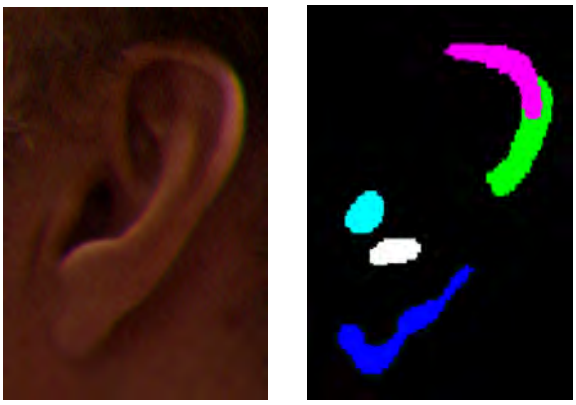


Figure 13.6: The CHOG detector can be trained to locate different regions in ear images. The example shows the detected image regions for the upper helix (purple), lower helix (Green), lobule region (blue), tragus (cyan) and antitragus (white).

accuracy for tragus, antitragus and the lobule was low. Another limitation of this approach was, that we were unable to define a threshold for a minimal filter response that would be appropriate for all subjects. In some cases, the filter response for the ear region was even lower than for other objects, such as earrings.

Even in cases where the detection was successful, the shape of the ear region is not a suitable feature. The intra-class variations between the shape of the regions and the number of wrongly detected regions are too large. We therefore decided to discard this approach.

Chapter 14

Future Work

14.1 Unsolved Challenges and Missing Data

The evaluation results from the GES-3D project indicate that ear recognition is far from being a solved problem, when applied under realistic conditions. The performance rates fr academic datasets that are reported in the literature tend to be overly good. The main reason for this is, that the images within a certain dataset are usually collected under the same conditions an with the same capture device. The demographics of the data subjects are also strongly biased, because they are recruited from the university. As such, the datasets are dominated by young people. Compared to face recognition, these capture scenarios for ear recognition datasets are rather easy, which explains the gap between the performance in literature and the conformance of the GES-3D system.

The exact performance rates of the experiments on any academic dataset are only of secondary importance, as they only serve as a proof on concept. Progress in research on ear recognition systems vitally depends on the availability new and more challenging public databases.

For keeping research in the field active, new databases for benchmarking ear recognition systems should be collected and released. For instance, such a dataset could contain higher demographic variations, left and right ears from the same subject, full 3D models of the ear (meshes), or images that were acquired over a longer period of time. We are aware that, besides privacy and data protection issues, the collection of biometric datasets requires long-running projects with sufficient personnel to execute the data collection and prepare the data for publication. New datasets would not only accelerate pace of work in ear biometrics, but would also offer the opportunity take the existing systems to a new level.

Apart from the need for more challenging datasets, some interesting and yet unsolved aspects of ear biometrics are:

- Reliable detection of landmarks in ear images under pose variation and occlusion. Especially the detection and masking of occluded areas would be an important contribution for ear recognition in surveillance scenarios.
- Further techniques for accelerating search operations in large databases for 2D and 3D ears. In particular for forensic image analysis, it would be helpful to use shape or surface information for ear classification.
- Evaluate the suitability of binary ear descriptors for template protection. Binary feature vectors allow us to perform fast searches and can be used with existing template protection techniques.
- Obtain a better understanding of the differences between the left and the right ear of the same subject. These differences could be used for creating feature vectors that encode only the differences between the left and the right ear, which would be harder to forge.
- A database that allows us to investigate ageing effects in ear biometrics and/or the impact of pose variation on the appearance of the outer ear.

14.2 Alternative sources for 3D images

The Kinect camera system has drawn some attention from the computer vision community. Driven by the new imaging possibilities the community developed systems for face recognition using the 2D and the 3D video stream of the camera. In the context of this work, we also evaluated the Kinect for ear recognition, but found the depth resolution of the camera to be insufficient for capturing enough details of the ear structure.

With the release of the Kinect 2 [1], another low-cost device that delivers a 2D and a 3D video stream has become available. In combination with 3D reconstruction software, such as reconstructMe [2], research groups have new opportunities for collecting data and for providing new challenges for the community. Future work will show, if the increased depth resolution of the Kinect 2 will be sufficient to serve as the input for a 3D ear recognition system. If his is the case, low-cost devices, such as the Kinect, could be interesting for manufacturers of ATMs, in order to obtain tamper-proof surveillance images.

Another interesting 3D imaging technique are light field cameras. Lytro has introduced a consumer model that offers the opportunity to change the focus to any arbitrary position of the image after the image has been captured. Raytrix, who is another vendor for light field cameras [3], offers a number of camera devices and software packages for calculating 3D representations from a single light field camera picture using the different foci in the image. However, we found that one of the more expensive Raytrix models is needed in order to get a reliable depth image.

Structure from motion is a noteworthy approach for computing 3D representations of an object based on a video that is showing the subject from different perspectives. Structure from motion is computationally expensive and needs dense textural details in order to provide good reconstructions of an object's surface structure. We have conducted a number experiments with structure from motion using mobiles phones and found that skin texture does not provide enough details for getting a sufficiently smooth and detailed representation of the object's surface. A description of the experimental setup and more details on the result are provided in Appendix D.

14.3 Beyond Forensics: New Applications for Ear Recognition

The next step towards a fully functional ear recognition system is to explore the limitations of ear recognition approaches in the literature in different scenarios. This thesis makes a start, by comparing the performance rate that could be achieved using an academic dataset and another dataset that is far more challenging. Academic datasets are mostly collected from college students in a very controlled scenario. At least one more challenging dataset is available (WPUTEDB), but it is rarely used because the authors want to keep up with their competitors.

Apart from forensic identification a possible use case for ear biometrics could be authentication for mobile phones. We can think of two different scenarios here. A first possibility is to automatically unlock the phone when answering a phone call. When the user moves the phone towards her ear, the frontal camera would take a picture and answer the call automatically, when the owner is verified. A second possible scenario is, to use ear prints in order to keep a minimum level of trust during the phone call. The contact-sensitive display would return a pattern of contact points that is dependent on the structure of the owner's ear. This pattern could be used for authenticating a subject while making a phone call. The cellphone could periodically evaluate the ear prints during the phone call and keep the phone unlocked as long as the owner using it.

[1] see http://www.microsoftstore.com/store/msusa/en_US/pdp/Kinect-for-Windows-v2-Sensor/productID.298810500

[2] see http://reconstructme.net/

[3] see http://www.raytrix.de/index.php/Cameras.html

Given that there would be a standard for the storage of ear template for passport documents, ear recognition systems could be used as an amendment to face recognition in automatic border control. Even though the ear might be occluded in different cultures, it could be used as an additional characteristic in all cases where it is visible. Current automatic border control systems (self-service gates) would, in principle, allow to capture profile images at during the passport verification process.

14.4 Conclusion

Ear recognition is an promising biometric characteristic with forensics as its main application. Our work shows that we can achieve a high recognition performance under laboratory conditions. Ear recognition systems can easily be integrated with existing face recognition systems for applications, where users are not actively interacting with the identification system. The uniqueness of the outer ear, as well as the differences between the left and the right ears, are valuable information that should be used in order to improve the performance in scenarios, where identification systems have to process half profile of full profile images.

In this work we provided an elaborate overview of existing work in ear recognition and addressed several open challenges in the field. We investigate the value of depth information for ear recognition during segmentation and recognition and propose two different algorithms for combining depth and texture to a composite descriptor for ear segmentation and recognition. We have also shown that normalization of ear images without the use of any further contextual or semantic information form is possible.

Further, we provide results on the robustness of different segmentation and texture-based recognition techniques to signal degradation. The possibility of fast identity retrieval in large datasets, is crucial for any biometric characteristic and has not been addressed for ear recognition yet. We investigate the possibilities of classification binning) and sequential search for ear recognition systems using texture features. Our results show that unsupervised classification of the ear is possible. In order to perform a sequential search, we can derive binary representations of texture descriptors and use them for a re-screening on larger databases. We show that the true-positive identification rate of a system using pre-screening is superior to the baseline system using exhaustive search.

The main application of ear recognition systems will remain to be forensics in the upcoming years. Similarly to other biometric characteristics, more applications are likely to emerge after automated ear recognition has become more established in forensics.

Part IV

Appendix

Appendix A

Ear Occlusion Study

A.1 Introduction

Between September 2012 and July 2013, we conducted a study on the visibility of the outer ear at Darmstadt central station. Our goal was to get an estimation of the average chance to obtain a clear view of the outer ear in a surveillance scenario.

During the time of the study, we counted the number persons with occluded ears, who were walking into the main building through the front door. Additionally to the counting, whether the ear is visible or occluded, we also made notes about the gender and the weather conditions outside. In total, we observed 5431 people.

We also differentiated between men and women and the type of the occlusion. For the occlusion type, we were using six categories, which are as follows:

- **Hair partly:** At most one third of the outer ear is occluded with hair
- **Hair full:** The ear is fully occluded with hair
- **Headdress/Hat:** The ear is occluded with a headdress or the person is wearing a hat with a brim.
- **Large earrings:** The person is wearing large earrings that either fully occlude the lobule or other parts of the ear.
- **Attached earrings:** The person is wearing earrings with an oval shape that are attached to the lobule (these earrings can cause errors in the segmentation process)
- **Earphones:** The person is wearing earphones that either occlude the concha (in-ear headphones) or full occlude the ear.

A.2 Occlusion per Gender

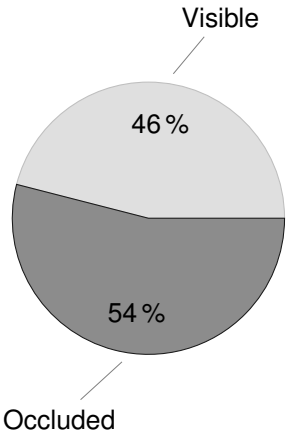


Figure A.1: Occlusion of outer ear for all subjects in the study

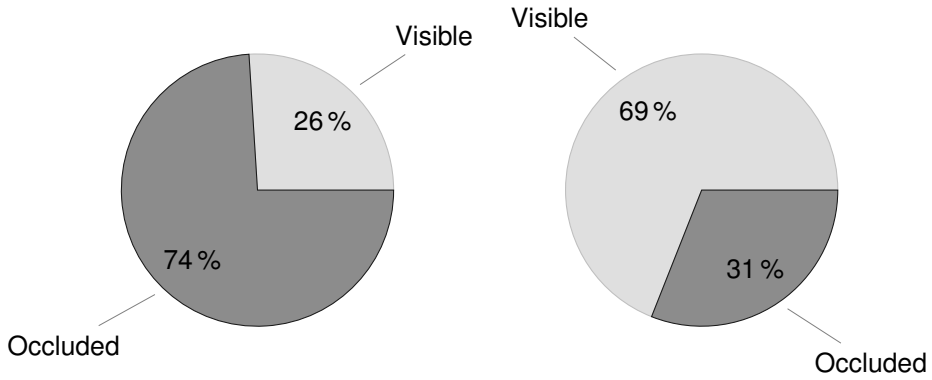


Figure A.2: Occlusion of outer ear for all women (left) and men (right) in the study and for all types of occlusions.

A.3 Occlusion Types per Gender

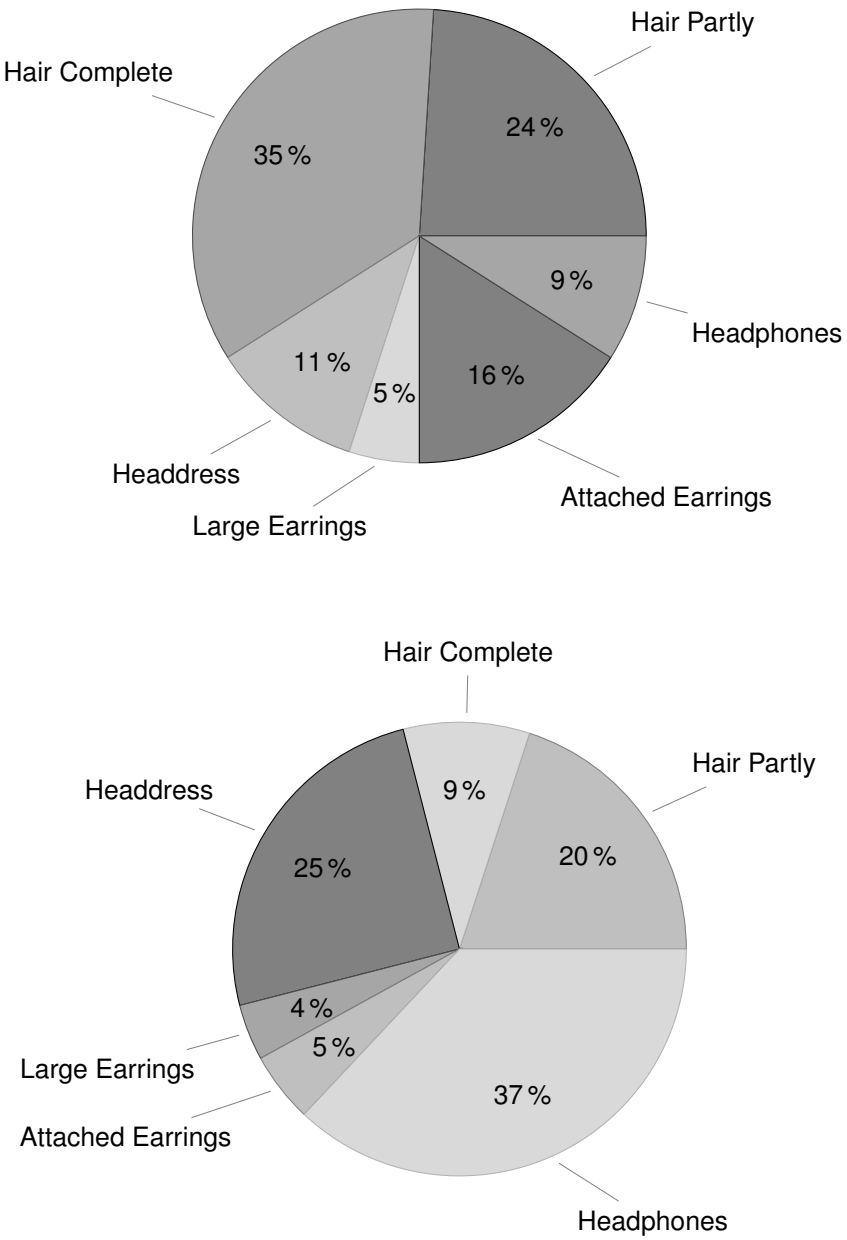


Figure A.3: Occlusion types for women (upper pie chart) and men (lower pie chart).

A.4 Impact of Environmental Conditions

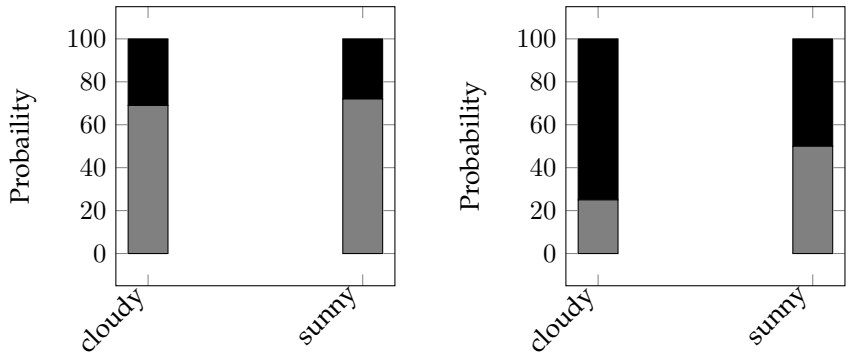


Figure A.4: Impact of weather conditions on occlusion of the outer ear. (gray = visible, black = occluded)

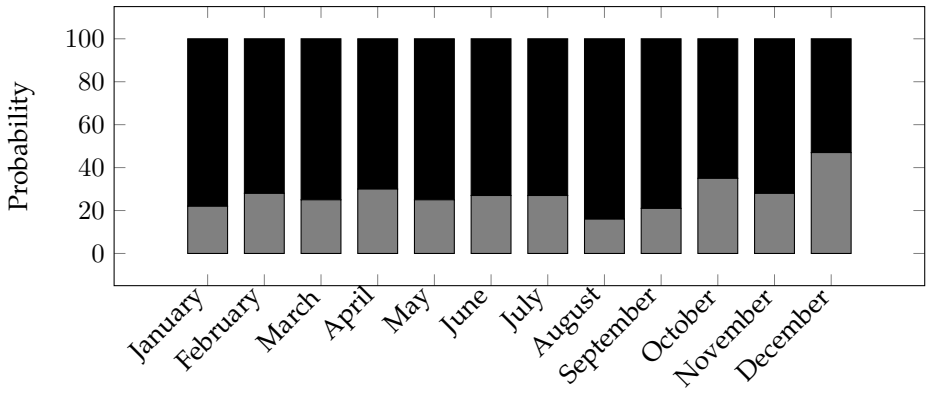


Figure A.5: Probability of observing an occluded ear per month for women. (gray = visible, black = occluded)

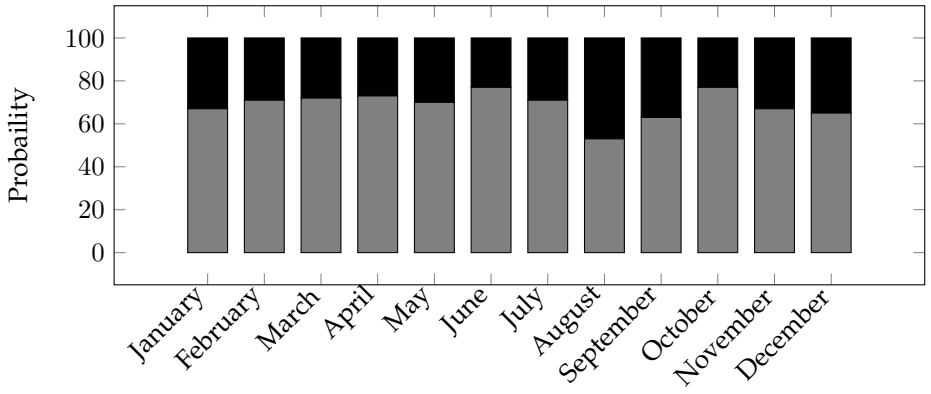


Figure A.6: Probability of observing an occluded ear per month for men.(gray = visible, black = occluded)

A.5 Conclusion

This study gives us a small indication of the average chance to obtain ear images in public venues. We are aware that our results are bias by several factors, such as the venue and the time of the day. We observe that the average probability of observing a visible ear is much larger for men than for women. The visibility slightly changes due to weather conditions.

It is important to remember that the total number of observations is relatively small and that more data is required to get a better impression about the visibility of the outer ear in public. We also restricted ourselves to one single venue, a train station, which certainly has an impact on our results. It is likely that the number of subjects wearing headphones could be smaller in a venue where people have to interact with each other. The number of people wearing hats and scarves is likely to be smaller inside a building (e.g. a shopping mall).

We finally note that all of our observations were made during work hours between 9:00 AM and 3:00PM. This implies hat we have observed many people who were heading to or coming back from their work place, university or school. It is likely that this also has an influence on the way how many subjects are dressed. Finally, the total number of people to be observed within a certain time frame was smaller in August (Summer break) than in all other months. It would be interesting to compare our observations with additional surveys that were conducted in the late evening.

Appendix B

Standardization and Common Terms in Biometrics

B.1 Generic Biometric System

The general components and work flow in a biometric system is described in the ISO/IEC 2382-37:2012 standard document [89]. In this thesis, we particularly focus on the signal processing subsystem, the matching subsystem and the decision subsystem.

We evaluate ear recognition systems in two different modes, which are the verification and the identification mode. Verification refers to an operational mode, where the presented images (also referred to as the probe image) is associated with an identity claim. The system verifies or falsifies this claim, meaning that it will make binary decision about whether or not, the claim is true. In identification mode, we only have a probe image without an identity claim. The system will return the most likely candidate identity that belongs to this images, based on the images in the database.

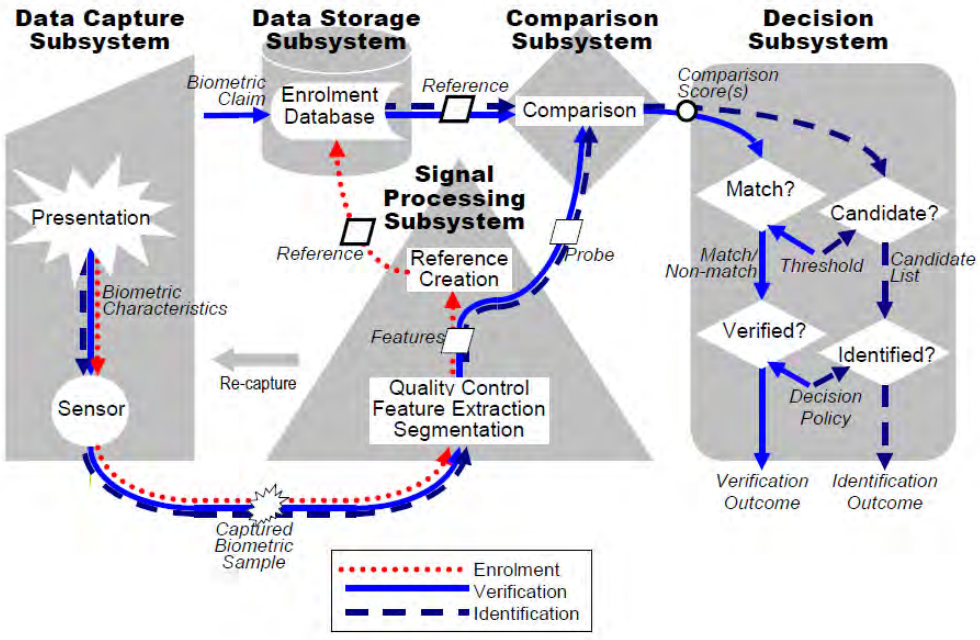


Figure B.1: A generic biometric system as defined in the ISO/IEC SC37 SD11 standard document [89]

B.2 Harmonized Vocabulary

The ISO-IEC standard [91] defines a number of terms that are well-established within the biometrics research community. Based on the components of a biometric system, as described in the previous section, we use the following terms within this thesis.

B.2.1 General Terms

When describing a biometric systems, we make use a number of standardized terms. These terms are defined in the ISO/IEC JTC SC37 SD11 [91] standard. The following list is an except from the standard. For a complete overview of the biometrics vocabulary, the reader is referred to the original standard document.

Probe: A biometric sample or biometric feature set input to an algorithm for use as the subject of biometric comparison to a biometric reference(s).

Reference: one or more stored biometric samples, biometric templates or biometric models attributed to a biometric data subject and used as the object for biometric comparison.

Features: Numbers or labels extracted from biometric samples and used for comparison.

Template: A set of stored biometric features comparable directly to probe biometric features.

Enrolment: The act of creating and storing a biometric enrolment data record in accordance with an enrolment policy.

Comparison: The estimation, calculation or measurement of similarity or dissimilarity between biometric probe (s) and biometric reference(s)

B.2.2 System Performance and Failures

When describing the performance of a biometric system, the standard proposes the following performance indicators. All of these definitions are taken from [91].

Failute to Enrol (FTE): Proportion of the population for whom the system fails to complete the enrolment process. This failure can occur anywhere during the capture process, segmentation process or the feature extraction process.

False Match Rate (FMR): Proportion of zero-effort impostor attempt samples falsely declared to match the compared non-self template. Note that this performance indicator only measure the algorithm performance and not the system performance.

False Non-Match Rate (FNMR): Proportion of genuine attempt samples falsely declared not to match the template of the same characteristic from the same user supplying the sample. Note that this performance indicator only measure the algorithm performance and not the system performance.

False Accept Rate (FAR): Proportion of verification transactions with wrongful claims of identity that are incorrectly confirmed. This performance indicator refers to the complete system, which means that it also incorporates the FTE.

False Reject Rate (FRR): Proportion of verification transactions with truthful claims of identity that are incorrectly denied. This performance indicator refers to the complete system, which means that it also incorporates the FTE.

Equal Error Rate (EER): An operational point where the FAR and FRR are equal. Often used for reporting the performance in verification mode.

Rank-n identification rate (R1 or IR): identification rate proportion of identification transactions by users enrolled in the system in which the user's correct identifier is among those returned for reporting the performance in verification mode. n denotes the maximum rank of the true positive in a sorted list of candidates.

Preselection Error (PSE): Error that occurs when the corresponding enrolment template is not in the pre-selected subset of candidates when a sample from the same biometric characteristic on the same user is given.

Penetration Rate (PEN): Measure of the average number of pre-selected templates as a fraction of the total number of templates.

Appendix C

An Update on Related Work in Ear Recognition Since 2012

This section is intended to provide an overview of selected contributions from other researchers to the field of ear recognition in from mid-2012 until the end of 2014. It puts recent work into the context of the state of the art, that is presented in the survey paper in Chapter 3. We follow the breakdown of work in the original survey paper and present approaches for segmentation, 2D ear recognition and 3D ear recognition separately.

C.0.3 Segmentation

Because many publicly available ear datasets contain cropped and party also normalized ear images, segmentation has moved out of the focus in many works. The only public database, that allows for segmentation experiments is the UND-J2 collection [202]. Lei at al. [112] showed that the Tree Structured Model approach by [222] can also be applied to ear detection. Additionally, the approach detects reliable landmarks with an average error between 4.5 pixels for UND-F and 5.5.pixels for UND-J2. Numerous segmentation methods have been proposed in the last decade that achieved a detection performance close to 100%, which brings us to the conclusion that the segmentation task on UND-J2 can be regarded as a solved problem.

We observe that the community started to apply existing segmentation techniques, such as in the work of Zhang et al. [216]. They detect the nose tip and then scan the silhouette for the ear pit region, which is the region with the farthest distance from the camera within the region of interest. As far as this can be judged from the paper, this approach is mostly the same as proposed earlier by Yan and Bowyer [202].

Jamil et al. [95] propose an ear recognition system that uses Biased normalized cut (BNC) for segmenting the ear region from a pre-cropped ROI (illustrated in Figure C.1). BNC is a graph-based technique where adjacent ear regions with similar illumination conditions are merged until a stopping condition is reached. The highest reported detection rate is 95%.

C.0.4 2D Ear recognition

As we already saw in Chapter 3, there is a number of public datasets available for testing ear recognition algorithms. The most popular dataset, however is still the UND-J2 dataset [201], followed by the IITK database [107].

Zhang et al. have conducted extensive research on Gabor Filters and their application to ear biometrics. Starting with [212], they have introduced an approach that utilizes multi-scale Gabor filter for ear images under different yaw poses. The compare randomly selected poses from the same person between 0 degrees and 60 degrees. The reported rank-1 recognition rate is larger than 95%. In a later publication Zhang et al. [213] extract fixed-length features using Gabor filters and reduce the dimensionality with non-negative sparse representation. Similar to this approach, Yuan and Mu [208] also use a bench of Gabor filters as local descriptors, but use KFDA for dimensionality reduction instead. Both approaches use the Euclidean distance between two feature vector representations in the feature subspace

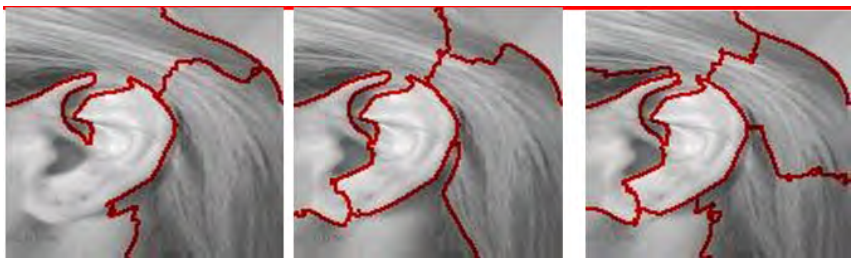


Figure C.1: Illustration of how biased normalized cuts [95]. BNC divides the image into adjacent homogeneous regions. We see that the ear is a clearly distinguishable region.

for comparison. Jamil et al. [95] feature extraction from the segmented ear region is also done with Log-Gabor filters. The output of the filter is quantized in order to obtain a binary representation. These so-called ear codes are compared using the hamming distance. The proposed system is tested with a custom database that contains 200 images from 50 subjects.

Landmark-based features have also been further pursued, as for example in [159]. Prakash and Gupta enhance the contrast of pre-cropped ear images and extract SURF features from multiple enrolment samples. A feature-level fusion creates a joint feature vector from these samples that contains all pairwise different feature points from the enrolment. A nearest neighbor classifier finally computes the Euclidean distances between matching key points.

Xu an Zheng [199] propose a landmark extraction method that is similar to the approaches of Choras [52]. They extract edges using the canny edge detector and then fit an axis through the ear that connects the lobule and the upper helix. Based on this, they compute aspect ratios that describe the shape of the ear and compare them by using a weighted Euclidean distance.

C.0.5 3D Ear recognition

Progress in 3D ear recognition is still driven by algorithms that have been developed around the UND-J2 dataset [201]. This dataset remains to the only public dataset that contains 2D and 3D ear images and consists of full profile images where only the head of a person is shown. Because of the lack of alternative data, most of the 3D ear recognition approaches in the literature are actually designed for depth images.

Zhang et al. [216] use the depth information directly after normalizing the input images and reduce the dimensionality with sparse representation. Wang and Mu [187] propose an ear recognition approach that uses automatically detected key points. The key points are detected by observing the variation of depth data within local image patches. Afterwards, the key point descriptors are used for initializing ICP for registration and comparison of a pair of 3D ear images. Another approach for extracting surface patches from depth images is described in [111]. Let et al. use the curvedness information and the shape index of local image patches for automatically detecting points on interest on the ear surface. The features consists of spatial information and a weighting factor. In the experiments, the proposed interest points descriptors outperforms similar approaches, such as spin images and a previous approach by the same working group, called SPHIS [221]. The reported performance is 93% rank-1 performance for spin images versus 97.4% rank-1 performance for the proposed method. A similar approach was already proposed earlier by Zhou and Cadavid [219]. After key point detection, Wang and Mu align two ear images using only the detected key points and ICP. The alignment error of ICP is finally used as the similarity score.

Although the usage of depth images enables the creation of computationally fast system, any information of the the different views of the ear in different poses is lost. Dimov

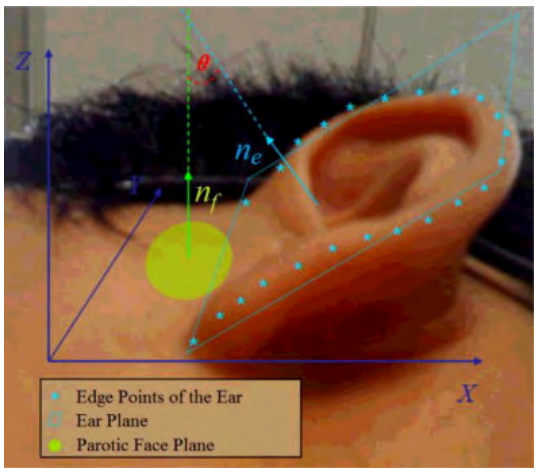


Figure C.2: Concept of ear pariodic feature extraction[117]. From the perspective show in the figure, we can determine the face and the ear plane and extract the angle between them.

et al. [61] propose to render a cleanly segmented 3D ear in different poses and then extract a set of appearance features form the rendered images. The performance rates are not very impressive and we think that the authors encounter similar problems than we did during our experiment in GES-3D (see Section 2.6). The idea to create a set of rendered images seem straight forward and promising, but also adds the requirement for an accurate pose estimation.

Cantoni et al. [43] propose the usage of extended Gaussian Images (EGI) for representing the structure of a 3D ear mesh without the need to create rendered images. The authors collected a set of ear models of 11 different subjects ans extracted a fixed length EGI histogram descriptor and provide a proof of concept implementation of their idea and also show the descriptor's robustness to uniform noise. Because the descriptor is a holistic representation of the full 3D object, it is invariant to any pose variations by design. On the other hand, the descriptor requires that bot, probe and reference data are 3D meshes, which is a requirement that is hardly fulfilled by any system n use today.

An interesting approach for indexing in large ear databases is described in [117] by Liu et al. who investigate the uniqueness of the angle between the ear and the face plane. A special experimental setup is designed, where 3D ear images of 250 subjects are collected from a special perspective that reveals the angle between the face and the ear (see Figure C.2). Liu et al. provide statistical evaluations of the uniqueness features and show that their features can reduce the search space in large datasets to 10% of the original database.

Appendix D

3D Face Reconstruction and Multimodal Person Identification from Video Captured Using Smartphone Camera

Abstract

In this paper, we propose a novel approach for reconstructing 3D face in real-life scenarios. Our main objective is to address the most challenging issue that involves reconstructing depth information from a video that is recorded from frontal camera of the smartphone. Such videos recorded using smartphones impose lot of challenges, such as motion blur, non-frontal perspectives and low resolution. This limits the applicability of state-of-the-art algorithms, which are mostly based on landmark detection.

This situation is addressed with the Scale-Invariant Feature Transformation (SIFT) followed by feature matching to generate consistent tracks. These tracks are further processed to generate a 3D point cloud using Point/Cluster based Multi-view stereo (PMVS/ CMVS). The usage of PMVS/CMVS will however fail to generate a dense 3D cloud points on the weak surfaces of face, such as cheeks, nose and forehead. This issue is addressed by multi-view reconstruction of these weakly supported surfaces using Visual-Hull. The effectiveness of our method is evaluated on a newly collected dataset, which simulates a realistic identification scenario using a smartphone.

D.1 Introduction

Face recognition has received substantial attention from both academia and industry for more than three decades. During the last decade, significant improvements in terms of accuracy were achieved with 2D face images that are captured in controlled scenarios [195]. In practice, however, face recognition still faces a number of challenges, such as pose variation, non-uniform illumination, expressions and occlusions. Recent improvements in the field of face recognition have made an attempt to address some of the above mentioned issues [2, 171].

Especially in surveillance applications, the usage of 3D models for compensating pose variations and non-uniform illumination has gained significant attention. In particular, a number of approaches has been introduced, that reconstruct 3D information from videos. Moreover, 3D face recognition is of paramount interest because of its various applications in the field of computer vision including face recognition, facial expression analysis, avatar, model based image coding and border control applications [2, 37].

However, the success of 3D face reconstruction depends on accurately inferring the depth information from a set of images in order to build a 3D face model. There are wide variety of approaches to build a 3D face model based on 2D images, namely: Active Appearance Model (AAM) [122, 22, 54], Parametrized Appearance Model (PAM) [31] and combined 2D and 3D active appearance model [197].

Even though these approaches appear to be promising, they exhibit drawbacks such as (1) the need for accurate landmark selection, (2) requirement of extensive training, or (3) the need for multiple and properly calibrated cameras. Recent improvements involve

creating an accurate 3D face reconstruction from video using either a single image [99, 142] or a single camera [75].

Among these two approaches, single camera based 3D face reconstruction is more appealing, especially for biometrics. The reason for this are (1) it allows a basic form of liveness detection, that makes the system robust against photo attacks. (2) it overcomes the additional computational burden in selecting the best frame from a video to carry out 3D face reconstruction from single image. (3) It also makes additional depth information that is hidden in the spatio-temporal dependence between single video frames explicit, such that it can be processed by the biometric system with the goal of becoming more accurate. (4) With the fact that the video also contains profile views of the person, we can implement a multimodal approach that involves the outer ear as an additional biometric characteristic.

Motivated by these advantages, we consider the problem of 3D face reconstruction using a frontal smartphone camera. More precisely, we use the spatio-temporal dependency of video frames that show a face from multiple perspectives, for computing depth information. The videos were captured by the subjects themselves with the frontal camera of a smartphone. For our test database, we use the Google Nexus S. The resolution of the frontal camera is 0.3 megapixels. For capturing video, the subject holds the smartphone before his face and starts the process of capturing. Then he turns his head from left to the right and back. Hence, the captured videos contain frontal, half profile an full profile yaw-poses with relatively stable roll and pitch. Using these videos, 3D reconstruction is

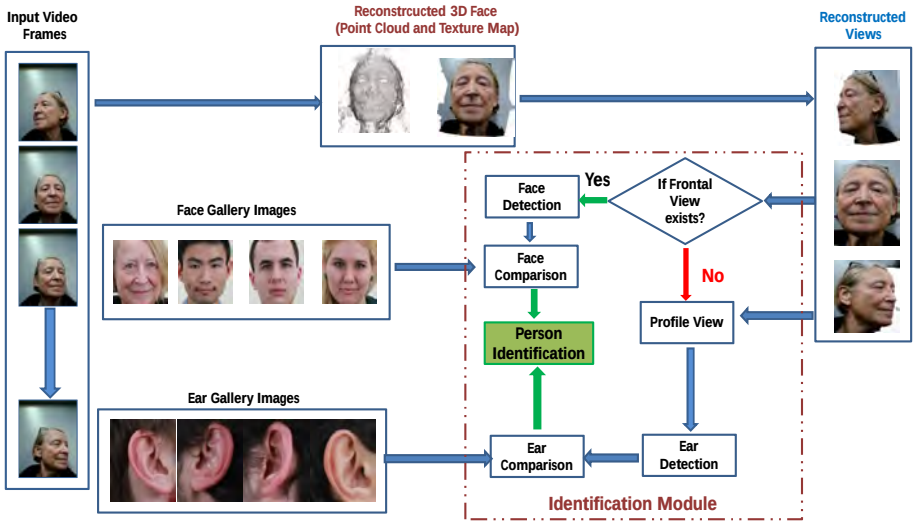


Figure D.1: Overview of the proposed person identification system with 3D face reconstruction from video

a challenging problem, because there are hardly any stable feature points that are visible throughout the video. Moreover, no ground truth information or training data is available for calibration. The videos also contain motion blur and varying scales, because the subject's hand slightly shakes during video capture and the distance between the smartphone and the camera changes throughout the video capturing process.

In this work, we present a system to quickly perform the 3D reconstruction from videos that are captured using the frontal camera of a smartphone. The proposed pipeline is derived largely from existing state of the art schemes. It employs SIFT feature extraction [195], Structure from Motion (SFM) algorithms [2] and Point/ Cluster based Multi-view stereo (PMVS/CMVS) [67]. Using these techniques, we obtain a dense cloud reconstruction of the 3D face.

Due to motion blur, high reflectivity and lack of sufficiently structured texture information, the dense reconstruction of the face from video results in weakly supported surfaces. We propose to address this issue by employing the multi-view reconstruction based on the visual-Hull [96] to accurately reconstruct the 3D face surface. To this extent, we employed the multi-view reconstruction algorithm proposed in [96] that can take the 3D cloud points from PMVS/CMVS and render the strong surface with rich texture information.

The proposed pipeline for this precise application has remarkable advantages, such as: (1) no need of landmark detection and tracking, (2) no need of camera calibration, (3) no need of training data (4) user convenience, because data will be acquired by the user using frontal camera of smartphone (Nexus S). The proposed reconstruction pipeline forms a new concept for multimodal identification on smartphone devices with high user convenience. In this work, we present the results of our first study on such a system, and also explore confronted shortcomings when using existing algorithms.

Thus the main contributions of our work can be summarized as follows: (1) Coupling the available techniques to form a new pipeline to address the application of 3D face reconstruction using a frontal camera of the smartphone (Nexus S). Further, we provide insights on the lessons learnt while building the whole pipeline for this precise application. (2) Extensive evaluation of the proposed pipeline on a newly collected database under realistic conditions and its use in 3D face reconstruction and person identification based on either face or ear. (3) We provide an outlook on possible future applications and research involving our reconstructed 3D models and the collected dataset. (4) To the best of our knowledge, this is the first work that addresses the new and exciting application area of 3D biometric authentication using a smartphone.

The rest of the paper is organized as follows: Section D.2 describes the proposed pipeline for 3D face reconstruction, Section D.3 describes the experimental procedure and results and Section D.4 draws conclusion and proposals for future research directions.

D.2 Proposed Scheme

Figure D.1 shows the block diagram of the proposed scheme (or methodology) for person identification based on 3D face reconstruction. Given the video, the proposed method begins with feature extraction and matching to accurately construct the tracks. In the next step, these tracks are processed using incremental SFM to compute the 3D cloud points. Since the constructed 3D point clouds fail to accurately render face surface, we propose to use the multi-view reconstruction algorithm based on Visual Hull [96] to accurately construct the 3D face surface and texture. We proceed further to perform the person identification using either face or ear biometric. To perform this, we first check whether we can detect frontal face from the 3D reconstructed face. If yes, we carry out the person identification based on the face, else we perform the person identification using ear that can be detected from the reconstructed 3D face profile. In the following section, we discuss each step involved in building our proposed scheme.

D.2.1 Feature extraction and matching

The commonly used feature extraction scheme for the 3D face reconstruction involves locating facial landmark points that outline the structure of eyes, nose, mouth, eyebrows and the facial boundary. However, accurate landmark detection and tracking requires [143] extensive training and robustness to various viewpoints (both full profile and half profile). These drawbacks will limit the applicability of the 3D face reconstruction schemes (like AAM, PAM, etc.) based on landmark point detection and tracking especially for our precise application of 3D face reconstruction from video captured using smartphone. Hence, we employed the Scale-Invariant Feature Transformation (SIFT) [195] to extract the features

from each frame of the recorded video. In this work, we employed the GPU implementation of SIFT by considering its speed and user friendly interface [195]. Here, the choice of SIFT feature extraction looks appealing for our precise application by considering its robustness against: (1) image resolution (as we are processing low resolution video with 0.3 megapixel) (2) various viewpoints (3) self-occlusion and (4) clutters.

In the next step, we carry out the feature matching between images based on the nearest neighbor search [2]. The obtained matches are then pruned and verified using RANSAC based estimation of the fundamental or essential matrix [179] [8]. Finally, we combine all matches into tracks to form a connected set of key points across multiple images. In order to prune out the inconsistent tracks, we perform the track selection that contains at least two key points for the next stage that involves in 3D cloud reconstruction.

D.2.2 3D surface reconstruction

After obtaining the consistent tracks, we proceed further to run the structure from motion algorithm for recovering a 3D location for each track. This can be carried out by using bundle adjustment, which minimizes the re-projection error between each tracks and its corresponding image [8]. In order to overcome the problem of SFM getting struck in a local minima [179], we carry out an incremental reconstruction procedure by adding one frame at a time. In this work, we employed the VisualSFM [196] an open source GUI that integrates three packages namely: SIFT on GPU (siftGPU), incremental SFM system and multi core bundle adjustment [196]. Further, VisualSFM runs very fast by exploiting the multi-core acceleration for feature extraction, matching and bundle adjustment. In addition, Visual SFM also integrates an interface to run both PMVS/CMVS for 3D dense reconstruction. Hence, given the video sequence, we run the VisualSFM to get the 3D dense reconstruction of the face surface.

Figure D.2: Illustration of 3D Clouds obtained using (a) PMVS/CMVS (b) Visual-Hull (c) after texture rendering

Figure D.2(a) shows the 3D cloud point obtained using PMVS/CMVS using VisualSFM. Here, it can be observed that, the use of PMVS/ CMVS will fail to construct the strong surfaces for the 3D face reconstruction. Thus, in order to overcome this problem, we employed the multi-view reconstruction algorithm based on Visual-Hull [96] to reconstruct the difficult face surfaces (specially cheeks, nose and ear) that are not sampled densely using PMVS/CMVS. The main idea of the visual hull algorithm involves in computing the free space support that detects the highly supported surface boundary and reconstruct the weak surface by optimizing the t-weights in an s-t graph [96]. Thus the employed visual hull algorithm [96] performs the multi view reconstruction by accepting the 3D cloud surfaces generated from PMVS/CMVS. Figure D.2(b) shows the multi-view reconstruction of the weak face surfaces especially, cheeks and nose region while Figure D.2(c) shows the final reconstructed image after texture rendering based on the point clouds [96].

D.2.3 Face/Ear based person identification

After accurately reconstructing the 3D head model from a video, we proceed further to identify the subject by comparing the frontal view of the reconstructed face with the enrolled samples. We are also addressing the 3D profile (half and/or full) face reconstruction. Hence, we are also interested to explore the contribution of ear recognition to the system's overall robustness. To this extent, we compare the outer ears, in cases where the accurate 3D frontal face reconstruction failed. In this way, we are making an attempt to explore the multi-modality from the reconstructed 3D structure to identify a person (or subject) based on either face or ear.

We collected the enrollment samples in a controlled illumination (or studio) environment using Canon EOS 550D DSLR camera. The usage of two different cameras and the fact that the enrollment images and the probe images were taken under different conditions, simulating real-life intra/inter-class variance between the images. For each subject, we captured one frontal view for face with a neutral expression and two full profile views that represent left and right ear. We then perform a series of pre-processing steps that include detecting and cropping a face region using Viola-Jones face detector [183].

Figure D.3: Examples of the enrolled samples (a) Enrolled samples (b) Corresponding face images

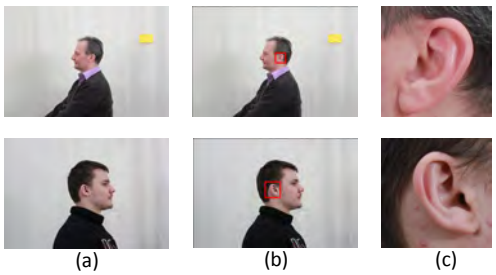


Figure D.4: Examples of enrolled Ear samples (a) Enrolled samples (b) Ear detection results (c) Corresponding Ear images

The detected face region is then resized to 120×120 pixels to reduce the computation cost. Finally, we apply a Gaussian filter ($\sigma = 2$) to remove noise high frequency noise. A similar procedure is followed for extracting the the ear region from the enrollment images. The ear detector was trained using OpenCV 2.4 with 2000 positive and 300 negative samples. The positive training images were taken from manually cropped images from the publicly available UND-J2 database [6], the IIT-Kanpur database [107] and the AMI ear database [70]. The negative training samples were chosen randomly from the INRIA person detection image set [57]. Figure D.3 shows the example of the enrolled samples and their corresponding processed face samples while Figure D.4 shows the enrolled and processed

ear samples. In order to perform the person identification, we compare the reconstructed 3D face with the enrolled sample of the subject, we first correct the pose of all reconstructed faces towards a frontal view to have yaw, pitch and roll values to zero. Then we project the head model to the 2D plane. In this work, we compare the frontal view from the reconstructed 3D face (after projecting to 2D) with the gallery sample (which is also a frontal view) using Sparse Representation Classifier (SRC) [194].

Figure D.5: Illustration of multi-view reconstruction using proposed scheme

A similar procedure is also followed to compare the ear detected from reconstructed 3D profile face with the enrolled sample of the ear. Hence, in this work, we compare either left or right (not both) ear with the corresponding left or right enrolled sample using Sparse Representation Classifier (SRC) [194]. In this work, we choose the SRC by considering its robustness and performance on both face [194] and ear [105] biometrics. Other existing classifiers such as SVM, Kernel Discriminant analysis and etc. can also be used at this stage. However, the study on these classifiers remains beyond the scope of this work which focuses mainly on the 3D face reconstruction using smartphone camera. The ear to compare is the first ear that is visible in the video stream (see next section for details).

D.3 Experimental Results

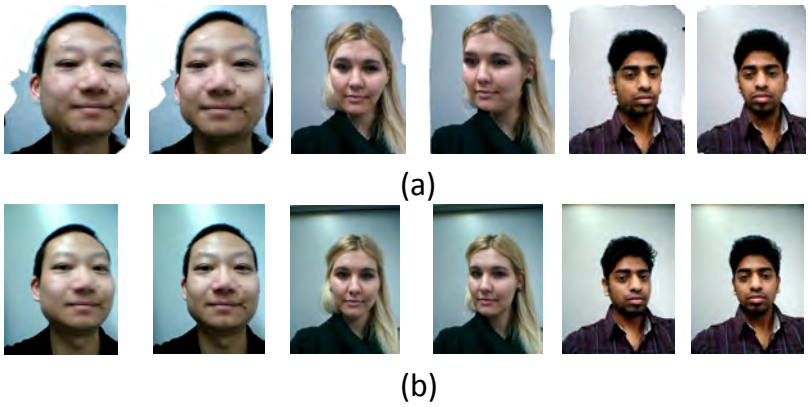


Figure D.6: Results for 3D reconstruction (a) 3D reconstruction (b) Corresponding video frame

In this section, we describe both qualitative and quantitative results of our proposed pipeline for fast and accurate 3D face reconstruction and person identification using the frontal camera of Google Nexus S smartphone. First we discuss our new face video dataset. We then present the qualitative results of the proposed 3D face reconstruction scheme

and finally, we presents the quantitative results of the multimodal person identification, which includes face and ear recognition. Our newly collected dataset consists of 25 sub-

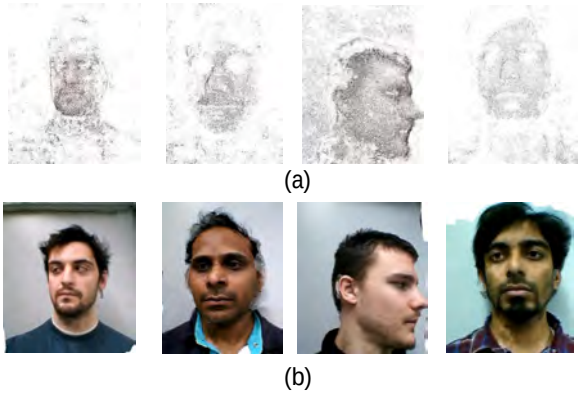


Figure D.7: Illustration of (a) 3D Cloud points (2) Corresponding 3D image rendered using proposed scheme

jects among which there are 19 males and 6 females. Each subject is asked to rotate his head by holding the smartphone in hand. This particular setting for the data collection appears quite appealing from the user's point of view, however it poses large number of challenges for the accurate 3D face recognition. The recorded video from the frontal camera of Google Nexus S smartphone is of moderate quality with 0.3 megapixel VGA, which makes it further interesting and challenging. Every subject is asked to record a video of himself, where he rotates his head from frontal pose to the left, the right and back to the frontal pose. The subjects could chose by themselves, whether they turn their heads left or right first. Each video recording takes between 2 -3 seconds. We then decompose each of these videos into its corresponding frames before performing the 3D face reconstruction. Thus, the collected dataset has any one of the following rotation patterns for each subject:(1) rotation of head starting from frontal face till left full profile face (2) rotation of head starting from frontal face till right profile face (3) rotation of head starting from left full profile till right full profile or vice versa.

Figure D.6 shows the 3D reconstruction of the half profile face. As it can be clearly seen, our method manages to reconstruct the shape of the cheeks and the outer ear shape. Moreover, we see that, one can simulate self-occlusion by adjusting the reconstructed model to the according pose.

Figure D.7 illustrates the qualitative results of the proposed pipeline for 3D reconstruction that shows both 3D point cloud and corresponding 3D frontal face after texture rendering. Figure D.5 illustrates the full profile reconstruction of the rotated face. These results further justify the efficiency of the proposed pipeline for 3D reconstruction. The proposed scheme has shown a remarkable performance in terms of reconstructing 3D frontal view on 17 subjects out of 25 subjects in total. For the remaining 8 subjects, the proposed scheme can accurately reconstruct the 3D profile view. One possible reason is the fact that the profile can be reconstructed, even though the reconstruction of the frontal view failed due to lack of frontal face images. This can happen, if the subjects have moved their head too fast.

Figure D.8 shows the ear detection and localization from the reconstructed 3D profile view. The person identification is carried out by comparing the projected ear (to 2D) with its corresponding gallery samples. In order to provide the comprehensive comparison, we manually choose the profile sample from the video frame and compare the same with the gallery samples.

Table D.1 shows the performance of the face recognition on using frontal view from the

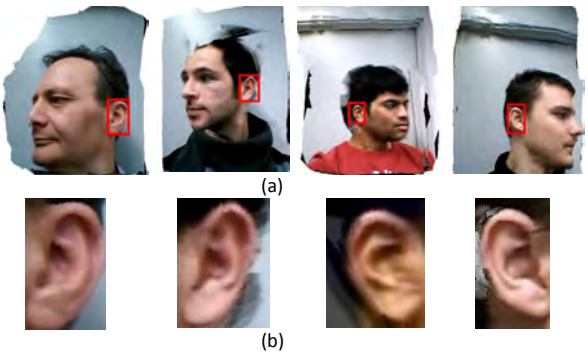


Figure D.8: Illustration Ear detection from reconstructed 3D profile face image (a) Ear detection results (b) Corresponding segmented ear image (2D)

3D face and frontal view of the face from a recorded video. In this work, all the results are presented in terms of closed-set identification rate (rank 1) that is obtained by comparing $1 : N$ subjects in the dataset, therefore, a higher value of identification rate corresponds to better accuracy of the algorithm. For each subject, we acquired three enrolment samples under controlled illumination conditions, namely: frontal face with neutral expression, full profile with left ear and full profile with right ear. Thus, for the comparison, we first check for the existence of a frontal face and, if we are able to detect it, we compare the reconstructed frontal view to the frontal view from the gallery. Otherwise, we compare the detected and extracted ear from the reconstructed profile view with its corresponding ear from the enrolment samples (either left or right).

We have presented the results of the proposed scheme for both unimodal and multimodal recognition using face and ear. For the comprehensive comparison, we have also compared the quantitative results of the proposed 3D reconstruction scheme with the 2D video frame. It can be observed from the Table D.1 that, the proposed scheme shows the best performance among the possible combinations, with an identification rate of 82.25% on the frontal face and 75% on the ear. Furthermore, the proposed method also shows an outstanding performance of 80% when using both face and ear images, by indicating an improvement of 12% as compared to the 2D video frame. The degraded performance of the 2D video can be attributed to the presence of noises due to motion blur, illumination etc. The multimodal results reported in this paper are obtained by associating the correctly identified subjects from both frontal face and ear modality (OR rule). These quantitative results further justify the applicability and efficiency of the proposed scheme. In this work,

Table D.1: Qualitative performance of the proposed scheme

Modality	Methods	No. of subjects	Identification Rate (%)
Unimodal (Face / Ear)	3D face	17	82.35
	Video face	17	70.58
	3D ear	8	75.00
	Video ear	8	62.50
Multimodal (Face and Ear)	3D face + 3D ear	25	**80.00**
	Video face + Video ear	25	**68.00**

we carried out the 2D comparison because of the lack of 3D enrollment data. Even though one can argue that, the use of 2D video can also provide equally good results, but, it is well

know that, the 2D based biometric systems are highly vulnerable for attacks (eg. Photo attacks) and hence fails to prevents spoofing. While the use of 3D is very difficult to spoof as compared to that of 2D [15]. Thus, the proposed scheme opens up more avenues for secured applications using smartphones. Kindly refer http://youtu.be/VQTGh5AjM38 for additional comprehensive results.

D.4 Conclusion

We have presented a new scheme for 3D reconstruction and recognition from videos that are acquired using the frontal camera of the Google Nexus S smartphone. This scheme can be seen as a proof of concept for future authentication systems, which are more accurate and more robust against spoofing than existing approaches. The proposed scheme is derived from the available techniques, which are coupled in a novel way to achieve a significantly better performance in 3D reconstruction and accurate person recognition.

The proposed scheme stands different when compared to the state of art schemes by overcoming the requirement of landmark points detection, exhaustive training, camera calibration and multiple cameras. Further, the proposed scheme is extensively evaluated on a newly collected database comprising of 25 subjects by considering real-life scenarios. The proposed scheme has shown good performance in reconstructing 3D frontal faces from 17 different subjects out of 25.

Our further work involves improving the system to achieve better reconstruction and matching speed, selection of key frames for improving the overall speed and also on exploring the surface reconstruction scheme to build more robust and accurate system. Moreover, we expect further performance improvement by including both ears into the pipeline.

Bibliography

[1] A. Abate, M. Nappi, D. Riccio, and S. Ricciardi. Ear recognition by means of a rotation invariant descriptor. In *18th International Conference on Pattern Recognition, ICPR 2006.*, volume 4, pages 437 –440, 0-0 2006. 43, 45

[2] A. F. Abate, M. Nappi, D. Riccio, and G. Sabatino. 2d and 3d face recognition: A survey. *Pattern Recognition Letters*, 28(14):1885–1906, 2007. 183, 184, 186

[3] A. Abaza and M. A. F. Harrison. Ear recognition: a complete system. In *SPIE 8712, Biometric and Surveillance Technology for Human and Activity Identification*, 2013. 124, 132, 136

[4] A. Abaza, C. Hebert, and M. Harrison. Fast learning ear detection for real-time surveillance. In *Fourth IEEE International Conference on Biometrics: Theory Applications and Systems (BTAS 2010)*, pages 1 –6, September 2010. 38, 40

[5] A. Abaza and A. . Ross. Towards understanding the symmetry of human ears: A biometric perspective. In *Fourth IEEE International Conference on Biometrics: Theory Applications and Systems (BTAS)*, 2010. 32, 54

[6] A. Abaza, A. Ross, C. Hebert, M. A. F. Harrison, and M. S. Nixon. A survey on ear biometrics. *ACM Computing Surveys (CSUR)*, 45(2):22, 2013. 3, 33, 40, 41, 142, 143, 187

[7] M. Abdel-Mottaleb and J. Zhou. Human ear recognition from face profile images. In D. Zhang and A. Jain, editors, *Advances in Biometrics*, volume 3832 of *Lecture Notes in Computer Science*, pages 786–792. Springer Berlin / Heidelberg, 2005. 43, 45

[8] S. Agarwal, N. Snavely, I. Simon, S. M. Seitz, and R. Szeliski. Building Rome in a day. In *IEEE 12th International Conference on Computer Vision*, pages 72–79. IEEE, 2009. 186

[9] T. Ahonen, A. Hadid, and M. Pietikainen. Face description with local binary patterns: Application to face recognition. *Pattern Analysis and Machine Intelligence*, 28:2037–2041, 2006. 10, 115

[10] T. Ahonen, E. Rahtu, V. Ojansivu, and J. Heikkila. Recognition of blurred faces using local phase quantization. In *International Conference on Pattern Recognition*, 2008. 21, 97, 104, 115, 119, 136, 139, 143, 145

[11] H. A. Al Nizami, J. P. Adkins-Hill, Y. Zhang, J. R. Sullins, C. McCullough, S. Canavan, and L. Yin. A biometric database with rotating head videos and hand- drawn face sketches. In *Proceedings of the 3rd IEEE international conference on Biometrics: Theory, applications and systems*, BTAS'09, pages 38–43, Piscataway, NJ, USA, 2009. IEEE Press. 37

[12] M. Alaraj, J. Hou, and T. . Fukami. A neural network based human identification framework using ear images. In *TENCON 2010 - 2010 IEEE Region 10 Conference*, November 2010. 44, 48

[13] I. Alberink and A. Ruifrok. Performance of the fearid earprint identification system. *Forensic Science International*, 166(2-3):145 – 154, 2007. 33, 66

[14] L. Alvarez, E. Gonzalez, and L. Mazorra. Fitting ear contour using an ovoid model. In *39th Annual 2005 International Carnahan Conference on Security Technology (CCST '05)*, October 2005. 38, 41

[15] A. Anjos and S. Marcel. Counter-measures to photo attacks in face recognition: a public database and a baseline. In *International Joint Conference on Biometrics (IJCB)*, pages 1–7. IEEE, 2011. 191

[16] S. Ansari and P. Gupta. Localization of ear using outer helix curve of the ear. In *International Conference on Computing: Theory and Applications*, pages 688 –692, March 2007. 38, 40, 56

[17] B. Arbab-Zavar and M. Nixon. On shape-mediated enrolment in ear biometrics. In G. Bebis, R. Boyle, B. Parvin, D. Koracin, N. Paragios, S.-M. Tanveer, T. Ju, Z. Liu, S. Coquillart, C. Cruz-Neira, T. Mller, and T. Malzbender, editors, *Advances in Visual Computing*, volume 4842 of *Lecture Notes in Computer Science*, pages 549–558. Springer Berlin / Heidelberg, 2007. 38, 39, 41

[18] B. Arbab-Zavar and M. Nixon. Robust log-gabor filter for ear biometrics. In *International Conference on Pattern Recognition (ICPR)*, December 2008. 38, 43, 45

[19] B. Arbab-Zavar, M. Nixon, and D. Hurley. On model-based analysis of ear biometrics. In *First IEEE International Conference on Biometrics: Theory, Applications, and Systems, 2007. (BTAS 2007)*, pages 1 –5, sept. 2007. 42, 43, 46

[20] S. Attarchi, K. Faez, and A. Rafiei. A new segmentation approach for ear recognition. In J. Blanc-Talon, S. Bourennane, W. Philips, D. Popescu, and P. Scheunders, editors, *Advanced Concepts for Intelligent Vision Systems*, volume 5259 of *Lecture Notes in Computer Science*, pages 1030–1037. Springer Berlin / Heidelberg, 2008. 38, 40, 56

[21] G. Badrinath and P. Gupta. Feature level fused ear biometric system. In *Seventh International Conference on Advances in Pattern Recognition (ICAPR)*, pages 197 –200, feb. 2009. 44, 46

[22] S. Baker, I. Matthews, and J. Schneider. Automatic construction of active appearance models as an image coding problem. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 26(10):1380–1384, 2004. 183

[23] H. Bay, T. Tuytelaars, and L. V. Gool. Surf: Speeded up robust features. In *Proceedings of the 9th European Conference on Computer Vision*, 2006. 47

[24] S. Belongie, J. Malik, and J. Puzicha. Shape context: A new descriptor for shape matching and object recognition. In *Neural Information Processing Systems Conference (NIPS)*, pages 831–837, December 2000. 161

[25] D. E. Bernstein and J. D. Jackson. The daubert trilogy in the states. *Law and Economics Working Paper Series*, 2004. 5

[26] A. Bertillon. *La Photographie Judiciaire: Avec Un Appendice Sur La Classification Et L'Identification Anthropometriques*. Gauthier-Villars, Paris, 1890. 3, 32, 124, 132

[27] P. Besl and N. McKey. A method for registration of 3d shapes. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, volume 14, pages 239–256. IEEE, December 1992. 17

[28] P. J. Besl. *Surfaces in Range Image Understanding*. Spinger, 1988. 57, 69, 82, 105

[29] B. Bhanu and H. Chen. *Human Ear Recognition by Computer*, chapter 3D Ear Detection from Side Face Range Images, pages 21 – 59. Springer, 2008. 56

[30] S. Billeb, C. Rathgeb, M. Buschbeck, H. Reininger, and K. Kasper. Efficient two-stage speaker identification based in universal background models. In *Intl. Conference of the Biometrics Special Interest Group*, 2014. 114, 115, 159

[31] M. J. Black and A. D. Jepson. Eigentracking: Robust matching and tracking of articulated objects using a view-based representation. *International Journal of Computer Vision*, 26(1):63–84, 1998. 183

[32] N. Boodoo-Jahangeer and S. Baichoo. LBP-based ear recognition. In *Bioinformatics and Bioengineering*, 2013. 143

[33] Bundeskriminalamt. Qualitätstatndardbeschreibung zur fertigung und anzeige von dididigital erkennungsdienstlichen llichtbilder in inpol. 14

[34] Bundeskriminalamt, 2012. 4

[35] M. Burge and W. Burger. *Ear Biometrics*, chapter 13, pages 273–285. Springer US, 1998. 43

[36] X. Burgos-Artizzu, P. Perona, and P. Dollár. Robust face landmark estimation under occlusion. In *ICCV*, 2013. 93, 156

[37] C. Busch and M. Brauckmann. Towards a more secure border control with 3d face recognition. In *Norsk informasjonssikkerhetskonferanse (NISK)*, 2012. 183

[38] C. Busch, A. Pflug, X. Zhou, M. Dose, M. Brauckmann, J. Helbig, A. Opel, P. Neugebauer, K. Leowski, H. Sieber, and O. Lotz. Multi-biometrische gesichtserkennung. In *13. Deutscher IT-Sicherheitskongress*, May 2013. 9

[39] J. Bustard and M. . Nixon. Toward unconstrained ear recognition from two-dimensional images. *Systems, Man and Cybernetics, Part A: Systems and Humans*, 40:486, April 2010. 44, 46

[40] S. Cadavid and M. Abdel-Mottaleb. 3d ear modeling and recognition from video sequences using shape from shading. In *19th International Conference on Pattern Recognition (ICPR)*, pages 1 –4, December 2008. 50, 52

[41] S. Cadavid, M. Mahoor, and M. Abdel-Mottaleb. Multi-modal biometric modeling and recognition of the human face and ear. In *IEEE International Workshop on Safety, Security Rescue Robotics (SSRR)*, pages 1 –6, November 2009. 51

[42] J. Canny. A computational approach to edge detection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, PAMI-8(6):679 –698, nov. 1986. 69

[43] V. Cantoni, D. Dimov, and A. Nikolov. 3d ear analysis by an egi representation. In V. Cantoni, D. Dimov, and M. Tistarelli, editors, *Biometric Authentication*, Lecture Notes in Computer Science, pages 136–150. Springer International Publishing, 2014. 181

[44] R. Cappelli, M. Ferrara, and D. Maltoni. Minutia cylinder-code: A new representation and matching technique for fingerprint recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(12):2128–2141, Dec 2010. 115

[45] R. Cappelli, M. Ferrara, and D. Maltoni. Fingerprint indexing based on minutia cylinder-code. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(5):1051–1057, May 2011. 115

[46] K. Chang, K. W. Bowyer, S. Sarkar, and B. Victor. Comparison and combination of ear and face images in appearance-based biometrics. *IEEE Transactions in Pattern Anallysis and Machine Intelligene*, 25:1160–1165, September 2003. 43, 48

[47] H. Chen and B. Bhanu. Contour matching for 3d ear recognition. In *Proceedings of the Seventh IEEE Workshop on Applications of Computer Vision (WACV/MOTION)*, 2005. 38, 39, 50, 51, 52

[48] H. Chen and B. Bhanu. Shape model-based 3d ear detection from side face range images. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops (CVPR)*, page 122, June 2005. 38, 39, 155

[49] H. Chen and B. Bhanu. Human ear recognition in 3d. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(4):718 –737, April 2007. 38, 39, 50, 67, 68, 80

[50] M. Choras. Image feature extraction methods for ear biometrics–a survey. In *Computer Information Systems and Industrial Management Applications, 2007. CISIM '07. 6th International Conference on*, pages 261 –265, june 2007. 33, 80

[51] M. Choras. Image pre-classification for biometrics identification systems. In J. Peja? and K. Saeed, editors, *Advances in Information Processing and Protection*, pages 361–370. Springer US, 2008. 56

[52] M. Choras. Perspective methods of human identification: Ear biometrics. *Opto-Electronics Review*, 16:85–96, 2008. 42, 43, 47, 180

[53] A. Conan Doyle. *A Chapter on Ears*. Solis Press, 2012. 3

[54] T. F. Cootes, G. V. Wheeler, K. N. Walker, and C. J. Taylor. Coupled-view active appearance models. In *Proceedings of the British machine vision conference*, volume 1, pages 52–61, 2000. 183

[55] A. Cummings, M. Nixon, and J. Carter. A novel ray analogy for enrolment of ear biometrics. In *Fourth IEEE International Conference on Biometrics: Theory Applications and Systems (BTAS)*, September 2010. 38, 39, 41

[56] N. Dalal. *Finding Peple in Images and Vides*. PhD thesis, Institut National Polytechnique de Grenoble, 2006. 83, 115

[57] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition*, volume 1, pages 886–893, June 2005. 98, 103, 105, 127, 136, 143, 145, 187

[58] N. Damer and B. Fuhrer. Ear recognition using multi-scale histogram of oriented gradients. In *Intelligent Information Hiding and Multimedia Signal Processing*, 2012. 102, 143

[59] J. Daugman. How iris recognition works. *Circuits and Systems for Video Technology, IEEE Transactions on*, 14(1):21–30, Jan 2004. 115, 117

[60] M. De Marsico, N. Michele, and D. . Riccio. Hero: Human ear recognition against occlusions. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, page 178, August 2010. 44, 46

[61] D. Dimov and V. Cantoni. Appearance-based 3d object approach to human ears recognition. In V. Cantoni, D. Dimov, and M. Tistarelli, editors, *Biometric Authentication*, Lecture Notes in Computer Science, pages 121–135. Springer International Publishing, 2014. 181

[62] P. Dollar, P. Welinder, and P. Perona. Cascaded pose regression. In *Computer Vision and Pattern Recognition*, 2010. 10, 21, 93, 94, 103, 143, 156

[63] J. Dong and Z. Mu. Multi-pose ear recognition based on force field transformation. In *Second International Symposium on Intelligent Information Technology Application (IITA)*, volume 3, pages 771 –775, December 2008. 44, 45

[64] T. C. Faltemier, K. W. Bowyer, and P. J. Flynn. Rotated profile signatures for robust 3d feature detection. In *Automatic Face and Gesture Recognition*, 2008. 35, 67, 74, 75, 76, 81, 85, 126, 135

[65] H. Freeman. On the encoding of arbitrary geometric configurations. *IRE Transactions on Electronic Computer*, 10:260–268, 1961. 161

[66] D. Frejlichowski and N. Tyszkiewicz. The west pomeranian university of technology ear database a tool for testing biometric algorithms. In A. Campilho and M. Kamel, editors, *Image Analysis and Recognition*, volume 6112 of *Lecture Notes in Computer Science*, pages 227–234. Springer Berlin / Heidelberg, 2010. 35, 127

[67] Y. Furukawa and J. Ponce. Accurate, dense, and robust multiview stereopsis. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(8):1362–1376, 2010. 184

[68] R. Gadde. Iris indexing and ear classification. Master's thesis, West Virginia University, 2012. 143

[69] J. Gentile, N. Ratha, and J. Connell. An efficient, two-stage iris recognition system. In *Biometrics: Theory, Applications, and Systems, 2009. BTAS '09. IEEE 3rd International Conference on*, pages 1–5, Sept 2009. 114, 115

[70] E. Gonzalez, L. Alvarez, and L. Mazorra. Normalization and feature extraction on ear images. In *International Carnahan Conference on Security Technology*, 2012. 92, 103, 107, 147, 187

[71] M. Grgic, K. Delac, and S. Grgic. Scface – surveillance cameras face database. *Multimedia Tools Appl.*, 51(3):863–879, February 2011. 36

[72] Y. Guo and Z. i. W. Xu. Ear recognition using a new local matching approach. In *International Conference on Image Processing*, 2008. 44, 47, 102, 143

[73] L. Gutierrez, P. Melin, and M. Lopez. Modular neural network integrator for human recognition from ear images. In *The 2010 International Joint Conference on Neural Networks (IJCNN)*, October 2010. 44, 49

[74] F. Hao, J. Daugman, and P. Zielinski. A fast search algorithm for a large fuzzy database. *IEEE Trans. Information Forensics and Security,*, 3:203–212, 2008. 115, 142

[75] T. Hara, H. Kubo, A. Maejima, and S. Morishima. Fast-accurate 3d face model generation using a single video camera. In *Pattern Recognition (ICPR), 2012 21st International Conference on*, pages 1269–1272. IEEE, 2012. 184

[76] P. Hardeep, P. B. Swadas, and M. Joshi. A survey on techniques and challenges in image super resolution reconstruction. *International Journal of Computer Science and Mobile Computing*, 2013. 25

[77] D. Hartung, A. Pflug, and C. Busch. Vein pattern recognition using chain codes, spacial information and skeleton fusing. In *GI-Sicherheit*, March 2012. 9

[78] X. He and N. Yung. Corner detector based on global and local curvature properties. *Optical Engineering*, 47(5), 2008. 61

[79] A. Hoogstrate, H. V. D. Heuvel, and E. Huyben. Ear identification based on surveillance camera images. *Science & Justice*, 41(3):167 – 172, 2001. 3, 33, 92, 124, 132

[80] D. J. Hurley, M. S. Nixon, and J. N. Carter. Force field energy functionals for image feature extraction. *Image and Vision Computing*, 20(5-6):311 – 317, 2002. 42, 43

[81] A. V. Iannarelli. *Ear identification*. Paramount Publishing Company, 1964. 3, 32, 56, 66, 80, 92, 102, 132, 141, 142, 143, 159

[82] M. Ibrahim, M. Nixon, and S. Mahmoodi. The effect of time on ear biometrics. In *International Joint Conference on Biometrics (IJCB)*, pages 1 –6, October 2011. 32, 66

[83] R. Imhofer. Die bedeutung der ohrmuschel für die feststellung der identität. *Archiv für die Kriminologie*, 26:150–163, 1906. 3, 32

[84] S. Islam, M. Bennamoun, and R. Davies. Fast and fully automatic ear detection using cascaded adaboost. In *Applications of Computer Vision, 2008. WACV 2008. IEEE Workshop on*, pages 1 –6, jan. 2008. 38, 67, 68

[85] S. Islam, M. Bennamoun, A. Mian, and R. Davies. A fully automatic approach for human recognition from profile images using 2d and 3d ear data. In *Proceedings of 3DPVT - the Fourth Internatinoal Symposium on 3D Data Processing, Visualization and Transmission*, 2008. 50, 51, 52

[86] S. Islam, R. Davies, M. Bennamoun, and A. Mian. Efficient detection and recognition of 3d ears. *International Journal of Computer Vision*, 95:52–73, 2011. 40, 51, 52

[87] S. M. Islam, R. Davies, and M. Mian, A. S.an Bennamoun. A fast and fully automatic ear recognition approach based on 3d local surface features. In *Proceedings of the 10th International Conference on Advanced Concepts for Intelligent Vision Systems*, ACIVS '08, pages 1081–1092, Berlin, Heidelberg, 2008. Springer-Verlag. 50

[88] S. M. S. Islam, M. Bennamoun, R. Owens, and R. Davies. Biometric approaches of 2d-3d ear and face: A survey. In T. Sobh, editor, *Advances in Computer and Information Sciences and Engineering*, pages 509–514. Springer Netherlands, 2008. 33

[89] ISO/IEC 2382-37:2012 Biometrics. *ISO/IEC 2382-37:2012 General Biometric System*. International Organization for Standardization, 2012. 7, 14, 17, 175

[90] ISO/IEC JTC1 SC37 Biometrics. *ISO/IEC 19795-1:2006. Information Technology – Biometric Performance Testing and Reporting – Part 1: Principles and Framework*. International Organization for Standardization and International Electrotechnical Committee, Mar. 2006. 119

[91] ISO/IEC JTC1 SC37 Biometrics. *ISO/IEC JTC 1/SC 37 N 3663 Working Draft: Harmonized Biometric Vocabulary*. International Organization for Standardization, may 2010. 176

[92] ISO/IEC TC JTC1 SC37 Biometrics. *ISO/IEC 19795-1:2006. Information Technology – Biometric Performance Testing and Reporting – Part 1: Principles and Framework*. International Organization for Standardization and International Electrotechnical Committee, Mar. 2006. 135, 146

[93] A. Jain, A. Ross, and S. Prabhakar. An introduction to biometric recognition. *Circuits and Systems for Video Technology, IEEE Transactions on*, 14(1):4–20, Jan 2004. 114

[94] A. K. Jain, S. Prabhakar, and L. Hong. A multichannel approach to fingerprint classification. *IEEE Trans. Pattern Anal. Mach. Intell.*, 21(4):348–359, Apr. 1999. 115

[95] N. Jamil, A. AlMisreb, and A. A. Halin. Illumination invariant ear authentication. In *Conference on Robot PRIDE - Medical and Rehabilitation Robotics and Instrumentation*, 2014. 179, 180

[96] M. Jancosek and T. Pajdla. Multi-view reconstruction preserving weakly-supported surfaces. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3121–3128. IEEE, 2011. 185, 186

[97] E. Jeges and L. Mt. Model-based human ear localization and feature extraction. *ICMED*, 1(2):101–112, 2007. 38, 39, 42, 43, 47

[98] J. Kannala and E. Rahtu. Bsif: Binarized statistical image features. In *IEEE International Conference on Pattern Recognition*, pages 1363–1366, 2012. 104, 115, 119

[99] I. Kemelmacher-Shlizerman and R. Basri. 3d face reconstruction from a single image using a single reference face shape. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 33(2):394–405, 2011. 184

[100] R. Khorsandi and M. Abdel-Mottaleb. Gender classification using 2-D ear images and sparse representation. In *Workshop on Applications of Computer Vision,*, 2013. 143

[101] T. F. Kiely. *Forensic Evidence: Science and the Criminal Law, Second Edition*, chapter Ear Impressions, pages 368–370. CRC Press, 2005. 5

[102] D. Kisku, H. Mehrotra, P. Gupta, and J. Sing. Sift-based ear recognition by fusion of detected keypoints from color similarity slice regions. In *International Conference on Advances in Computational Tools for Engineering Applications (ACTEA)*, pages 380 –385, July 2009. 44, 46, 102

[103] J. Klontz and A. Jain. A case study on unconstrained facial recognition using the boston marathon bombings suspects. *Computer*, 46(11), 2013. 3

[104] A. Kong, D. Zhang, and M. Kamel. A survey of palmprint recognition. *Pattern Recognition*, 42(7):1408 – 1418, 2009. 114, 115

[105] A. Kumar and T.-S. T. Chan. Robust ear identification using sparse representation of local texture descriptors. *Pattern Recognition*, 2012. 188

[106] A. Kumar, M. Hanmandlu, M. Kuldeep, and H. Gupta. Automatic ear detection for online biometric applications. In *Third National Conference on Computer Vision, Pattern Recognition, Image Processing and Graphics (NCVPRIPG)*, pages 146 –149, December 2011. 38, 41, 44

[107] A. Kumar and C. Wu. Automated human identification using ear imaging. *International Conference on Pattern Recognition*, 45(3):956–968, March 2012. 35, 44, 45, 103, 107, 147, 179, 187

[108] H.-K. Lammi. Ear biometics. Technical report, Lappeenranta University of Technology, Department of Information Technology, 2004. 33

[109] P. Lancaster and K. Salkauskas. Surfaces generated by moving least squares. *Mathematics of Computation*, 37:1–18, 1981. 103

[110] J. Lei, J. Zhou, and M. Abdel-Mottaleb. Gender classification using automatically detected and aligned 3d ear range data. In *Biometrics (ICB), 2013 International Conference on*, pages 1–7, June 2013. 3

[111] J. Lei, J. Zhou, and M. Abdel-Mottaleb. A novel shape-based interest point descriptor (sip) for 3d ear recognition. In *International Conference on Image Processing (ICIP)*, pages 4176–4180, 2013. 180

[112] J. Lei, J. Zhou, M. Abdel-Mottaleb, and X. You. Detection, localization and pose classification of ear in 3d face profile images. In *International Conference on Image Processing (ICIP)*, 2013. 179

[113] E. Levina and P. J. Bickel. Maximum likelihood estimation of intrinsic dimension. In *Neural Information Processing Systems*, 2004. 145

[114] H. Liu and D. Liu. Improving adaboost ear detection with skin-color model and multi-template matching. In *3rd IEEE International Conference on Computer Science and Information Technology (ICCSIT)*, volume 8, pages 106 –109, July 2010. 38

[115] H. Liu and J. Yan. Multi-view ear shape feature extraction and reconstruction. In *Third International IEEE Conference on Signal-Image Technologies and Internet-Based System (SITIS)*, pages 652 –658, December 2007. 43, 48, 51

[116] H. Liu and D. Zhang. Fast 3d point cloud ear identification by slice curve matching. In *3rd International Conference on Computer Research and Development (ICCRD)*, page 224, March 2011. 50, 51

[117] Y. Liu, B. Zhang, and D. Zhang. Ear-parotic face angle: A unique feature for 3d ear recognition. *Pattern Recognition Letters*, 2014. 181

[118] G. D. Lowe. Object recognition from local scale-invariant features. In *IEEE International Conference on Computer Vision (ICCV 1999)*, volume 2, pages 1150 –1157. IEEE Computer Society, 1999. 46

[119] L. Lu, Z. Xiaoxun, Z. Youdong, and J. Yunde. Ear recognition based on statistical shape model. In *First International Conference on Innovative Computing, Information and Control*, pages 353–356, 2006. 43, 48, 102

[120] S. Maity and M. Abdel-Mottaleb. 3d ear segmentation and classification through indexing. *IEEE Transactions on Information Forensics and Security*, PP(99):1–1, 2014. 159

[121] D. Maltoni, D. Maio, A. Jain, and S. Prabhakar. *Handbook of Fingerprint Recognition*. Springer-Verlag, 1st edition, 2009. 142

[122] I. Matthews and S. Baker. Active appearance models revisited. *International Journal of Computer Vision*, 60(2):135–164, 2004. 183

[123] L. Meijerman. *Inter- and Intra-Individual Variation in Earprints*. PhD thesis, University of Leiden, 2006. 5

[124] L. Meijerman, S. Sholl, F. D. Conti, M. Giacon, C. van der Lugt, A. Drusini, P. Vanezis, and G. Maat. Exploratory study on classification and individualisation of earprints. *Forensic Science International*, 140(1):91 – 99, 2004. 3, 32

[125] L. Meijerman, C. Van Der Lugt, and G. J. Maat. Cross-sectional anthropometric study of the external ear. *Journal of Forensic Sciences*, 52(2):286–293, 2007. 3, 5, 32, 80

[126] A. Mhatre, S. Palla, S. Chikkerur, and V. Govindaraju. Efficient search and retrieval in biometric databases. In *SPIE Defense and Securiry Symposium*, 2005. 159

[127] S. Mika, G. Ratsch, J. Weston, B. Scholkopf, and K. Muller. Fisher discriminant analysis with kernels. In *Neural Networks for Signal Processing IX, 1999. Proceedings of the 1999 IEEE Signal Processing Society Workshop.*, pages 41–48, 1999. 104

[128] B. Moreno, A. Sanchez, and J. Velez. On the use of outer ear images for personal identification in security applications. In *IEEE 33rd Annual 1999 International Carnahan Conference on Security Technology*, pages 469 – 476, August 2002. 43

[129] Z. Mu, L. Yuan, Z. Xu, D. Xi, and S. Qi. Shape and structural feature based ear recognition. In S. Li, J. Lai, T. Tan, G. Feng, and Y. Wang, editors, *Advances in Biometric Person Authentication*, volume 3338 of *Lecture Notes in Computer Science*, pages 311–364. Springer Berlin / Heidelberg, 2005. 43, 47

[130] R. Mukherjee and A. Ross. Indexing iris images. In *International Conference on Pattern Recognition*, 2008. 142

[131] R. Mukundan, S. Ong, and P. Lee. Image analysis by tchebichef moments. *IEEE Transactions on Image Processing*, 10(9):1357 –1364, September 2001. 48

[132] L. Nanni and A. Lumini. A multi-matcher for ear authentication. *Pattern Recognition Letters*, 28:2219–2226, December 2007. 43, 49

[133] I. Naseem, R. Togneri, and M. Bennamoun. Sparse representation for ear biometrics. In G. Bebis, R. Boyle, B. Parvin, D. Koracin, P. Remagnino, F. Porikli, J. Peters, J. Klosowski, L. Arns, Y. Chun, T.-M. Rhyne, and L. Monroe, editors, *Advances in Visual Computing*, volume 5359 of *Lecture Notes in Computer Science*, pages 336–345. Springer Berlin / Heidelberg, 2008. 44, 49

[134] National Institute for Standardization (NIST). *Best Practice Recommendation for the Capture of Mugshots*, September 1997. 14, 15

[135] NY Daily News. Mezuzah arsonist snagged by an ear thanks to facial recognition technology, April 2013. 124

[136] F. C. P. O. of Germany. Police crime statistics - annual report. Technical report, Federal Criminal Police Office of Germany, 2013. 4

[137] N. A. Ogale. A survey of techniques for human detection from video. *Survey, University of Maryland*, 2006. 125, 133

[138] T. Ojala, M. Pietikainen, and T. Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 24:971–987, 2002. 97, 104, 127, 143, 144

[139] V. Ojansivu and J. Heikkil. Blur insensitive texture classification using local phase quantization. In *Image and Signal Processing*. Springer Berlin Heidelberg, 2008. 136

[140] G. Oxley. *Forensic Human Identification*, chapter Facial Recognition and Imagery Analysis, pages 257 – 270. CRC Press, 2007. 5, 6

[141] M. Ozuysal, M. Calonder, V. Lepetit, and P. Fua. Fast keypoint recognition using random ferns. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(3):448–461, 2010. 94

[142] M. Pamplona Segundo, L. Silva, and O. R. P. Bellon. Improving 3d face reconstruction from a single image using half-frontal face poses. In *19th IEEE Conference on Image Processing (ICIP)*, pages 1797–1800. IEEE, 2012. 184

[143] U. Park and A. Jain. 3d model-based face recognition in video. *Advances in Biometrics*, pages 1085–1094, 2007. 185

[144] G. Passalis, I. Kakadiaris, T. Theoharis, G. Toderici, and T. Papaioannou. Towards fast 3d ear recognition for real-life biometric applications. In *IEEE Conference on Advanced Video and Signal Based Surveillance (AVSS 2007)*, pages 39 –44, September 2007. 50, 51

[145] F. Perronnin and J.-L. Dugelay. Clustering face images with application to image retrieval in large databases. In *Proc. SPIE Conf. Biometric Technology for Human Identification II*, volume 5779, pages 256–264, 2005. 142

[146] A. Pflug, P. Back, and C. Busch. Towards an ear detection that is robust against rotation. In *The 46th Annual IEEE International Carnahan Conference on Security Technology*, 2012. 9, 10, 11, 67, 68, 79, 95

[147] A. Pflug and C. Busch. Ear biometrics: a survey of detection, feature extraction and recognition methods. *Biometrics, IET*, 1(2):114–129, 2012. 8, 9, 31, 92, 103, 115, 124, 132

[148] A. Pflug and C. Busch. Segmentation and normalization of human ears using cascaded pose regression. In *19th Nordic Conference on Secure IT Systems*, 2014. 9, 10, 21, 91, 119

[149] A. Pflug, D. Hartung, and C. Busch. Feature extraction from vein images using spatial information and chain codes. *Information Security Technical Report*, 17(12):26 – 35, 2012. Human Factors and Bio-metrics. 9

[150] A. Pflug, P. Paul, and C. Busch. A comparative study on texture and surface descriptors for ear biometrics. In *Proceedings of the International Carnahan Conference on Security Technology (ICCST)*, 2014. 9, 101, 115

[151] A. Pflug, C. Rathgeb, U. Scherhag, and Bus. Binarization of histogram models: An application to efficient biometric identification. In *In Proceedings of International Conference on Cybernetics*, 2015. 9, 11, 113

[152] A. Pflug, A. Ross, and C. Busch. 2d ear classification based on unsupervised clustering. In *In Proceedings of International Joint Conference on Biometrics (IJCB)*, 2014. 9, 12, 115, 141

[153] A. Pflug, J. Wagner, C. Rathgeb, and C. Busch. Impact of severe signal degradation on ear recognition performance. In *Proceedings of Biometrics & Forensics & De-identification and Privacy Protection (BiForD)*, 2014. 9, 11, 131

[154] A. Pflug, A. Winterstein, and C. Busch. Ear detection in 3D profile images based on surface curvature. In *Proceedings of IEEE International Conference on Intelligent Information Hiding and Multimedia Signal Processing*, 2012. 8, 10, 55, 67, 68, 69, 81

[155] A. Pflug, A. Winterstein, and C. Busch. Robust localization of ears by feature level fusion and context information. In *Proceedings of the International Conference on Biometrics (ICB)*, 2013. 8, 10, 65

[156] S. Prakash and P. Gupta. An efficient ear recognition technique invariant to illumination and pose. *Telecommunication Systems Journal, special issue on Signal Processing Applications in Human Computer Interaction*, 30:38–50, 2011. 36, 44, 47

[157] S. Prakash and P. Gupta. An efficient technique for ear detection in 3d: Invariant to rotation and scale. In *The 5th IAPR International Conference on Biometrics (ICB)*, 2012. 38, 40, 67, 68, 80, 155

[158] S. Prakash and P. Gupta. An effient ear localization technique. *Image and Vision Computing*, 30:38–50, 2012. 38, 40

[159] S. Prakash and P. Gupta. An efficient ear recognition technique invariant to illumination and pose. *Telecommunication Systems*, 52(3):1435–1448, 2013. 180

[160] P. Promila and V. Laxmi. Palmprint matching using lbp. In *Computing Sciences (ICCS), 2012 International Conference on*, pages 110–115, Sept 2012. 115

[161] K. Pun and Y. Moon. Recent advances in ear biometrics. In *Automatic Face and Gesture Recognition, 2004. Proceedings. Sixth IEEE International Conference on*, pages 164 – 169, May 2004. 33

[162] R. Raghavendra, K. Raja, A. Pflug, B. Yang, and C. Busch. 3d face reconstruction and multimodal person identification from video captured using a smartphone camera. In *Proceedings of the 13th IEEE Conference on Technologies for Homeland Security (HST)*, November 2013. 9

[163] M. Rahman, R. Islam, N. I. Bhuiyan, B. Ahmed, and A. Islam. Person identification using ear biometrics. *International Journal of The Computer, The Internet and Management*, 15:1–8, August 2007. 43, 47

[164] K. Ramesh and K. Rao. Pattern extraction methods for ear biometrics - a survey. In *World Congress on Nature Biologically Inspired Computing (NaBIC)*, pages 1657 –1660, December 2009. 33

[165] R. Raposo, E. Hoyle, A. Peixinho, and H. Proenca. Ubear: A dataset of ear images captured on-the-move in uncontrolled conditions. In *Computational Intelligence in Biometrics and Identity Management (CIBIM), 2011 IEEE Workshop on*, pages 84 –90, april 2011. 37

[166] N. Ratha, K. Karu, S. Chen, and A. Jain. A real-time matching system for large fingerprint databases. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 18(8):799–813, Aug 1996. 115

[167] A. Ross and R. Mukherjee. Augmenting ridge curves with minutiae triplets for fingerprint indexing. In *SPIE Biometrics*, 2007. 142

[168] P. J. Rouseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:5365, 1987. 148

[169] A. Sana, P. Gupta, and R. Purkait. Ear biometrics: A new approach. In *Advances in Pattern Recognition*, 2007. 43, 46

[170] C. Sanderson, A. Bgdeli, T. Shan, S. Chen, E. Berglund, and B. C. Lovell. Intelligent CCTV for Mass Transport Security: Challenges and Opportunities for Video and Face Processing. *Electronic Letters on Computer Vision and Image Analysis*, 6:30–41, 2007. 124

[171] A. Scheenstra, A. Ruifrok, and R. Veltkamp. A survey of 3d face recognition methods. In *Audio-and Video-Based Biometric Person Authentication*, pages 325–345. Springer, 2005. 183

[172] B. Scholkopf, A. Smola, and K.-R. Mueller. Kernel principal component analysis. In *Advances in Kernel Methods - Support Vector Learning*, pages 327–352. MIT Press, 1999. 98, 104

[173] C. Sforza, G. Grandi, M. Binelli, D. G. Tommasi, R. Rosati, and V. F. Ferrario. Age- and sex-related changes in the normal human ear. *Forensic Science International*, 187(1-3):110.e1 – 110.e7, 2009. 32, 125, 134

[174] H.-C. Shih, C. Ho, H. Chang, and C.-S. Wu. Ear detection based on arc-masking extraction and adaboost polling verification. In *Fifth International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP)*, volume ear detection, pages 669 –672, September 2009. 38

[175] H. Sieber. Das Gesichtserkennungssystems (GES) im Bundeskriminalamt. Presentation, June 2011. 4

[176] P. Singh and R. Purkait. Observations of external earan indian study. *HOMO - Journal of Comparative Human Biology*, 60(5):461 – 472, 2009. 32, 66

[177] H. Skibbe and M. Reisert. Circular fourier-hog features for rotation invariant object detection in biomedical images. In *IEEE International Symposium on Biomedical Imaging*, 2012. 161

[178] L. Spreeuwers. Fast and accurate 3d face recognition. *International Journal of Computer Vision*, 93(3):389–414, 2011. 160

[179] P. Torr and A. Zisserman. Robust computation and parametrization of multiple view relations. In *Computer Vision, 1998. Sixth International Conference on*, pages 727–732. IEEE, 1998. 186

[180] B. Victor, K. Bowyer, and S. Sarkar. An evaluation of face and ear biometrics. In *16th International Conference on Pattern Recognition (ICPR)*, volume 1, pages 429 – 432 vol.1, 2002. 43, 48

[181] T. Vikram, K. Chidananda Gowda, D. Guru, and S. Urs. Face indexing and retrieval by spatial similarity. In *Image and Signal Processing*, 2008. 142

[182] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages I–511 – I–518 vol.1, 2001. 19, 82, 127

[183] P. Viola and M. J. Jones. Robust real-time face detection. *International Journal of Computer Vision*, 57(2):137 – 154, May 2004. 10, 187

[184] V. Štruc and N. Pavešić. The complete gabor-fisher classifier for robust face recognition. *EURASIP J. Adv. Signal Process*, 2010:31:1–31:13, Feb. 2010. 138

[185] J. Wagner. Influence of image compression on ear biometrics. Master's thesis, Hochschule Darmstadt, 2014. 157, 158

[186] J. Wagner, A. Pflug, C. Rathgeb, and C. Busch. Effects of severe signal degradation on ear detection. In *Proc.of 2nd Int'l Workshop on Biometrics and Forensics (IWBF14)*, 2014. 9, 11, 123, 132, 133

[187] K. Wang and Zhi-Chun. A 3d ear recognition method based on auricle structural feature. *International Journal of Computer Science Issues (IJCSI)*, 10(5), 2013. 180

[188] S.-z. Wang. An improved normalization method for ear feature extraction. *International Journal of Signal Processing, Image Processing and Pattern Recognition*, 6:49 – 56, 2013. 92

[189] X. Wang and W. Yuan. Human ear recognition based on block segmentation. In *International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC)*, pages 262 –266, October 2009. 44, 46

[190] X. Wang and W. Yuan. Gabor wavelets and general discriminant analysis for ear recogniton. In *8th World Congress on Intelligent Control and Automation (WCICA)*, page 6305, August 2010. 44, 45

[191] X.-q. Wang, H.-y. Xia, and Z.-l. Wang. The research of ear identification based on improved algorithm of moment invariants. In *Third International Conference on Information and Computing (ICIC)*, page 58, July 2010. 44, 46

[192] Y. Wang, Z. chun Mu, and H. Zeng. Block-based and multi-resolition methods for ear recognition using walelste transform and uniform local binary patterns. In *19th International Conference on Pattern Recognition (ICPR)*, pages 1 –4, December 2008. 44

[193] Z.-q. Wang and X.-d. Yan. Multi-scale feature extraction algorithm of ear image. In *International Conference on Electric Information and Control Engineering (ICEICE)*, page 528, April 2011. 44, 47

[194] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma. Robust face recognition via sparse representation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(2):210–227, 2009. 188

[195] C. Wu. SiftGPU: A GPU implementation of scale invariant feature transform (SIFT). `http://cs.unc.edu/~ccwu/siftgpu`, 2007. 183, 184, 185, 186

[196] C. Wu. Visualsfm: A visual structure from motion system. `http://ccwu.me/vsfm/`, 2011. 186

[197] J. Xiao, S. Baker, I. Matthews, and T. Kanade. Real-time combined 2d+ 3d active appearance models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2. IEEE Computer Society; 1999, 2004. 183

[198] Z. Xie and Z. Mu. Ear recognition using lle and idlle algorithm. In *19th International Conference on Pattern Recognition (ICPR)*, pages 1 –4, December 2008. 44, 48

[199] Y. Xu and W. Zeng. Ear recognition based on centroid and spindle. *Procedia Engineering*, 29(0):2162 – 2166, 2012. 2012 International Workshop on Information and Electronics Engineering. 180

[200] P. Yan and K. Bowyer. A fast algorithm for icp-based 3d shape biometrics. In *Fourth IEEE Workshop on Automatic Identification Advanced Technologies*, pages 213 – 218, October 2005. 50, 51, 52

[201] P. Yan and K. Bowyer. Biometric recognition using 3D ear shape. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 29:1297 – 1308, August 2007. 35, 38, 40, 50, 51, 56, 62, 67, 69, 74, 75, 76, 77, 80, 81, 82, 85, 95, 103, 107, 126, 135, 147, 179, 180

[202] P. Yan and K. W. Bowyer. An automatic 3d ear recognition system. In *3DPVT'2006*, pages 326–333, 2006. 118, 126, 135, 179

[203] H. Yang and I. Patras. Sieving regression forest votes for facial feature detection in the wild. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 1936–1943, Dec 2013. 92

[204] M. Yaqubi, K. Faez, and S. Motamed. Ear recognition using features inspired by visual cortex and support vector machine technique. In *International Conference on Computer and Communication Engineering (ICCCE)*, pages 533 –537, May 2008. 44, 49

[205] A. Yazdanpanah and K. Faez. Normalizing human ear in proportion to size and rotation. In *Emerging Intelligent Computing Technology and Applications*, volume 5754 of *Lecture Notes in Computer Science*, pages 37–45. Springer, 2009. 92

[206] L. Yuan, Z. chun Mu, Y. Zhang, and K. Liu. Ear recognition using improved non-negative matrix factorization. In *18th International Conference on Pattern Recognition (ICPR)*, volume 4, pages 501 –504, 2006. 43

[207] L. Yuan and Z. Mu. Ear recognition based on 2d images. In *First IEEE International Conference on Biometrics: Theory, Applications, and Systems (BTAS)*, pages 1 –5, September 2007. 43, 48

[208] L. Yuan and Z. Mu. Ear recognition based on gabor features and kfda. *The Scientific World Journal*, 2014, March 2014. 179

[209] L. Yuan and Z.-C. Mu. Ear detection based on skin-color and contour information. In *International Conference on Machine Learning and Cybernetics*, volume 4, pages 2213 –2217, August 2007. 38, 41

[210] T. Yuizono, Y. Wang, K. Satoh, and S. Nakayama. Study on individual recognition for ear images by using genetic local search. In *Proceedings of the 2002 Congress on Processing Scociety of Japan (IPSJ) Kyushu Chapter Symposium*, pages 237–242, 2002. 43, 49

[211] H. Zeng, J.-Y. Dong, Z.-C. Mu, and Y. Guo. Ear recognition based on 3d keypoint matching. In *IEEE 10th International Conference on Signal Processing (ICSP)*, page 1694, December 2010. 50, 52, 102

[212] B. Zhang, H. Mu, Z. andZeng, and H. H. Ear recognition based on gabor scale information. In *International Conference on Wavelet Analysis and Pattern Recognition*, 2013. 179

[213] B. Zhang, Z. Mu, H. Zeng, and S. Luo. Robust ear recognition via nonnegative sparse representation of gabor orientation information. *The Scientific World Journal*, 2014. 179

[214] D. Zhang, Z. Guo, G. Lu, D. Zhang, and W. Zuo. An online system of multispectral palmprint verification. *Transaction on Instrumentation and Measurement*, 59(2):480–490, Feb 2010. 118

[215] H. Zhang and Z. Mu. Compound structure classifier system for ear recognition. In *IEEE International Conference on Automation and Logistics*, pages 2306 –2309, September 2008. 44, 48

[216] L. Zhang, Z. Ding, H. Li, and Y. Shen. 3d ear identification based on sparse representation. *PLoS ONE*, 9(4), 2014. 179, 180

[217] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld. Face recognition: A literature survey. *ACM Comput. Surv.*, 35(4):399–458, Dec. 2003. 114

[218] J. Zhou, S. Cadavid, and M. . Abdel-Mottaleb. Histograms of categorized shapes for 3d ear detection. In *International Conference on Biometrics: Theory Applications and Systems*, November 2010. 10, 38, 41, 56, 67, 68, 81, 86, 88, 103, 155

[219] J. Zhou, S. Cadavid, and M. Abdel-Mottaleb. A computationally efficient approach to 3d ear recognition employing local and holistic features. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 98 –105, June 2011. 50, 51, 102, 180

[220] J. Zhou, S. Cadavid, and M. Abdel-Mottaleb. Exploiting color sift features for 2d ear recognition. In *18th IEEE International Conference on Image Processing (ICIP)*, pages 553 –556, Sepmtember 2011. 102

[221] J. Zhou, S. Cadavid, and M. Abdel-Mottaleb. An efficient 3-d ear recognition system employing local and holistic features. *Information Forensics and Security, IEEE Transactions on*, 7(3):978–991, June 2012. 180

[222] X. Zhu and D. Ramanan. Face detection, pose estimation and landmark localization in the wild. In *Computer Vision and Pattern Recognition*, 2012. 21, 179

[223] K. Zuiderveld. *Graphics Gems IV*, chapter Contrast Limited Adaptive Histogram Equalization, pages 474–485. Academic Press, 1994. 103, 135, 143