Ear-based biometric authentication

by

Aviwe Kohlakala

Thesis presented in partial fulfilment of the requirements for the degree of Master of Science (Applied Mathematics) in the Faculty of Science at Stellenbosch University

Supervisor: Dr J. Coetzer

April 2019

The financial assistance of the Ball Family towards this research is hereby acknowledged. Opinions expressed and conclusions arrived at, are those of the author and are not necessarily to be attributed to the Ball Family.

Declaration

By submitting this thesis electronically, I declare that the entirety of the work contained therein is my own, original work, that I am the sole author thereof (save to the extent explicitly otherwise stated), that reproduction and publication thereof by Stellenbosch University will not infringe any third party rights and that I have not previously in its entirety or in part submitted it for obtaining any qualification.

Name: ....Aviwe Kohlakala...........

Date: ....April 2019...........

Abstract

In this thesis novel semi-automated and fully automated ear-based biometric authentication systems are proposed. Within the context of the semi-automated system, a region of interest (ROI) that contains the entire ear shell is manually specified by a human operator. However, in the case of the fully automated system the ROI is automatically detected using a suitable convolutional neural network (CNN), followed by morphological post-processing. The purpose of the CNN is to classify sub-images as either foreground (part of the ear shell) or background (homogeneous skin, jewellery, or hair). Independent of the ROI-detection procedure, each grey-scale input image, in its entirety, is subjected to Gaussian smoothing, followed by edge detection through an appropriate Canny-filter, and morphological edge dilation. The detected ROI serves as a mask for retaining only those edges associated with prominent contours of the ear shell. Features are subsequently extracted from each binary contour image using the discrete Radon transform (DRT). The aforementioned features are normalised in such a way that they are translation, rotation and scale invariant. A Euclidean distance measure is employed for the purpose of feature matching. Ear-based authentication is finally achieved by constructing a ranking verifier. Exhaustive experiments are conducted on two large international datasets. It is assumed that only one reference ear is available for each individual enrolled into the system. An experimental protocol is adopted that appropriately partitions the respective datasets based on ears that belong to training, validation, ranking and evaluation individuals. It is demonstrated that the proficiency of the novel systems developed in this thesis compares favourably to those of existing systems.

Uittreksel

In hierdie tesis word nuwe semi- en vol-outomatiese oor-gebaseerde biometriese verifiëringstelsels voorgestel. Binne die konteks van die semi-automatiese stelsel word 'n fokusgebied (FG), wat die hele oorskulp bevat, deur 'n menslike operateur gespesifiseer. In die geval van die vol-outomatiese stelsel word bogenoemde FG egter outomaties deur 'n geskikte konvolusie-neuraalnetwerk (KNN) gevind, gevolg deur morfologiese na-verwerking. Die doel van die KNN is om sub-beelde as óf voorgrond (deel van die oorskulp) óf agtergrond (homogene vel, juweliersware, óf hare) te klassifiseer. Onafhanklik van die FG-herkenningsprosedure, word elke grysskaal-invoerbeeld in geheel aan Guassiese vergladding onderwerp, gevolg deur randherkenning met behulp van 'n geskikte Canny-filter, en morfologiese randverdikking. Die herkende FG dien as 'n masker wat slegs daardie randte wat met prominente kontoere van die oorskulp geassosieer word, behou. Kenmerke word vervolgens vanuit elke binêre kontoerbeeld met behulp van die diskrete Radon transform onttrek. Bogenoemde kenmerke word sodanig genormaliseer dat dit translasie-, rotasie- en skaal-invariant is. 'n Euklidiese afstandsmaat word vir die doel van kenmerkpassing aangewend. Oor-gebaseerde herkenning word laastens bewerkstellig deur van 'n rangorde-verifieerder gebruik te maak. Uitgebreide eksperimente word op twee groot internasionale datastelle uitgevoer. Daar word aanvaar dat slegs een verwysingsoor vir elke geregistreerde individu beskikbaar is. 'n Eksperimentele protokol wat die onderskeie datastelle sinvol op grond van afrigtings-, bekragtigings-, ordenings- en evalueringsindividue verdeel, word gevolg. Daar word aangetoon dat die vaardigheid van die nuwe stelsels wat in hierdie tesis ontwikkel is, goed met dié van bestaande stelsels vergelyk.

Acknowledgements

I would like to express my sincere gratitude to the following people and organisations:

- My supervisor, Dr Hanno Coetzer, for his invaluable insight, guidance, patience, unwavering support, immense knowledge and valuable critiques of this research work. This study would not have been possible without his input.
- The Universidad de Las Palmas de Gran Canaria for allowing the use of their ear database.
- The Hong Kong Polytechnic University Department of Computing for also allowing the use of their ear database.
- The Postgraduate Funding Department of Stellenbosch University, for their support and financial assistance.
- The Ball family, for their financial support throughout my postgraduate studies.
- My family and close friends, for their unconditional love and support.

Contents

List of Figures

List of Tables

List of Acronyms

AER	Average error rate
ACC	Accuracy
BN	Batch normalisation
CNN	Convolutional neural network
DRT	Discrete Radon transform
FAR	False acceptance rate
FC	Fully-connected
FRR	False rejection rate
MLP	Multilayer perceptron
REC	Recall
RF	Receptive field
ReLU	Rectified linear unit
ROI	Region of interest
PRE	Precision
SLP	Single layer perceptron
SGD	Stochastic gradient descent
SGDM	Stochastic gradient descent with momentum

Nomenclature

Variables within the context of deep learning

η	Learning rate
γ	Momentum value
∇E	Gradient of the loss function
σ	Standard deviation
f	Activation function
f_i	Activation function associated with the i-th node
l	Iteration number for a one full pass (forward and backward)
p_j	Probability of j-th class (output of softmax function)
w_{ij}	Weight associated with the i-th node within hidden layer j
x_i	Input for node i
y_i	Output for node i

Vectors within the context of deep learning

$\mathbf{b}$	Bias vector
$\mathbf{w}$	Weight vector

Variables within the context of the discrete Radon transform

β	Number of non-overlapping beams per angle

δ_{ij} The contribution of the i-th pixel towards the j-th beam-sum

Θ Total number of angles

R_j The j-th beam-sum

Subscripts

i, j, m, n Integer index

Chapter 1

Introduction

1.1 Background and motivation

In a modern society where digital social interaction is becoming increasingly commonplace and where financial transactions are routinely conducted through digital means, a reliable automated biometric system that is able to establish or verify an individual's identity is of paramount importance. A biometric system is in essence a pattern recognition system which uses a specific physiological or behavioural characteristic of a person for the purpose of establishing or verifying an individual's identity by first extracting prominent features from a questioned sample (image) and then comparing these features against a stored feature set or trained statistical model. Traditional means for personal authentication such as access cards, personal identification numbers (PINs) or passwords, can be can be stolen, duplicated, lost or forgotten. Due to the aforementioned limitations associated with traditional modes of personal authentication, the development of biometric systems is proving to be an efficient solution in overcoming the aforementioned shortcomings. Biometric systems are also inherently more reliable than most traditional modes of personal authentication due to measurable biometric traits such as universality, uniqueness, collectability and permanence.

A human ear constitutes a stable structure which does not change significantly as a result of aging and may be regarded as one of the most distinctive human biometric traits since it possesses all of the aforementioned attributes of uniqueness, collectability, permanence and universality (Iannerelli, 1989). The human ear furthermore constitutes a large, passive, and non-intrusively acquirable biometric trait, that remains relatively invariant despite changes in facial expression, the wearing of eye glasses or the application of make-up, and may therefore be considered more reliable than most other facial features for the purpose of personal identification and verification (Chang *et al.*, 2003).

Mark Burge and Wilhelm Burger were responsible for the first attempt at an automated ear-based biometric authentication system in 1996. They employed a mathematical graph model for the purpose of automatically extracting features from ear images in order to match certain curves and edges (Burge & Burger, 1996). In 1999, Belé Moreno, Ángel Sanchez, and José Vélez presented a study on a fully automated ear-based recognition system which is based on various attributes, like localised feature points and the morphology of the outer ear (Moreni *et al.*, 1999). Numerous feature extraction and matching algorithms for ear recognition have been proposed by researchers ever since. A dichotomisation of these systems is presented in Chapter 2.

The remainder of this chapter is structured as follows: An overview of the scope and objectives of this study is presented in Section 1.2. This is followed by a brief synopsis of the proposed system (see Section 1.3). The abbreviated results are presented in Section 1.4, while the contributions of this study are listed in Section 1.5. An outline of this thesis is given in Section 1.6.

1.2 Scope and objectives

The aim of this thesis is to develop a novel, fully automated and proficient ear-based biometric authentication system. The scope of the thesis is limited to situations where

(1) a *single* reference ear image is available for each client enrolled into the system, and

(2) ear images that belong to *other* individuals than the client in question - these individuals are partitioned into *training*, *validation* and *ranking* individuals - are also available.

The ear images that belong to the *training* and *validation* individuals are used to respectively train and validate an appropriate convolutional neural network (CNN) for the purpose of discriminating between sub-images that contain contours typically associated with an ear and those that contain background information. This facilitates the detection of appropriate ROIs within the ear images associated with the *ranking* individuals, as well as the ear images that constitute the questioned and reference samples associated with the claimed individual. Radon transform-based features extracted from the detected prominent ear contours within the questioned sample are matched to those of the reference sample for the claimed individual, as well as to those associated with the *ranking* individuals. Authentication is ultimately based on the relative rank of the resulting distance associated with the reference sample for the claimed individual, when the aforementioned rank is compared to the respective ranks of the resulting distances associated with the ear images that belong to the *ranking* individuals.

The scope of the thesis is furthermore limited to situations where

(1) the two-dimensional plane in which each ear approximately resides is more or less parallel to the two-dimensional plane in which the camera lens approximately resides,

(2) the distance between the abovementioned planes is allowed to vary,

(3) each ear may be orientated (rotated) differently within the abovementioned plane, and

(4) each ear may be translated differently within the abovementioned plane.

The abovementioned delimitations imply that the head of an individual is allowed to shift, tilt up or down, or move towards or away from the camera lens, while ensuring that the other side of the head is restrained by, for example, allowing it to rest against a solid vertical surface. The head is therefore not allowed to tilt towards or away from the camera. Pronounced tilting of the head towards or away from the camera inevitably leads to a deterioration of the proposed system's ability to consistently detect prominent contours associated with the ear shell, due to occlusions, etc. The scope of this thesis is further restricted to biometric authentication based on *right* ears. A specific individual's left and right ears may differ slightly.

Although the aim of this thesis is to develop a fully automated ear-based biometric authentication system, the proficiency of a semi-automated system in which a human operator manually selects the ROI for each questioned ear, will also be investigated and reported on. The manually selected ROIs also serve as a ground truth for evaluating the automated CNN-based ROI detection protocol. The proficiency of the fully automated end-to-end ear-based biometric authentication system is finally investigated and reported on.

1.3 System design

The enrollment and authentication stages of the semi-automated and fully automated (end-to-end) ear-based biometric authentication systems proposed in this thesis are conceptualised in Figures 1.1 and 1.2.

1.3.1 Data

In this thesis experiments are conducted on two different datasets, that is (1) the Mathematical Analysis of Images (AMI) ear database and (2) the Indian Institute of Technology (IIT) Delhi ear database. In the case of the AMI ear database, each image is first converted from RGB to grey-scale, while the images in the IIT Delhi ear database were originally captured in grey-scale format. The resolutions of the grey-scale images associated with the AMI and IIT Delhi ear databases are 702×492 pixels and 272×204 pixels, respectively.

1.3.2 Image segmentation

A CNN-based approach is proposed to facilitate automatic ROI detection within the context of ear-based biometric authentication. The proposed CNN-based protocol, combined with appropriate morphological post-processing, is proficient in detecting a suitable ROI that contains the prominent contours associated with the ear shell. The automated ROI detection strategy proposed in this thesis is conceptualised in Figure 1.3

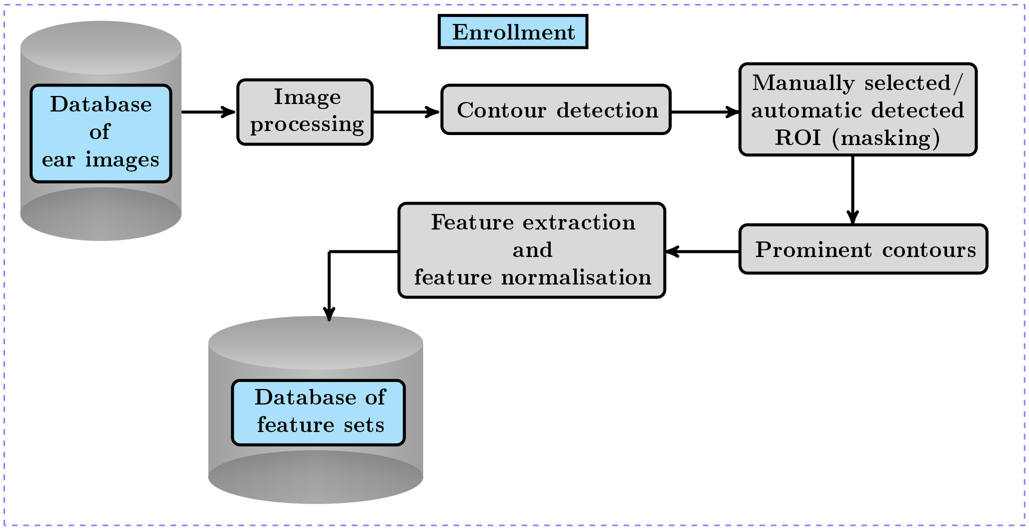


Figure 1.1: Conceptualisation of the *enrollment* stage of the semi-automated and fully automated ear-based biometric authentication systems proposed in this thesis.

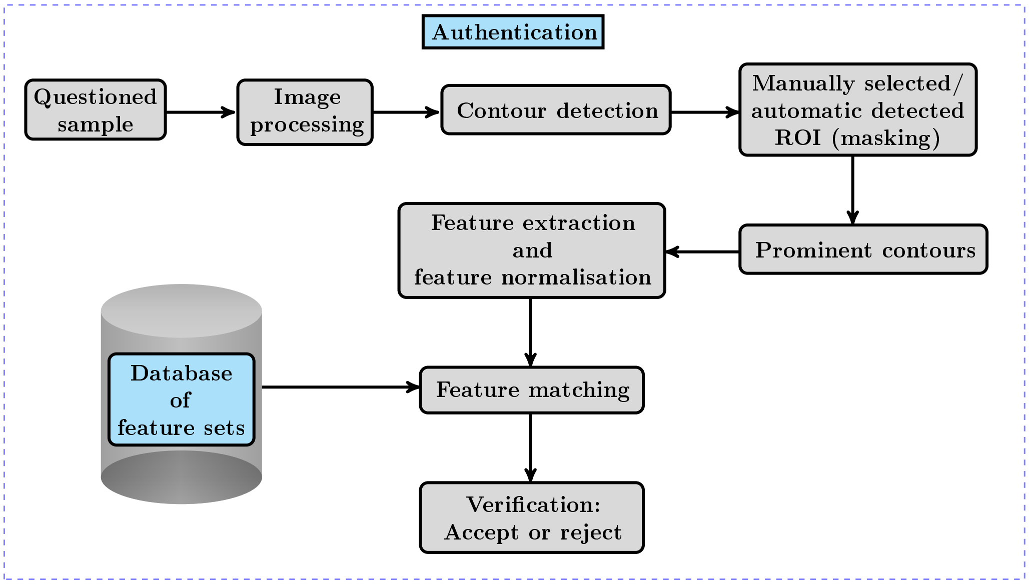

Authentication
Questioned sample
Image processing
Contour detection
Manually selected/ automatic detected ROI (masking)
Prominent contours
Feature extraction and feature normalisation
Database of feature sets
Feature matching
Verification: Accept or reject

Figure 1.2: Conceptualisation of the *authentication* stage of the semi-automated and fully automated ear-based biometric authentication systems proposed in this thesis.

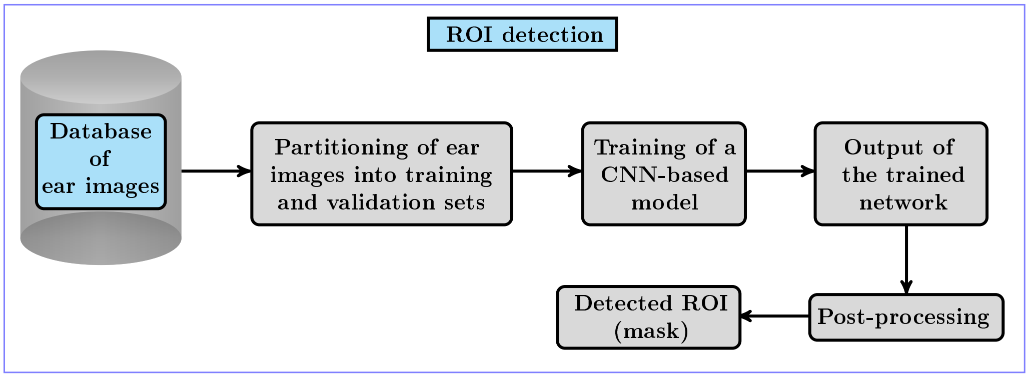

ROI detection
Database of ear images
Partitioning of ear images into training and validation sets
Training of a CNN-based model
Output of the trained network
Post-processing
Detected ROI (mask)

Figure 1.3: Conceptualisation of the proposed ROI detection protocol.

1.3.3 Preprocessing, contour detection and post-processing

A protocol for detecting prominent contours associated with the shell of a human ear is proposed. Appropriate preprocessing techniques are applied to the ear image in order to correct non-uniform illumination, suppress noise and enhance the contrast of the image. Prominent edges are detected through a Canny edge detector after which appropriate morphological operations are conducted in order to connect disconnected contours and remove small non-connected contours, while ROI-based masking is employed in order to ensure that contours associated with hair and jewellery are discarded.

1.3.4 Feature extraction and matching

A feature extraction strategy based on the calculation of the discrete Radon transform (DRT) of the contour image associated with the shell of a human ear is proposed. The extracted feature set is normalised in such a way that it constitutes a translational, rotational and scale invariant representation of the contours in question. After appropriate feature normalisation, template matching is achieved by calculating the average Euclidean distance between the corresponding feature vectors associated with the respective feature sets.

1.3.5 Verification

A rank-based verifier is finally employed in order to ascertain the authenticity of a questioned ear image.

1.4 Abbreviated results

As previously mentioned, the proficiency of the ear-based authentication systems developed in this thesis is estimated by considering two datasets namely the AMI and IIT ear databases. In this study *three* main algorithms are developed within the context of ear-based biometric authentication. Experiments are conducted to evaluate the proficiency of (1) the proposed automated ROI detection algorithm, as well as the respective proficiencies of (2) the semi-automated and (3) the fully automated ear-based biometric authentication systems developed in this thesis.

Within the context of the semi-automated ear-based biometric authentication system proposed in this thesis, two scenarios are investigated, that is (1) a scenario in which a questioned ear is only accepted when it has a ranking of one and (2) a scenario in which a questioned ear is accepted when it has a ranking better than or equal to an optimal ranking (which may be greater than one). For the first (rank-1) scenario, it is demonstrated that average error rates (AERs) of 2.4% and 6.59% are achievable within the context of the AMI and IIT Delhi ear datasets, respectively. In the case of the second (optimal ranking) scenario, it is however demonstrated that the above-mentioned error rates may be reduced to 1.91% and 5.07%, respectively.

What the CNN-based automated ROI detection algorithm developed in this thesis is concerned, it is demonstrated that 91% and 88% of the pixels are correctly classified as either ear pixels or background pixels within the context of the AMI and IIT Delhi ear databases, respectively.

Within the context of the fully automated ear-based biometric authentication system proposed in this thesis, only the scenario in which a questioned ear with a ranking of one is accepted, that is the rank-1 scenario, is investigated. For this scenario, AERs of 12.5% and 23% are reported for the AMI and IIT Delhi ear databases respectively. *An improvement on these results is however expected when other (optimal) rankings are also considered, but this was not investigated in this thesis due to time constraints.*

1.5 Contributions

To the best of our knowledge, the semi-automated and fully automated systems developed in this thesis employ an ensemble of techniques within the context of machine learning and template matching that has not been employed for ear-based biometric authentication on previous occasions, and may therefore be considered *novel*. This work may also form the basis of an investigation into an end-to-end deep learning-based approach to ear-based biometric authentication.

1.6 Thesis outline

The thesis is structured as follows:

Chapter 2: Literature study. A concise overview of existing research within the context of ear-based biometric authentication is presented in accordance to the systems proposed in this thesis. In particular, existing research on the automated segmentation of human ears and/or the detection of a ROI that encloses the ear in question, is scrutinised. Furthermore, existing research on feature extraction protocols and feature matching approaches within the context of ear-based recognition systems is laid out in this chapter.

Chapter 3: Image segmentation. The proposed CNN-based algorithm for the automatic detection of the ROI within the context of ear-based biometric authentication is described. Amongst other things, the parameters and data partitioning protocol utilised in the training of the CNN-based algorithm and the post-processing approach are discussed.

Chapter 4: Contour detection. The image processing algorithms that are utilised during the proposed contour detection protocol are discussed. Amongst other things, the Canny edge detector employed for the purpose of identifying prominent contours associated with the ear, followed by appropriate post-processing operations which ensure that noise and short edge segments are removed, are discussed in detail.

Chapter 5: Feature extraction, feature matching and verification. The proposed strategy that facilitates feature extraction from the contour image via the DRT is presented. An appropriate feature normalisation strategy to ensure scale, translation and rotation invariant representations of the original image is described. A feature matching protocol that is based on the Euclidean distance measure is discussed. Finally, a verification protocol that is based on the construction of a ranking verifier is introduced.

Chapter 6: Experiments. The datasets considered in this research and an outline of the experimental protocol employed in this thesis are discussed. This is followed by exhaustive experiments that gauge and analyse the proficiency of the algorithms proposed in this thesis. An overview of the software developed and hardware utilised in this thesis is also presented.

Chapter 7: Conclusion and future work. The research conducted in this thesis, as well as the experimental results are analysed and placed into perspective, after which avenues for future research are explored.

Chapter 2

Literature study

2.1 Introduction

As mentioned in the previous chapter numerous research studies on ear-based biometric authentication/recognition systems have been proposed on previous occasions. The most prominent pioneering work within this context is probably that by Iannarelli (Iannerelli, 1989). In this work the author examined 10000 ear images from which he extracted 12 geometric measurements based on the crus of the helix of the ear and concluded that these measurements are unique across individuals.

In this chapter a concise overview of relevant existing ear-based authentication systems is presented. The discussion provided on the aforementioned systems is therefore in some way related to the work presented in this thesis. The systems are therefore categorised into (1) the algorithms proposed for the automated segmentation of the ear or the detection of the region of interest (ROI) (see Section 2.2), (2) the techniques proposed for the purpose of extracting features from the ear (see Section 2.3) and (3) the proposed feature matching and verification paradigms for the purpose of ear-based authentication (see Section 2.3).

Since most existing ear-based authentication systems have not been evaluated on the same datasets than those considered in this thesis, it is not possible to directly compare the reported proficiency of these systems to those proposed in this thesis. Fortunately, a few existing systems have in fact been evaluated on the same datasets than those considered in this thesis which facilitates a more direct comparison in system proficiency in these cases. In Section 2.4 such a comparison is drawn within the context of the semi-automated system developed in this thesis. *It is important to note that the experimental protocol (data partitioning) may differ amongst the systems being compared.*

2.2 Automated ear segmentation

Automatic ear segmentation, that is the detection of the region of interest (ROI), involves the localisation of the ear shell within each ear image. In this chapter a concise overview of the work that has been conducted on automated ear segmentation is presented. An overview of existing ear detection techniques is presented in Table 2.1, along with the employed databases and the reported performance rates.

Abaza *et al.* (2010) proposed a modified Adaboost algorithm based on Haar features for automated real-time robust detection of the ear. An Adaboost algorithm is a pattern detection or classification strategy that combines a set of weakly effective classifiers to form a strong classifier. The proposed technique classifies images based on the value of rectangular features, operating on small sub-images. The input image is first rescaled and then divided into overlapping sub-images of size 24×16 pixels. The cascaded Adaboost algorithm is subsequently applied to each of the sub-images. The proposed system was evaluated on the University of Manchester Institute of Science and Technology (UMIST) database, the University of Notre Dame (UND) database, the West Virginia High Technology Foundation (WVHTF) database, the Facial Recognition Technology (FERET) database and the University of Science and Technology Beijing (USTB)-III dataset. Detection accuracies of 100%, 94.37% 93.86%, 84% and 93.75% are reported for the respective datasets.

Kumar and Wu (2012) proposed an automated ear detection algorithm based on basic image preprocessing techniques, morphological operations and Fourier descriptors. The proposed strategy involves the smoothing of the image with a Gaussian filter for the purpose of suppressing the effect of noise followed by histogram equalisation. Morphological operations (closing and opening techniques) are simultaneously applied after histogram equalisation. Otsu's threshold is employed on the preprocessed image to generate a binarised mask image. The resulting binary mask is subsequently employed for the purpose of ROI detection within the original grey-scale image, which is followed by morphological dilation of the grey-scale image. Morphological opening operations are subsequently applied for the purpose of noise elimination. Boundary tracing is finally employed and the shape of the ear is defined using Fourier descriptors. The proposed strategy was evaluated on the Indian Institute of Technology (IIT) Delhi ear dataset. Results for the proposed automated ear segmentation protocol are not available.

Vélez *et al.* (2013) presented a novel automated ear segmentation algorithm based on the combined use of the circular Hough transform and anthropometric ear proportions for the accurate detection of the ear region. This technique involves image preprocessing and contour detection followed by the localisation of the ear region. The input ear image is firstly converted from a RGB format to grey-scale format, after which a median filter is applied for noise removal. A Canny edge detector is employed for the purpose of detecting prominent contours. Morphological dilation is applied on the edge image using a disk-shaped structuring element of size 4×3 after which small connected components are removed. The proposed detection of the ear region is carried out by searching for circles through the application of the circular Hough transform. First a search is conducted for the upper helix region and once this region is detected, anthropometric ratios of the ear are considered for the detection of the remainder of the ear. To test the proposed ear detection technique the authors created three different image databases consisting of grey-scale, RGB and near infrared (NIR) images, respectively. Detection accuracies of 87.88%, 78.33% and 64% are reported for the respective datasets.

Yuan and Mu (2014) proposed an ear detection approach based on an improved Adaboost algorithm and the active shape model (ASM). The proposed technique detects the ear region under complex background conditions through the application of two steps, that is offline cascaded classifier training and online detection. For the improved Adaboost algorithm the authors propose a segment selection algorithm to choose the optimum threshold of weak classifiers. They also propose a strategy to reduce the false acceptance rate by changing the weight distribution of the weak classifiers and a new parameter is applied to improve the robustness of the detector and prevent overfitting. A single ear-detection technique is proposed on the basis of the asymmetry of the right and left ears. For the final segmentation of the ear region an automatic ear normalisation strategy based on the ASM is applied. The proposed techniques are evaluated on two datasets, that is the USTB-

III and the UND ear datasets. Detection accuracies of 96.46% and 94% are reported for the respective datasets.

Zhang and Mu (2017) proposed an automated ear detection technique that involves multiple scale faster region-based convolutional neural networks (Faster R-CNN) to detect ears from two dimensional profile images in uncontrolled conditions. The proposed technique involves the detection of three regions of different scales through the region proposal network (RPN) technique for the estimation of the location of the ear within the image. An ear region filtering technique is proposed to automatically eliminate false positives and for the accurate detection of the ear region via a threshold value method. The experiments for the proposed techniques were conducted on the Collection J2 of the University of Notre Dame Biometrics Database (UND-J2), and the University of Beira Interior Ear dataset (UBEAR). In addition to this they created their own dataset named WebEar which was also used for the purpose of conducting the experiments. Detection accuracies of 100%, 98.22%, and 98% are reported for the respective datasets.

Galdámez *et al.* (2017) proposed a CNN algorithm in conjunction with different object detectors based on the Viola-Jones framework for automated detection of the ear region. The authors used the Haar cascade classifier to identify the face profiles and proceeded to obtain the ear using the same Haar technique. The image ray transform (IRT) is computed in scenarios where the Haar technique fails to identify the ear. A Gaussian smoothing filter is applied in order to eliminate noise and remove gaps in the helix. The resulting image is then thresholded to obtain the final helix. An elliptical template is used to match the image. After the ROI is detected, preprocessing operations are performed. A RGB image is converted into grey-scale format and the image is normalised. Ear segmentation is performed by applying a mask. The Canny edge detector is employed for the detection of prominent contours within the detected ear region. The proposed system is evaluated on the Ávila's Police School database and on the Bisite videos database. Detection accuracies of 99.02% and 98.03% are reported for the respective datasets.

Publication	**Detection technique**	**Dataset**	**Accuracy (%)**
Abaza *et al.*, 2010	Modified AdaBoost	UMIST, UND WVHTF, FERET and USTB-III	100, 94.37 93.86, 84 and 93.75
Kumar & Wu, 2012	Local orientation and local gray level phase information	IIT Delhi	N/A
Vélez *et al.*, 2013	Modified AdaBoost	RGB, grey-scale and NIR images	87.88, 78.33 and 64
Yuan& Mu, 2014	Improved AdaBoost	USTB-III and UND	96.46 and 94
Zhang & Mu, 2017	Multiple scale faster R-CNN deep learning model	WebEar, UND-J2 and UBEAR	98, 100 and 98.22
Galdámez *et al.*, 2017	CNN techniques combined with Viola-Jones framework and IRT	Ávila's Police School and the Bisite	99.02 and 98.03

Table 2.1: A summary of existing two dimensional ear detection techniques, the databases employed and the reported detection accuracies. *The CNN-based ROI-detection algorithm proposed in this thesis achieves detection accuracies of 91% and 88% when evaluated on the AMI and IIT Delhi ear datasets.*

2.3 Feature extraction and matching

In this section a brief overview is presented of existing techniques that have been proposed for the extraction of a set of measurable features from ear images within the context of ear-based biometric authentication systems. A concise overview of the relevant template matching techniques proposed within the context of ear-based biometric authentication systems is provided and their respective performances are presented.

Choras (2008) proposed four novel techniques for feature extraction from two dimensional images based on geometrical strategies. The author extracted geometrical features from normalised contour images. The author furthermore proposed a feature extraction technique based on concentric circles centred at the centroid of the ear image. A contour tracing strategy based on extracting characteristic intersection points of the circles and the ear contours are used as feature points. An angle based contour representative technique is employed in which the angles between the centre point and the concentric circle intersecting points are employed for feature representation. A triangle ratio method determines the normalized distances between reference points and uses these distances for ear description. The author conducted studies on different databases and reports recognition rates between 86.2% and 100% for a database of 240 ear images (which includes 20 different views) from 12 subjects, and false rejection rates between 0% - 9.6% for a large databases of 102 ear images.

Tharwat *et al.* (2012) proposed the principal component analysis (PCA) algorithm for the purpose of extracting features from ear images. The authors propose four feature extraction techniques based on the PCA algorithm. In the first approach the whole image is used, while in the second, third and fourth approaches the ear image is first divided into non-overlapping sub-images. The images are centred by calculating the mean of each image. A covariance matrix is then constructed by calculating the eigenvalues and corresponding eigenvectors. For the second, third and fourth strategies the ear image is first divided into non-overlapping sub-images of four, nine and 16 blocks of size 64×64 pixels respectively. The PCA features are extracted from each sub-image. A minimum distance classifier is employed for template matching. The respective outputs of the classifiers are then combined on the abstract, score and rank levels. The cosine, Euclidean and city block distances are considered. The author conducted experiments on 102 grey-scale ear images (6 images per individual) and reports recognition rates within a range of 64.70% and 97.06%.

Shu-zhong (2013) proposed an improved normalisation technique for feature extraction based on a geometrical algorithm. The author proposes an angle normalisation strategy by employing geometrical parameters to ensure a translational, rotational and scale invariant representation of contour images. The proposed strategy is based on the extraction of an external ear shape feature. The author defines the connection of the highest and the lowest point on the outer ear contour as the long axis. The author then defines the long axis and the centre of mass as geometrical parameters for a feature vector representation and subsequently performs the angle normalisation by considering the geometrical parameters.

Yuan and Mu (2014) employed the Gabor filter for feature extraction and the kernel Fisher discriminant analysis (KFDA) for dimension reduction. The Gabor filter is applied on the ear images to extract spatially localised features of different directions and scales. Gabor-based feature extraction is implemented by convolving the ear image with the Gabor kernel function. Since the Gabor features are high dimensional, the full space KFDA algorithm is applied for feature reduction. A distance-based classifier is applied for classification purposes. The proposed approach was evaluated on the USTB and the UND ear datasets. The radial basis function (RBF) kernel was employed for classification purposes within the context of a rank-1 scenario. Recognition rates of 96.46% and 94% are reported for the USTB and the UND ear datasets respectively.

A novel feature extraction technique based on the fusion of the shape of the ear and the tragus was proposed by Annapurani *et al.* (2015). The authors extracted the shape of the ear by first performing preprocessing techniques, after which the preprocessed image is binarised. The connected components are calculated and the largest blob is classified as the ROI. The boundary of the blob is marked and the shape of the ear to be extracted is given by the maximum length of the marked boundary. The tragus is extracted by drawing a line connecting the maximum and the minimum coordinates. The centre region of the aforementioned line defines the tragus of the ear. The shape of the ear and the extracted tragus are fused to form a feature template. The Hamming distance and the Euclidean distance are employed for template matching. More specifically, the queried feature is compared to the enrolled feature of the claimed identity using the Hamming and Euclidean distances. Experiments are conducted on two datasets namely the AMI and IIT Delhi ear datasets. Within the context of the AMI ear dataset, accuracies of 99.97% and 100% are reported the Hamming and Euclidean distances respectively. For the IIT Delhi ear dataset an accuracy of 100% was reported for both the Hamming and Euclidean distances.

Rahman *et al.* (2016) employed the scale invariant feature transform (SIFT) algorithm for feature extraction and feature matching purposes. Features from the ear image are extracted using the SIFT algorithm (key-point location). Template matching or classification is done using a minimum distance classifier. The proposed system is evaluated on two datasets that is the IIT Delhi ear database and the AMI ear database, and recognition rates of 95.2% and 100% respectively are reported.

Omara *et al.* (2016a) proposed a novel feature extraction strategy based on the polar sine transform (PST). Preprocessing operations are applied to the input ear images followed by ear normalisation. The preprocessed images are then divided into overlapping circular sub-images of size 16×16 pixels with a step size of 2 pixels. The PST coefficients are computed to extract invariant features for each sub-image. The extracted features are then accumulated to form a single feature vector to represent the ear image. The authors employed a support vector machine (SVM) for classification purposes. The proposed approach was evaluated on the USTB-III database and a recognition rate of 96.67% is reported.

A novel geometrical feature extraction strategy was proposed by Omara *et al.* (2016b). This strategy involves image preprocessing using a Gaussian filter for eliminating noise effects in the images, which is followed by the detection of prominent contours via the Canny edge detector. Geometrical features are extracted from the contour image for the purpose of describing the outer helix. The proposed geometrical feature extraction technique involves two steps: (i) the location of the upper right, upper left, and lower left segments of the outer helix, and (ii) the detection of the minimum ear height line (EHL) and the extraction of the shape features. The dissimilarity of two ear images is measured by the Euclidean distance. The proposed approach is evaluated on the USTB-I and the IIT Delhi ear databases. Recognition rates of 98.33% and 99.60% were reported for the respective datasets.

2.4 Comparison with existing systems

In order to place the performance of the systems proposed in this thesis into perspective, the reported proficiency of the aforementioned systems are compared to those of existing state-of-the-art ear-based biometric authentication systems. This comparison is drawn within the context of the semi-automated system developed in this thesis. In Table 2.2 a brief summary of existing feature extraction and template matching techniques is presented along with the reported performance rates for the AMI and IIT Delhi ear datasets.

Accuracies				
Publication	**Feature extraction technique**	**Feature matching technique**	**AMI (%)**	**IIT Delhi (%)**
Our approach	*Discrete Radon transform*	*Euclidean distance*	**98.92**	**94.06**
Annapurani *et al.*, 2015	Fusion of the shape of the ear and the tragus	Hamming distance and Euclidean distance	99.97 and 100	100
Recognition rates				
Publication	**Feature extraction technique**	**Feature matching technique**	**AMI (%)**	**IIT Delhi (%)**
Rahman *et al.*, 2016	SIFT	A minimum distance classifier	100	95.20
Omara *et al.*, 2016	Geometrical features of the shape of the ear	Euclidean distance	···	99.6

Table 2.2: A summary of existing feature extraction and feature matching techniques, and the reported performance rates within the context of the AMI and IIT Delhi ear databases.

Chapter 3

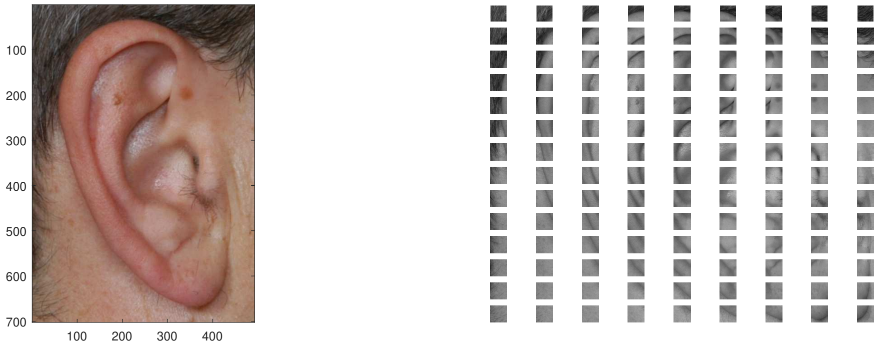
Image segmentation

3.1 Introduction

In this thesis novel semi-automated and fully automated ear-based biometric authentication systems are proposed. In the case of the semi-automated system a suitable region of interest (ROI), that contains the entire ear shell, is *manually* specified (selected). However, as part of the fully automated system, a suitable ROI has to be *automatically* detected. A convolutional neural network (CNN), followed by appropriate morphological post-processing, is proposed for this purpose.

Each ear image (see Figure 3.1 (a)) is partitioned into a number of overlapping sub-images (patches) by employing a sliding window (see Figure 3.1 (b)). The objective of the CNN is to classify each patch within a test image as either foreground or background. Foreground patches contain contours typically associated with the shell of a human ear, while background patches typically contain hair, jewellery and homogeneous skin. Ear images that are associated with so-called training and validation individuals are employed for the respective purposes of training the CNN (for ROI detection) and avoiding over-fitting.

It is important to note that ear images from *different* individuals are used for training, validation, ranking and evaluation purposes. The so-called ranking individuals are employed for the purpose of constructing a ranking verifier. Radon transform-based features extracted from a questioned ROI (within the evaluation set) is matched to the corresponding features extracted from a reference ROI (known to belong to the claimed individual), as well as



(a)

(b)

Figure 3.1: **(a)** An example of a RGB ear image of size 702×492 pixels. **(b)** A grey-scale version of the image depicted in (a) after being partitioned into 126 overlapping 82×82 sub-images (patches).

to the corresponding features extracted from ROIs belonging to the ranking individuals (known *not* to belong to the claimed individual). The resulting distances are ranked from small to large, after which the rank associated with the claimed individual is used to determine the questioned sample's authenticity.

The patches associated with the training and validation individuals are therefore manually annotated (labelled) and used to train and validate the CNN. The CNN is subsequently used to classify the (unseen) patches associated with the ranking and evaluation individuals.

This is followed by morphological post-processing for the purpose of ensuring that each detected ROI constitutes a fully-connected convex set of pixels that contains the entire ear shell. In order to quantify the proficiency of the proposed ROI-detection protocol, the amount of overlap between the manually specified (selected) ROIs and the automatically detected ROIs is estimated (and reported on) for the ranking and evaluation individuals. Within the context of ROI-detection, the ranking and evaluation sets may therefore be jointly referred to as the *test set*.

In this chapter a brief overview of the important concepts and algorithms associated with machine learning (in general) is first provided (see Section 3.2). This is followed by a general introduction to neural networks (see Section 3.3), after which the architecture and training of a typical CNN is discussed (see Section 3.4). Finally, in Section 3.5, the proposed ROI detection protocol within the context of ear-based biometric authentication is described in more detail, followed by an analysis of the results.

3.2 Machine learning

The main purpose of machine learning is to enable computers to learn and perform tasks with limited or no human intervention. A machine learning algorithm is simply defined as an algorithm that is able to learn from examples (observed or training data) without being *explicitly* programmed how to do so (Bishop, 2006). The algorithm enables the construction of a model that identifies certain patterns and structure in observed (training) data so as to predict the output for unseen (test) data. The basic protocol of a machine learning algorithm is therefore to receive and analyse input data in such a way that it is able to predict the output values within an acceptable range. As new data is fed into the system, the algorithm learns and optimises the model parameters in order to improve system performance.

Depending on which data is available, machine learning algorithms may be classified into one of the following paradigms: (1) *supervised learning*, (2) *unsupervised learning*, (3) *semi-supervised learning* and (4) *reinforcement learning* (Kotsiantis et al., 2007; Abraham & Sathya, 2013).

The underlying principle of supervised learning for predictive modelling is that the model learns to predict the output variables (y) from the input variables (x) using *labelled* data. Supervised learning algorithms may be further subcategorised into (1) *regression* and (2) *classification* models. Regression models predict *continuous* variables that link input-output pairs (Neter et al., 1996), while classification models assign the output variable to one of several *discrete* classes (Ren & Malik, 2003).

In an unsupervised learning scenario the algorithm finds structure from *unlabelled* training data by means of grouping the data into clusters or by arranging it in a more structured way. Semi-supervised learning constitutes a combination of supervised and unsupervised learning in which the algorithm considers *partially labelled* data. In the case of reinforcement learning the algorithms are goal-oriented and learn what actions to take in certain situations based on rewards and penalties.

One of the key problems being addressed in this chapter, that is the labelling of patches within an input ear image as either foreground or background, therefore constitutes a classification problem within the context of the supervised learning paradigm.

3.3 Neural networks and deep learning

Machine learning through a neural network, which contains a large number of hidden layers, is often referred to as *deep learning* (Glorot & Bengio, 2010). The concept of a hidden layer is discussed later in this section.

The basic building block of a neural network is a neuron (perceptron) as depicted in Figure 3.2. Each input value x_i is first multiplied with a corresponding weight w_i. The input-weight products are subsequently summed, after which a bias b is added to the weighted sum. A non-linear activation function f is finally applied to the resulting value in order to obtain the output y. This process can be mathematically formulated as follows,

$$y = f\big(b + \sum x_i w_i\big). \tag{3.1}$$

The neuron (perceptron) depicted in Figure 3.2 may for example be employed for the purpose of reaching a decision d as follows,

$$d = \begin{cases} 1, & \text{if } y \geq 0 \\ 0, & \text{if } y < 0. \end{cases} \tag{3.2}$$

A typical neural network contains a number of interconnected neurons (perceptrons) and consists of an input layer, an arbitrary number of hidden layers, an output layer, as well as a weight *matrix* **W** and a bias *vector* **b**. Each layer consists of a number of nodes, where each node (except for the input nodes) is associated with a neuron. When only two layers (that is an input layer and an output layer) are present, the network is referred to as a single layer perceptron (SLP). A SLP therefore contains no hidden layers, of which the simple "network" depicted in Figure 3.2 is an example. A network that contains at least three layers (that is a network with one or more hidden layers) is referred to as a multilayer perceptron (MLP). An example of an MLP that contains two hidden layers is presented in Figure 3.3.

In Figure 3.3 each node within a hidden layer (coloured blue), as well as the output node (coloured red) is able to receive, process and propagate data in the same way as is the case for the single neuron (perceptron) depicted

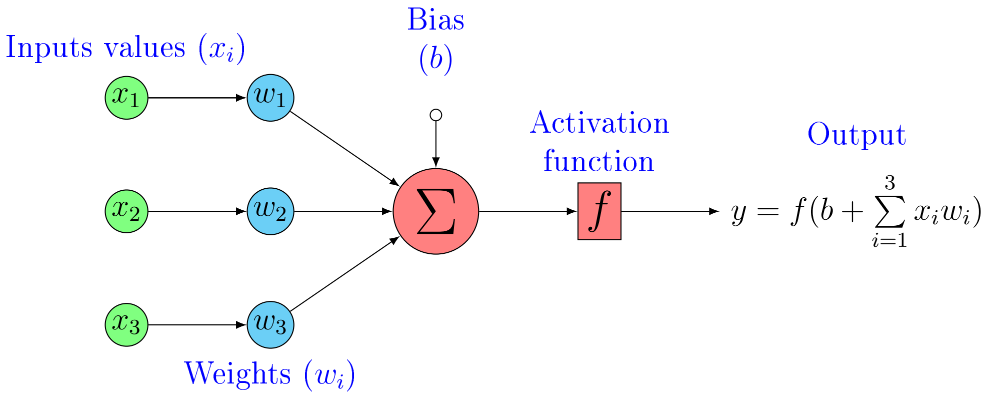


Figure 3.2: A neuron (perceptron) with three inputs values, x_1, x_2 and x_3.

in Figure 3.2. The output of one node therefore serves as input for the next. It is important to note that, during any given training iteration, the weight associated with propagating from node i to node j, that is w_{ij}, is the same irrespective of the layers involved. Furthermore, during a given training iteration, the bias associated with a specific node i, that is b_i, is the same across all the layers. Although a different activation function f_i may be associated with each node i, the function is kept fixed during training.

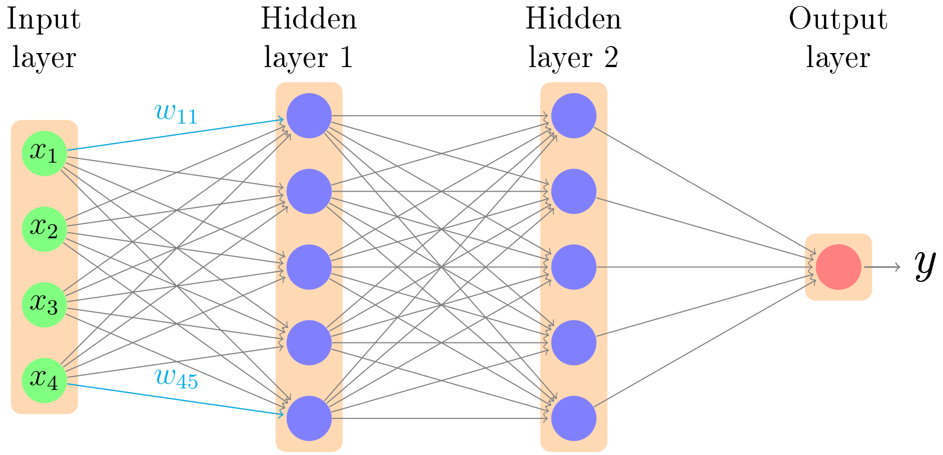


Figure 3.3: An example of a fully-connected neural network (MLP) with two hidden layers.

The main purpose of an activation function is to enable the network to learn more complex patterns by introducing non-linearity. Such an activation function does however have to be differentiable in order to facilitate back-propagation for optimisation purposes during training. The most popular activation functions include the logistic sigmoid function ($f(x) = \frac{1}{1+e^{-x}}$) (see Figure 3.4 (a)) which maps the input to the interval [0,1], as well as the hyperbolic tangent (tanh) function ($\tanh(x) = \frac{2}{1+e^{-2x}} - 1$) (see Figure 3.4 (b)) which maps the input to the interval [-1,1]. The tanh function may be expressed as a scaled version of the sigmoid function as follows $\tanh(x) = 2f(2x) - 1$. Both of these functions do however tend to saturate during training. This, in turn, leads to the exploding/vanishing gradient problem, which may fortunately be mitigated by introducing the so-called ReLU function. The ReLU function is discussed in more detail within the context of CNNs in Section 3.4.2.

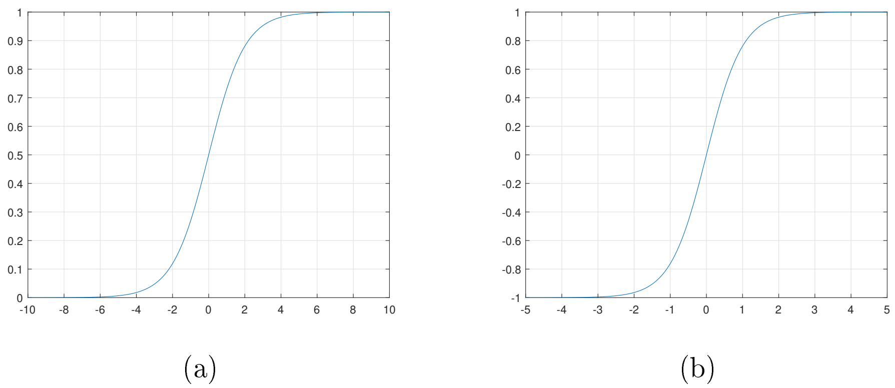


Figure 3.4: Popular activation functions. **(a)** The logistic sigmoid function. **(b)** The hyperbolic tangent (tanh) function.

During the training phase the input values are passed through the network, after which the *predicted* (network) output is compared to the *target* (desired) output. The difference between the predicted and target output may be quantified by an error (loss) function. This error is used to modify (update) the network parameters (that is the weights and biases) in such a way that the error gradually decreases over a number of training iterations through a back-propagation algorithm that (for example) employs stochastic gradient descent (SGD).

A SLP can only classify linearly separable functions, while a MLP is capable of also classifying non-linearly separable functions. MLPs are often referred to as deep feed-forward networks and form an important underlying component of CNNs. Deep feed-forward networks constitute directed acyclic graphs which implies that these models allow information to flow in one direction only, that is from the input layer through the hidden layers and finally to the output layer without any feedback connections or loops.

3.4 Convolutional neural networks

Convolutional neural networks (CNNs) are often simply referred to as convolutional networks. These networks represent a specialised class of neural networks that employs convolutional layers for the purpose of extracting pertinent information. Output in the form of a feature map is obtained by convolving the input data with a convolutional kernel (filter). The filters in the convolutional layers are automatically adjusted (updated) during training based on learned parameters in order to extract optimal features for the specific task at hand. In this section the architecture and training of a typical CNN are introduced within the context of handwritten digit recognition (see Figure 3.5). A discussion of the CNN employed in this thesis for the specific purpose of detecting a suitable ROI for ear-based biometric authentication (and its hyperparameters) is reserved for Section 3.5.

3.4.1 Architecture

A typical CNN consists of an input layer, one or more convolutional layers, a number of fully-connected (FC) layers (akin to the hidden layers discussed in the previous section), as well as a suitable classifier (which may for example be based on a softmax function). Additional specialised layers, like ReLU, pooling and dropout layers are often included to accelerate convergence and avoid overfitting.

The input layer contains the raw input data, that is the individual pixel values, as well as the width, height, and depth (number of channels) associated with the image that is to be processed by the network in question.

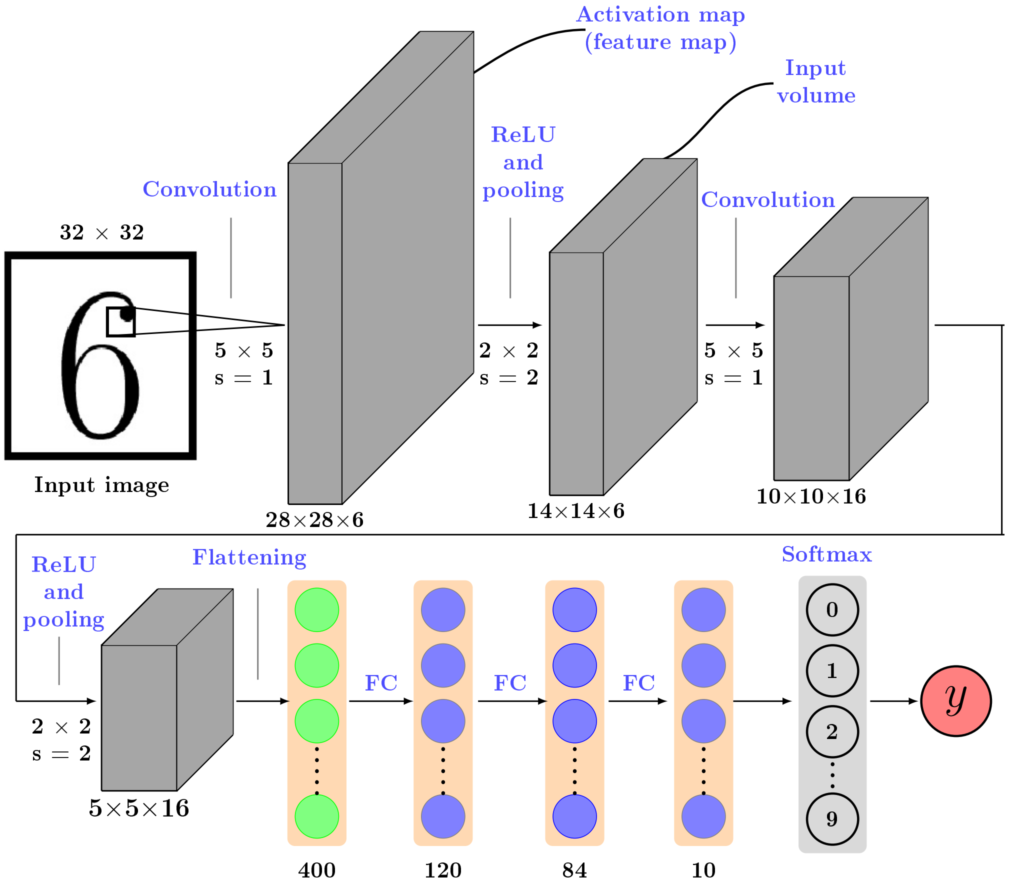


Figure 3.5: Architecture of a typical CNN suitable for handwritten digit recognition. The two convolutional layers are associated with 6 and 16 different kernels (filters) respectively. Each kernel has a size of 5 × 5 pixels, while a convolutional stride of $s = 1$ pixel is employed. The pooling layers consider 2 × 2 sub-images and employ a stride of $s = 2$ pixels. Three FC layers are present. This figure was redrawn from (LeCun et al., 1998).

Convolutional layer

Within the context of linear spatial filtering the concepts of convolution and correlation, which entail the process of sliding a kernel (filter) across an image, are closely related (Gonzalez & Wood, 2010). What is referred to as a convolution operation in most machine learning applications is often in actual fact a correlation operation. When the kernel (filter) is located on a specific part of an input image–this part of the image is referred to as the receptive field (RF)—the filter coefficients (weights) are multiplied with the corresponding pixel values. The sum of these products is subsequently assigned to the entry in the convolved output image that coincides with the centre of the RF (Goodfellow et al., 2016). The convolution process is conceptualised in Figure 3.6. More than one kernel (filter) is typically associated with each convolutional layer. The convolved output image for a specific kernel (filter) constitutes a specific channel within a multi-channel stack, referred to as an activation map (feature map).

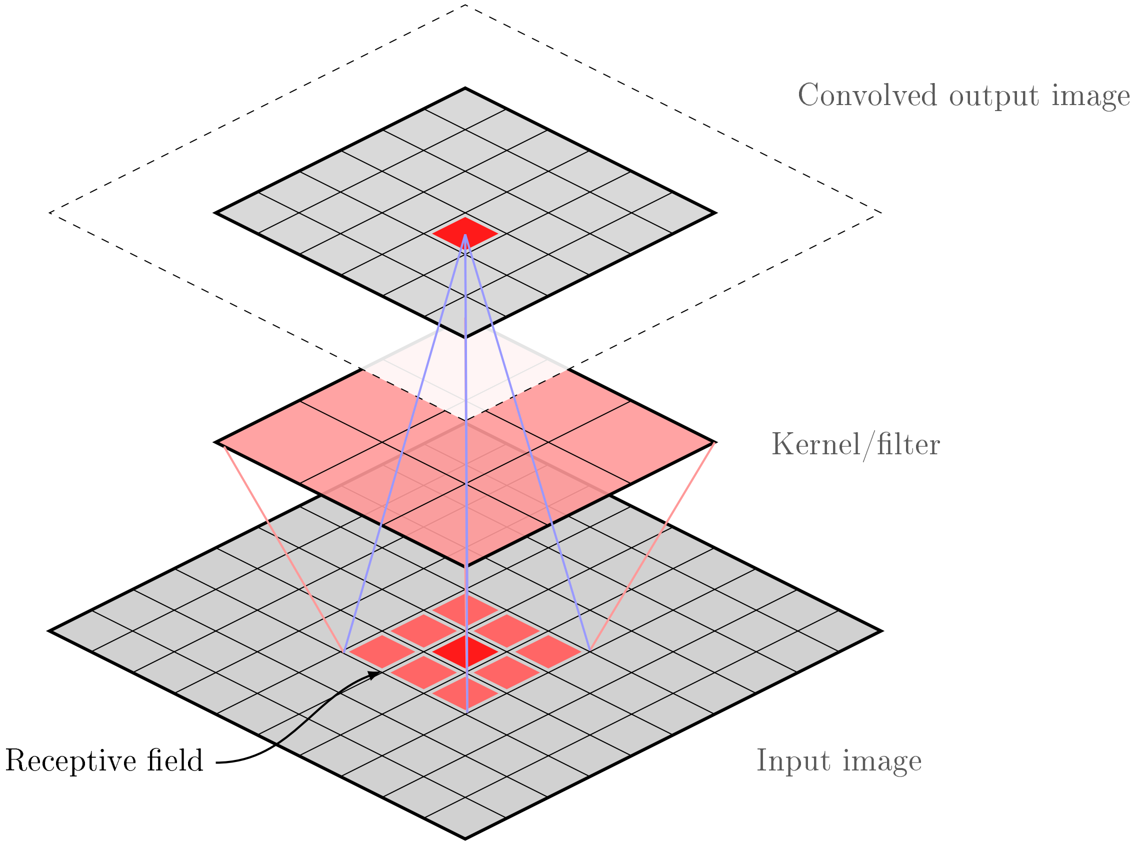


Figure 3.6: Conceptualisation of the convolution process within the context of a single-channel input image.

A multi-channel input image (for example an RGB image), or an activation map that has been processed by an operation like pooling, often serves as input to subsequent convolutional layers and is commonly referred to as the "input volume". A kernel (filter) that is convolved with such an input volume therefore typically constitutes a three-dimensional array with a depth equal to that of the input volume.

Networks intuitively learn filters that activate when they encounter certain types of features (located at specific spatial positions) in the input data. Within the context of a convolutional layer, each node is only connected to a local region within the input volume (that is the RF of the node). In this way a FC layer is essentially rendered locally-connected as conceptualised in Figure 3.7. This greatly reduces the number of parameters that has to be trained for each convolutional layer. The introduction of convolutional layers further reduces the parameter count by employing a technique referred to as parameter sharing as conceptualised in Figure 3.8. The aforementioned technique enables nodes connected to different local regions to have the same weight and ensures translational invariance.

Three main hyper-parameters determine the size and spatial arrangement of the output volume (feature map), namely the output depth, the stride and zero-padding.

- The **output depth** corresponds to the number of convolutional filters being employed, while the set of neurons that focuses at the same region within the input volume is referred to as the depth column.

- The **stride** refers to the number of pixels by which the centre of the kernel (filter) is adjusted after its application, while moving across the input volume as part of the convolution process. This reduces the spatial resolution, enables more efficient processing and contributes towards local translational invariance.

- **Zero-padding** involves the concatenation of zero-valued entries to the border of the input volume so as to ensure that the size of the input and output volumes remains similar.

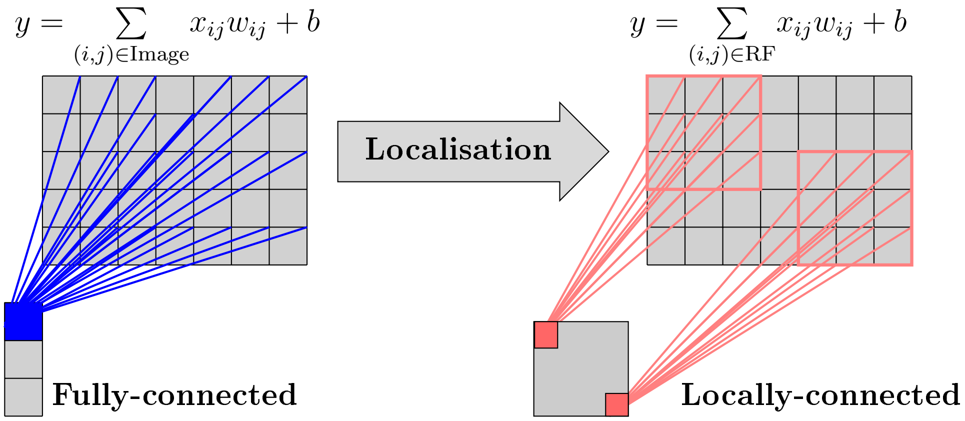

$y = \sum_{(i,j)\in\text{Image}} x_{ij}w_{ij} + b$
$y = \sum_{(i,j)\in\text{RF}} x_{ij}w_{ij} + b$
Localisation
Fully-connected
Locally-connected

Figure 3.7: Conceptualisation of rendering a FC layer locally-connected. This figure was redrawn from (Lee, 2008).

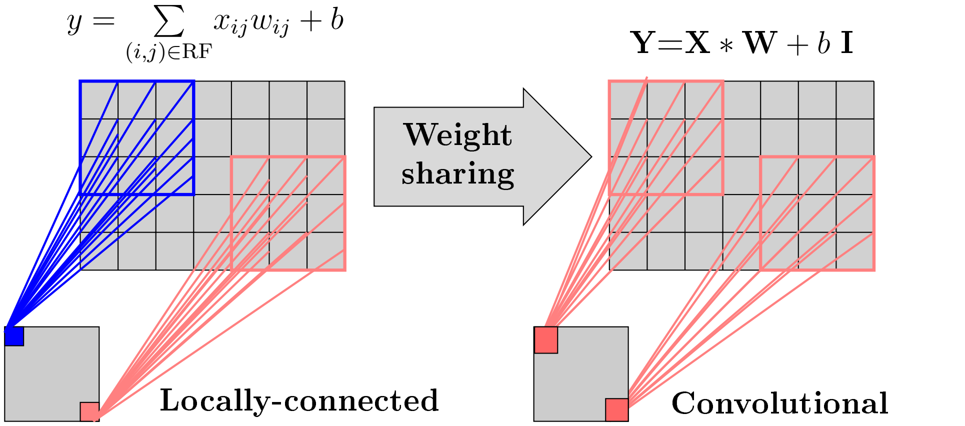

$y = \sum_{(i,j)\in\text{RF}} x_{ij}w_{ij} + b$
$\mathbf{Y}{=}\mathbf{X} * \mathbf{W} + b\ \mathbf{I}$
Weight sharing
Locally-connected
Convolutional

Figure 3.8: Conceptualisation of weight sharing, where the $*-$operator denotes convolution. This figure was redrawn from (Lee, 2008).

ReLU layer

Each convolutional layer is typically followed by a rectified linear unit (ReLU) layer. A ReLU layer applies an element-wise activation function to the feature map by setting all of the negative pixel values in the feature map equal to zero, that is $f(x) = \max(0, x)$ or

$$f(x) = \begin{cases} x, & \text{if } x \geq 0 \\ 0, & \text{if } x < 0. \end{cases} \tag{3.3}$$

The main purpose of the ReLU function (as depicted in Figure 3.9) is to introduce non-linearity to the input, therefore enabling neurons to learn non-linear representations. It is advantageous to incorporate ReLU layers into a CNN, since it reduces the training time by accelerating the convergence of the SGD algorithm (Krizhevsky et al., 2012). Since the ReLU function is not differentiable at the singular point $x = 0$, sub-derivatives facilitate the back-propagation algorithm (Simonyan et al., 2013; Goodfellow et al., 2016). The ReLU function has also been shown to perform better than other activation functions, like the logistic sigmoid and hyperbolic tangent functions, since it is less prone to saturation during training and mitigates the exploding/vanishing gradient problem (Krizhevsky et al., 2012).

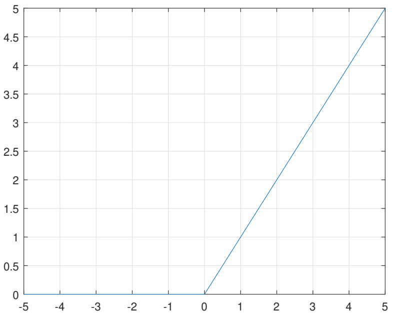


Figure 3.9: The ReLU function.

Pooling layer

Each ReLU layer is typically followed by a pooling layer. Spatial pooling, also referred to as sub-sampling, facilitates the reduction of the dimensionality (that is the height and width, but not the depth) of an input volume (feature map). Since images often have a so-called stationary property, which implies that prominent features within a certain sub-image are also likely to be relevant for other sub-images, large images may be adequately described by aggregate statistics for various non-overlapping sub-images. This process (referred to as pooling) is often achieved by simply computing the mean or maximum value of each sub-image, which is subsequently assigned (at the corresponding spatial position) to a sub-sampled version. In scenarios where 2 × 2 sub-images are considered, the height and width of the sub-sampled version will be half that of the original image (see Figure 3.10). A pooling layer therefore reduces the spatial resolution, the number of required parameters, computational complexity and overfitting (Goodfellow et al., 2016). It is important to note that pooling is applied to each channel of the rectified activation map independently.

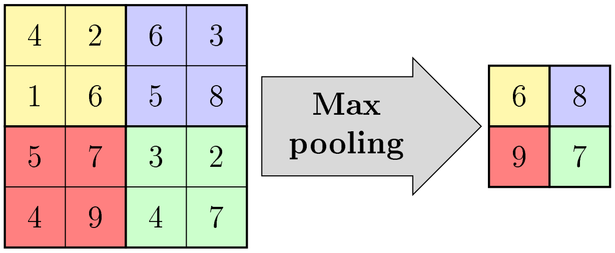


Figure 3.10: Example of the max pooling operation being performed on 2×2 sub-images with a stride of 2 pixels.

Fully-connected layer

Within the context of the FC layers, every neuron in a particular layer is connected to every neuron in the next layer (akin to the hidden layers conceptualised in Figure 3.3). The FC layers process the trained features (output of a flattened version of the final pooling layer) for the purpose of assigning the input image to one of several classes in an optimal way. The flattening operation transforms a three-dimensional input volume into a one-dimensional array. The last FC layer outputs an N-dimensional vector where N represents the number of classes. The aforementioned vector serves as input for a softmax layer.

Softmax layer

The softmax function, also referred to as the normalised exponential, takes any N-dimensional vector of arbitrary real values, that is the output of the last FC layer within the context of a CNN, and transforms it into an N-

dimensional vector of real values (probabilities) in the interval $[0, 1]$ that sum to 1. The softmax function is implemented as follows,

$$p_j = \frac{e^{y_j}}{\sum_{k=1}^{N} e^{y_k}} \qquad \text{for } j = 1, \ldots, N. \tag{3.4}$$

These probabilities may then be used to assign an input image to one of N disjoint classes (during the evaluation phase) and for the computation of the loss function (during the training phase).

3.4.2 Training and regularisation

During each forward pass of the training phase, information from the input values (within the context of a generic neural network) or the input image (within the context of a CNN) is propagated through the different layers of the network to obtain the predicted (network) output. The predicted output is then compared to the target (desired) output. The difference between the predicted and target output is referred to as the loss (error), which is subsequently used during the so-called backward pass to update the weights (parameters) within each layer of the network using a backpropagation algorithm like SGD (Li et al., 2012). The SGD algorithm iteratively updates the weights using a batch of training data so as to minimise the error function and may be denoted as follows

$$\mathbf{w}_{l+1} = \mathbf{w}_l - \eta \nabla E(\mathbf{w}_l), \tag{3.5}$$

where $\mathbf{w}$ denotes the weight vector, l the iteration number, η the learning rate and $\nabla E(\mathbf{w})$ the gradient of the loss function. It is important to note that within the context of SGD, the average loss across the entire batch (subset of the training data) is used to update the weights. The calibration of the learning rate is important within the context of the convergence of the network in the sense that being stuck in a local minimum or oscillations around the optimum value has to be avoided. The addition of a momentum term to the parameter update constitutes a popular strategy for mitigating the above-mentioned convergence issues (Sutskever et al., 2013). The SGD algorithm with momentum (SGDM) updates the weights as follows

$$\mathbf{w}_{l+1} = \mathbf{w}_l - \eta \nabla E(\mathbf{w}_l) + \gamma(\mathbf{w}_l - \mathbf{w}_{l-1}), \tag{3.6}$$

where the momentum value γ determines the influence of the previous update on the current iteration. It is important to note that during each training iteration a batch of data (that is a subset of the available training data) is presented to the network. The presentation of all of the available batches represents an epoch. All of the available batches are again presented to the updated network during subsequent epochs until sufficient convergence has been reached.

The remainder of this section is devoted to other techniques (that is the incorporation of batch normalisation and dropout layers) commonly employed within the context of deep learning for the specific purpose of combatting overfitting.

Regularisation refers to a number of techniques that are employed during training in order to prevent the network from overfitting to the training data and reduce the generalisation error. The concept of overfitting is illustrated in Figure 3.11. Overfitting occurs when a model learns noise and data specifically associated with the training data. This impacts negatively the model's ability to generalise to unseen data and typically leads to poor performance for the *evaluation* data. One of the techniques for combatting overfitting is to employ a *validation set*. Since the data in the validation set has not been used for training purposes, the local minimum in the error curve can be used as a stopping criterion.

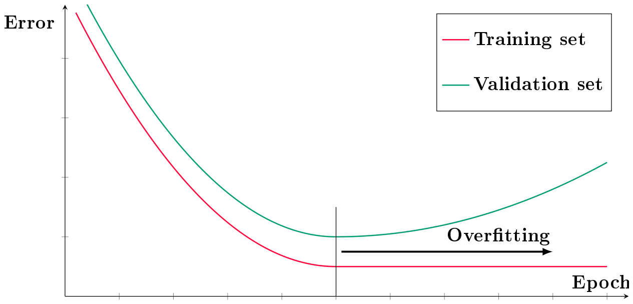


Figure 3.11: Conceptualisation of overfitting.

- A **batch normalisation (BN) layer** normalises an activation map across an entire batch, by subtracting the batch mean and dividing by the batch standard deviation. A BN layer therefore shifts and scales the relevant activation map by learnable parameters (Ioffe & Szegedy, 2015). BN layers are typically inserted between the convolutional and ReLU layers in order to accelerate convergence during training. BN has also been shown to reduce the sensitivity of the training procedure with respect to weight initialisation.

- A **dropout layer** simply removes a random set of activations within the layer in question by setting these activations equal to zero, thereby essentially forcing the network to be redundant. This implies that the network should be able to provide the correct classification or output even when certain activations are removed, therefore rendering the network less prone to overfitting, which in turn leads to better generalisation. A simple example of the dropout algorithm is provided in Figure 3.12.

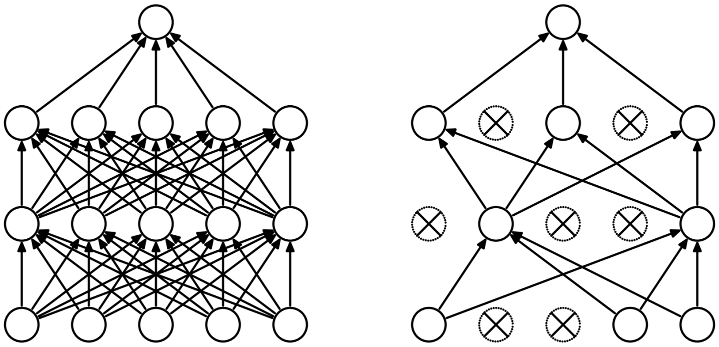
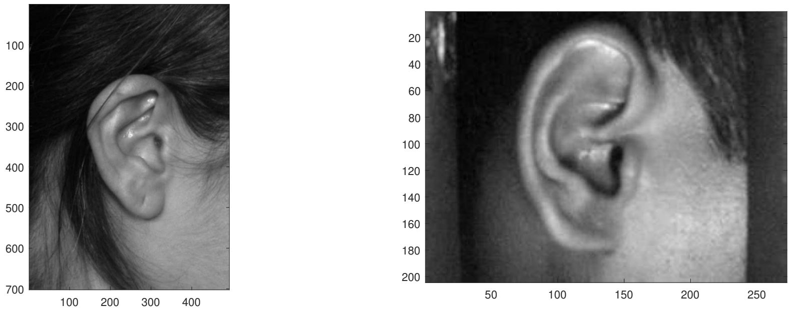


Figure 3.12: FC layers before and after the implementation of the droput algorithm. - Source: `https://www.doc.ic.ac.uk/~js4416/163/website/img/neural-networks/dropout.png` .

3.5 Detection of the region of interest

In this section, the architecture of and training protocol for the proposed CNN which facilitates the *automatic* detection of a suitable ROI that contains the entire ear shell, are discussed in detail. Recall that the purpose of the proposed CNN is to classify sub-images (patches) of an ear image as either *foreground* or *background*, which is followed by *morphological* post-processing. The *manually* selected ROI serves as a ground truth for evaluating this part of the proposed system.



(a) (b)

Figure 3.13: **(a)** A grey-scale image of size 702 × 492 pixels from the AMI ear database. **(b)** A grey-scale image of size 204 × 272 pixels from the IIT Delhi ear database.

Data

Two different datasets are independently considered for training respective CNNs and evaluating the proposed automated ROI detection protocol, that is the Mathematical Analysis of Images (AMI) ear database and the Indian Institute of Technology (IIT) Delhi ear database. These databases are discussed in more detail in Chapter 6. Examples from the AMI and IIT Delhi ear databases are depicted in Figures 3.13 (a) and (b) respectively. In the case of the AMI database, each image is first converted from RGB to grey-scale, while the images in the IIT Delhi database were originally captured in grey-scale format. For each dataset, ear images from different individuals are used for *training*, *validation* and *evaluation* purposes. The *training* set (seen data) is used to learn the parameters (weights) for the CNN in question, the *validation* set is used for avoiding overfitting by enforcing a stopping criterion, while the *evaluation* set is used to measure the performance of the CNN on unseen data.

While some results are presented here by means of selected examples, a more sophisticated *cross validation* algorithm is proposed and reported on in Chapter 6 in order to gauge the proficiency of the model, as well as its capability to generalise to unseen data.

Each ear image is subdivided into overlapping regions by sliding a 82×82 square window across the image in question. Each sub-image in the training and validation set is manually annotated as either positive (foreground) or negative (background). Typical examples of positive and negative sub-images from the AMI ear database are presented in Figures 3.14 and 3.15 respectively. These sub-images are saved to a database (that can be efficiently accessed) and serve as input for the network.

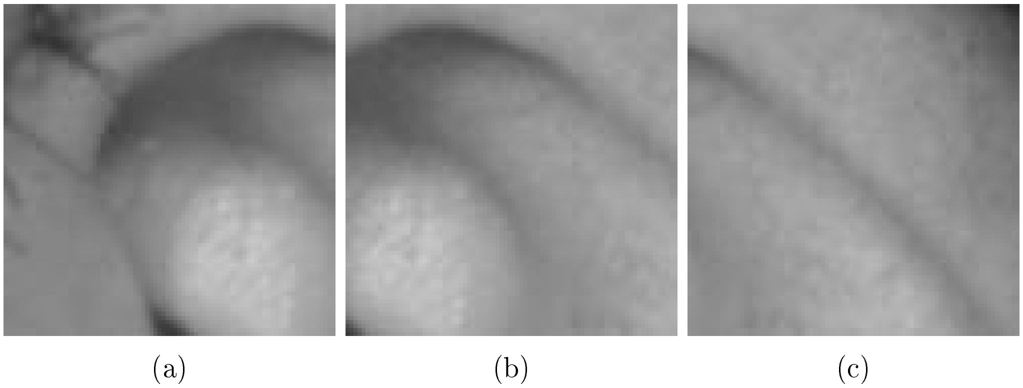


Figure 3.14: Examples of positive sub-images of size 82 × 82 pixels from the AMI ear database. These positively labelled sub-images are considered to be part of the foreground and contain contours typically associated with the shell of a human ear.

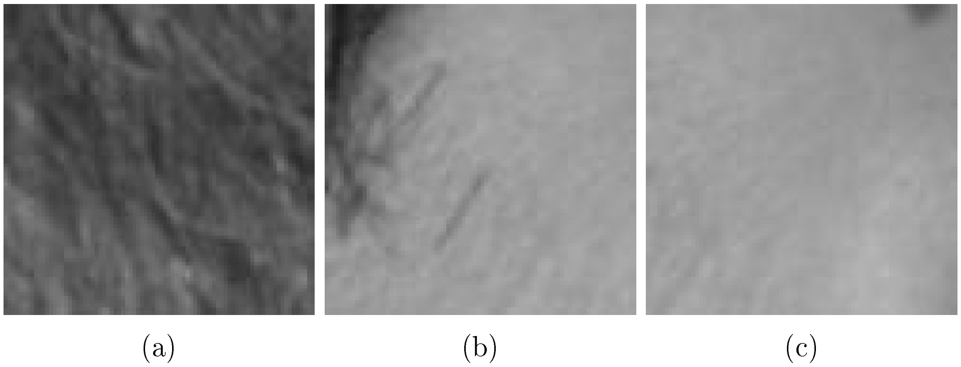


Figure 3.15: Examples of negative sub-images of size 82 × 82 pixels from the AMI ear database. These negatively labelled sub-images are considered to be part of the background and contain hair and/or homogeneous skin.

CNN architecture and training

The process of determining the appropriate hyper-parameters and best structure of a CNN for a particular task often relies on trial and error. After experimenting with a number of structures and hyper-parameters the network architecture depicted in Figure 3.16 was deemed optimal. The CNN consists of four convolutional layers, where each of these layers is followed by a BN, ReLU and/or max pooling layer. The final pooling layer is followed by two FC layers.

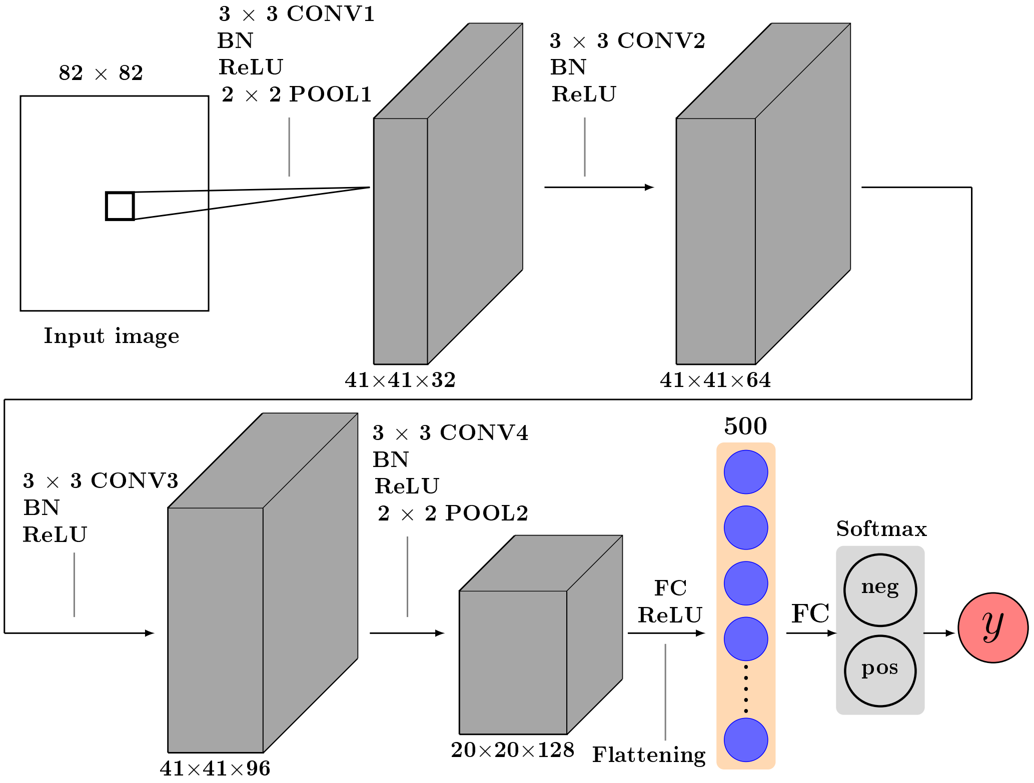


Figure 3.16: A depiction of the CNN architecture employed in this thesis for the purpose of automatically detecting a suitable ROI within an image of a human ear.

Table 3.1 provides a more detailed summary of the proposed network architecture.

The first convolutional layer (CONV1) processes an input image of size 82×82×1 with 32 different kernels (filters). Each kernel has a size of 3×3 pixels, while a convolutional stride of 1 pixel is employed. This layer therefore contains $82 \times 82 \times 32 = 215\ 168$ neurons, where each neuron has $3 \times 3{+}1{=}10$ trainable parameters (weights). The activation maps are normalised by incorporating a BN layer between the convolutional and ReLU layers in order to accelerate convergence and render the network less sensitive to parameter initialisation. The resulting activation maps are subsequently subjected to max pooling (POOL1) by considering 2× 2 sub-images and employing a stride of 2 pixels. This results in an output volume of size 41×41×32.

The second and third convolutional layers (CONV2 and CONV3) employ 64 and 96 different kernels (filters) respectively. Each kernel has a size of 3×3 pixels, while a convolutional stride of 1 pixel is employed. Each of the above convolutional layers are followed by a BN and ReLU layer, while no pooling layers are employed within the context of CONV2 and CONV3.

The fourth convolutional layer (CONV4) processes the output of CONV3 with 128 different kernels (filters). Each kernel has a size of 3×3 pixels, while a convolutional stride of 1 pixel is employed. After the application of a BN and ReLU layer, the output of CONV4 is down-sampled through max pooling (POOL2) by considering 2×2 sub-images and employing a stride of 2 pixels.

The first FC layer has 500 neurons. After the application of the ReLU function, dropout with a probability of 0.3 is enforced, thereby effectively setting the output of the relevant FC neurons to zero. The output of the second (last) FC layer is fed to a 2-way softmax function which results in a probability distribution across the positive and negative binary classes. The final layer constitutes a classification layer, which also calculates the cross entropy loss function during training.

The CNN is trained by employing the SGDM algorithm with a momentum value of $\gamma = 0.9$ and an initial learning rate of $\eta = 0.001$. During training, the initial learning rate is reduced after every 8 epochs. During each training iteration, which constitutes one forward and one backward pass, a batch that contains 128 training images is presented to the CNN. Recall that an epoch constitutes all the iterations required to traverse the entire training set. The training algorithm is run for a maximum of 100 epochs. In order to improve the convergence of the network, the weights of the first convolutional layer are initialised using normally distributed random numbers with a standard deviation of 0.0001.

It is important to note that the proposed CNN is trained from scratch. No fine-tuning of an existing pre-trained network (transfer learning) is therefore conducted.

After each epoch, the accuracy of the network is gauged by employing an independent validation set. In this way an early stopping criterion can be employed in order to avoid overfitting.

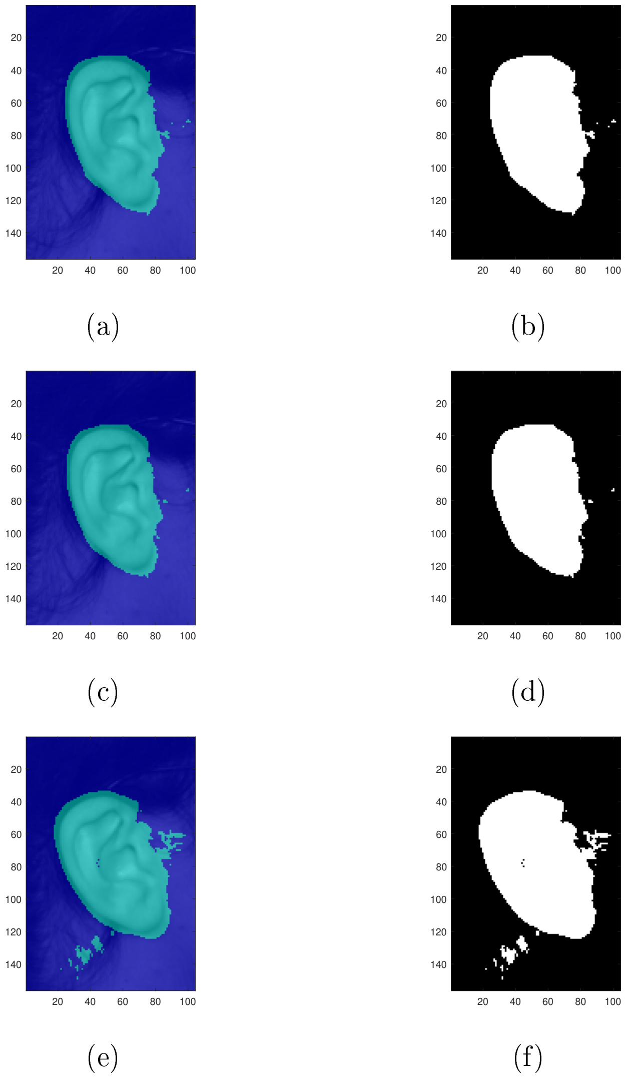
Layer	**Activation map**	**Kernel size**	**Stride (pixels)**	**Zero-padding (pixels)**
Input	$82 \times 82 \times 1$			
CONV1	$82 \times 82 \times 32$	3×3	1	1
POOL1	$41 \times 41 \times 32$	2×2	2	
CONV2	$41 \times 41 \times 64$	3×3	1	1
CONV3	$41 \times 41 \times 96$	3×3	1	1
CONV4	$41 \times 41 \times 128$	3×3	1	1
POOL2	$20 \times 20 \times 128$	2×2	2	
FC1	$1 \times 1 \times 500$			
Dropout (30%)	$1 \times 1 \times 500$			
FC2	2			

Table 3.1: The network architecture and hyper-parameters employed by the proposed system.

Selected illustrational results and post-processing

Selected results illustrating the proficiency of the proposed CNN-based model for the purpose of segmenting a human ear into foreground and background regions for the AMI ear database are presented in Figure 3.17. Figures 3.17 (a), (c) and (e) depict the probability that a sub-image belongs to the foreground (contains contours associated with the shell of an ear) as a shade of blue for ears belonging to three different individuals. After a threshold of 0.5 has been applied to the aforementioned probabilities, the acquired binary images are depicted in Figures 3.17 (b), (d) and (f) respectively. Although it is clear that the respective white regions (detected foreground) within the aforementioned binary images contain the entire ear shell, these images are still characterised by significant levels of noise, while the boundaries of the foreground regions are highly irregular.

In order to reduce the noise in the binary images depicted in Figure 3.17 and render the foreground boundaries more regular, the images are subjected to morphological post-processing. For this purpose morphological closing (with a circular structuring element of radius 10 pixels) is employed. The results are depicted in Figure 3.18.

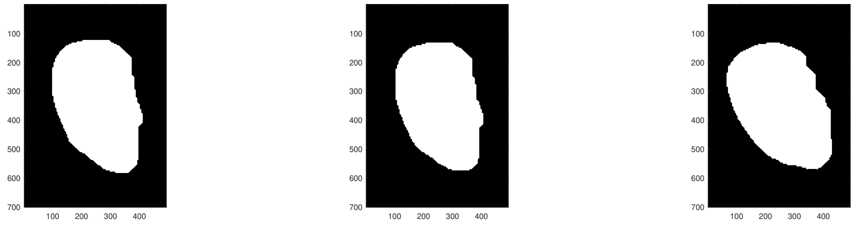


(a) (b)

(c) (d)

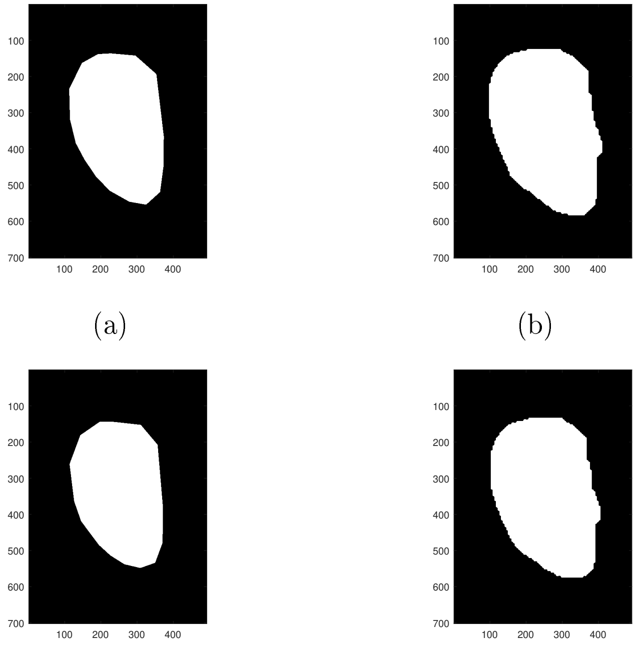
(e) (f)

Figure 3.17: (**Left**) Results of applying the proposed CNN-based model for the purpose of automated ROI detection within the context of the AMI ear database. The probability that a sub-image belongs to the foreground (contains contours associated with the shell of an ear) is represented by a shade of blue. (**Right**) Binary versions of the corresponding images on the left after a threshold of 0.5 has been applied.



(a) (b) (c)

Figure 3.18: The automatically detected ROIs after a morphological closing operation has been applied to the binary images depicted in Figure 3.17.



(a) (b)

(c) (d)

Figure 3.19: Qualitative depiction of the proficiency of the proposed automated ROI detection protocol within the context of the AMI ear database. (**Left**) Manually selected ROIs. (**Right**) Automatically detected ROIs corresponding to the images on the left.

In Figure 3.19 the manually selected and automatically detected ROIs are shown for comparison purposes. A quantitative analysis of the proficiency of the proposed automated ROI detection protocol is conducted and reported on in Chapter 6, in which case the manually selected ROI serves as a ground truth.

3.6 Concluding remarks

In this chapter, a CNN-based approach was proposed to facilitate automatic ROI detection within the context of ear-based biometric authentication. The CNN-based model was trained from scratch and (by visual inspection) achieves very satisfactory results in the sense that it appears to be robust with respect to noise, as well as variations in scale, location and orientation. The proposed CNN-based protocol, combined with appropriate morphological post-processing, is proficient in detecting a suitable ROI that contains the prominent contours associated with the entire ear shell. A quantitative analysis of the proposed ROI-detection protocol will be discussed in more detail in Chapter 6.

Chapter 4

Contour detection

4.1 Introduction

In this chapter a strategy is proposed for the detection of prominent contours associated with the entire ear shell. Recall that an algorithm capable of automatically detecting the region of interest (ROI) containing the entire ear shell was proposed in the previous chapter. One of the main objectives of the aforementioned ROI-detection protocol (or manual ROI selection) was to eliminate hair and jewellery, which are often associated with prominent edges, and may therefore be mistaken for ear contours. As depicted in Figure 4.1, the proposed contour detection protocol therefore involves four main stages, that is

(1) preprocessing (see Section 4.2),

(2) edge detection (see Section 4.3),

(3) post-processing (see Section 4.4), and

(4) ROI masking (see Section 4.5).

After all of the prominent contours, including those associated with hair and jewellery, have been detected within any given grey-scale input image (that is the output of stage 3), the detected ROI is employed as a mask in order to remove all of the undesired contours (as explained in Section 4.5). Features are subsequently extracted from the remaining (relevant) contours (that is the output of stage 4) by employing the Radon transform, as will be explained in the next chapter.

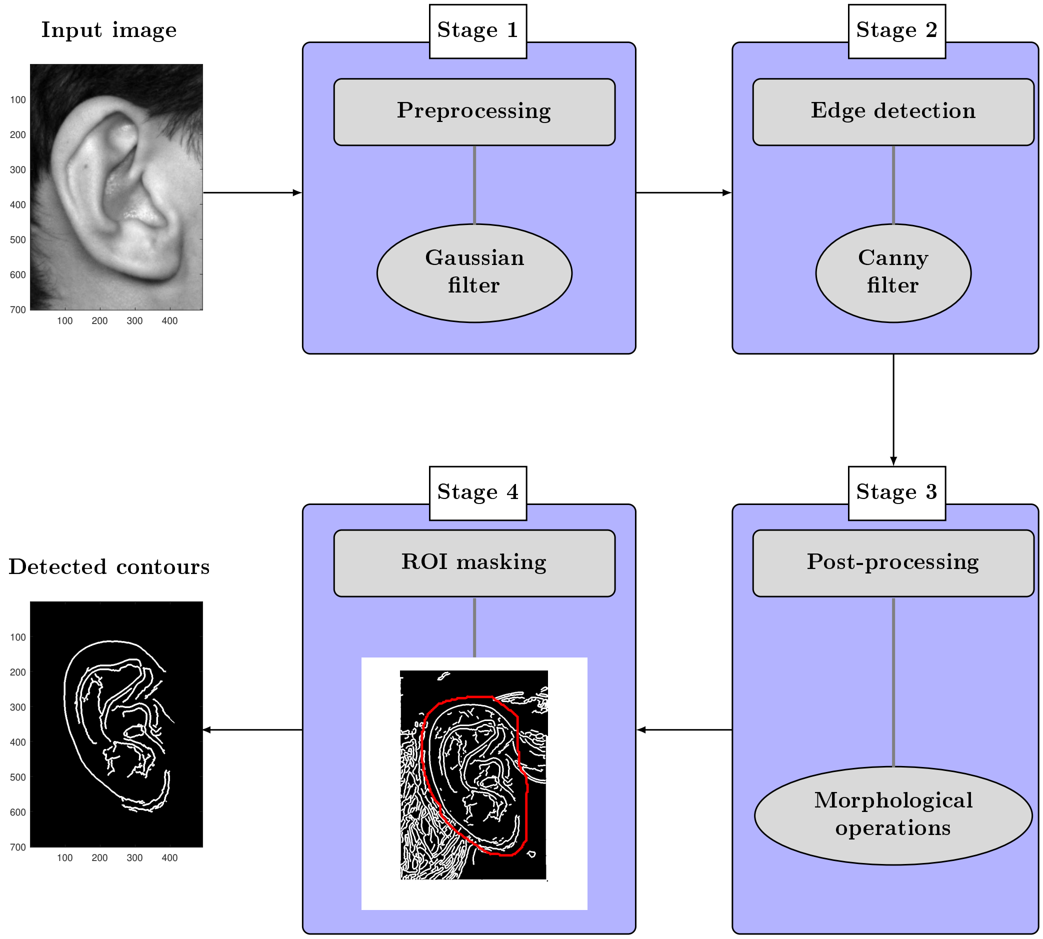


Figure 4.1: Schematic representation of the proposed contour detection protocol.

4.2 Preprocessing

The purpose of performing preprocessing techniques on an input image within the context of contour detection is to improve the quality of the image by suppressing undesired distortions while preserving sharp details such as edges associated with the ear shell. During preprocessing, input images from the Mathematical Analysis of Images (AMI) ear database are first converted from RGB to grey-scale format, while the input images in the Indian Institute of Technology (IIT) Delhi ear database were originally captured in grey-scale format. Within the context of linear spatial filtering, the application of a *Gaussian lowpass filter* to the aforementioned grey-scale images is therefore deemed appropriate.

A Gaussian lowpass filter is well suited for the purpose of suppressing the effect of noise in the ear images, while also preserving prominent edges to a large extent. An input ear image is smoothed by filtering the image with a Gaussian kernel (see Figure 4.2). This kernel is based on the following Gaussian function

$$g(m,n) = e^{-\left(\frac{m^2+n^2}{2\sigma^2}\right)}, \tag{4.1}$$

where σ denotes the standard deviation which determines the "spread" of the kernel, while m and n are positive integers. The resulting Gaussian kernel is normalised so that the filter coefficients sum to one.

The smoothing of an input ear image is therefore achieved by applying a Gaussian kernel of size 9×9, and with the standard deviation specified as $\sigma = 3$, to the image in question. The aforementioned parameters were found to be optimal in removing a sufficient amount of noise and non-prominent edges, while retaining prominent edges typically associated with the contours of the ear shell. Examples of ear images from the AMI database are presented in Figures 4.3 (a), (c) and (e), while the corresponding preprocessed (smoothed) versions are depicted in Figures 4.3 (b), (d) and (f) respectively.

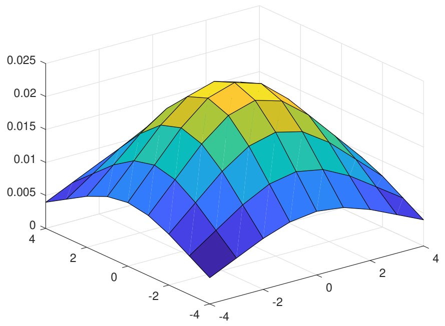
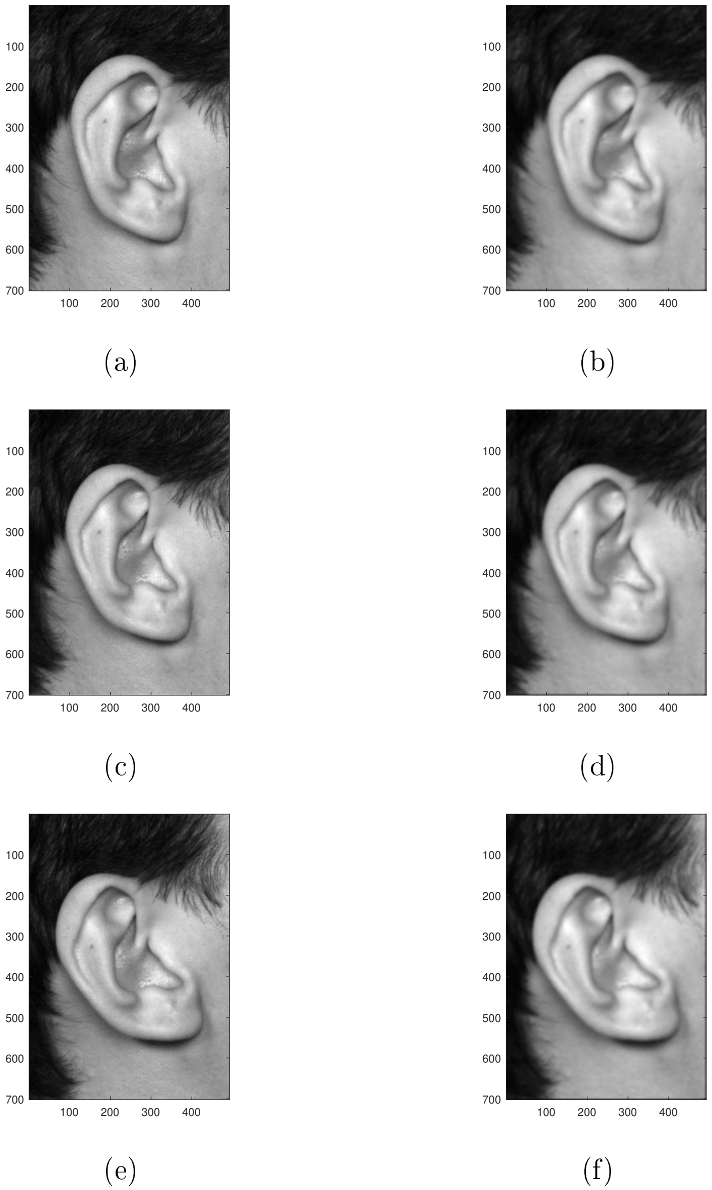


Figure 4.2: A Gaussian kernel of size 9×9 with $\sigma = 3$.



(a)

(b)

(c)

(d)

(e)

(f)

Figure 4.3: Preprocessing. (**Left**) Input images from the AMI ear database. These images are associated with the same individual, but the head is tilted in three different ways, that is down, front and up respectively. (**Right**) Smoothed versions of the corresponding images on the left after the application of the Gaussian filter depicted in Figure 4.2.

4.3 Edge detection

The Canny edge detector (Canny, 1986) is employed for the purpose of detecting prominent contours associated with the ear shell within the preprocessed (smoothed) versions of the ear images. The Canny algorithm involves a multi-step process (see Figure 4.4) that enables the inference of an optimal edge detector based on an estimation of the gradient magnitude and direction. This is followed by the identification of local maxima within an matrix (image) representing the gradient magnitude at different spatial positions.

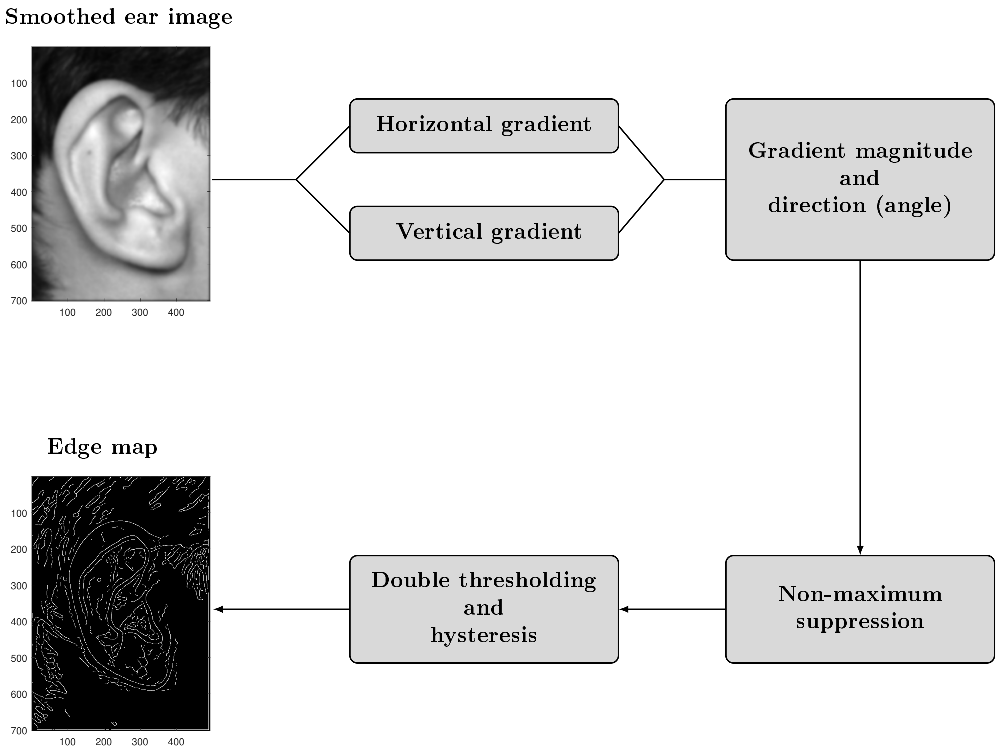


Figure 4.4: Conceptualisation of the Canny edge detection algorithm.

The magnitude and direction of the gradient are estimated from the *first derivatives* (in the horizontal and vertical directions) of the Gaussian kernel output, which are consequently represented by matrices (images).

A *non-maximum suppression* process is subsequently applied to the matrix (image) representing the gradient magnitude, therefore facilitating the identification of local maxima. These local maxima constitute candidate edges that are one pixel thick.

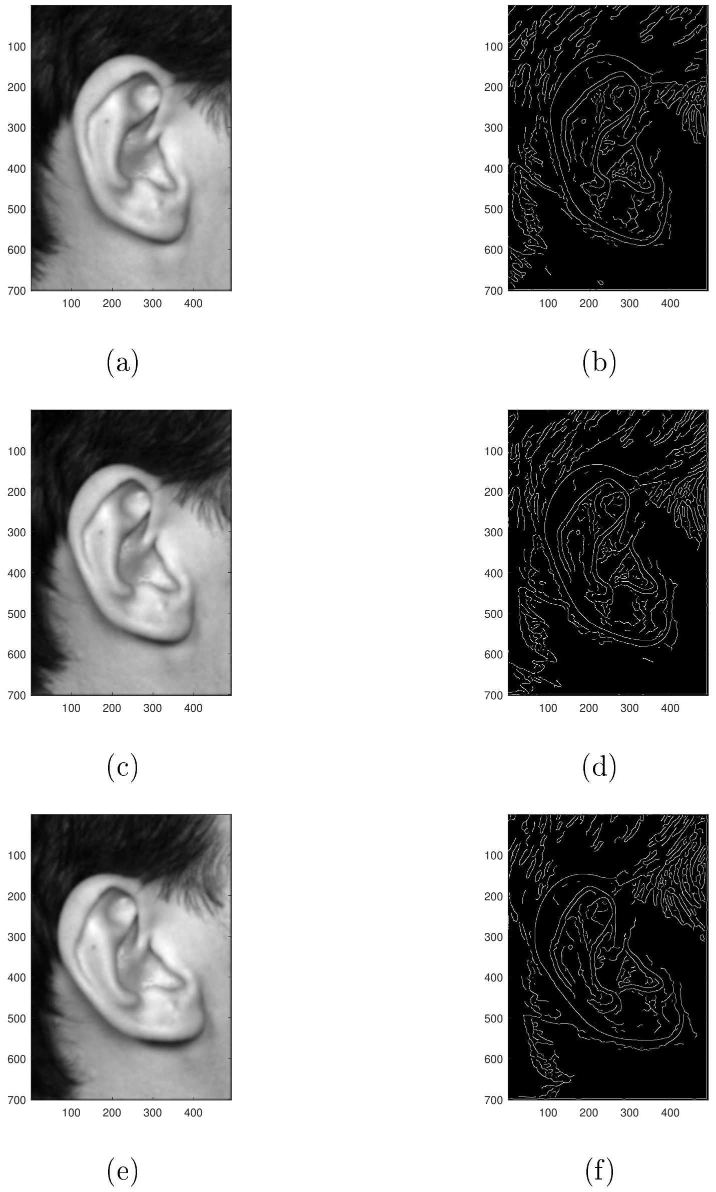
False edge segments that are associated with a low gradient magnitude (weak edges) may however still be present and require deletion. Isolated edge segments of average strength that are not collinear with strong edges (are associated with sufficiently different edge directions) are also deemed to be false segments and should therefore be deleted as well. Furthermore, edge segments of average strength that are collinear with strong edge segments should be connected with the aforementioned segments and therefore retained. All of the strong segments are retained without exception.

A *hysteresis tracking* process with different thresholds—these thresholds are obtained through Otsu's algorithm (Otsu, 1979)—is applied for the purpose of facilitating the aforementioned objectives of deletion and connection. Candidate edge pixels with a gradient magnitude below the lower threshold are deemed weak, while those with a gradient magnitude above the higher threshold are deemed strong. The other candidate edge pixels are deemed to be of average strength.

Typical edge detection results for the AMI ear database, through the application of the Canny algorithm to smoothed versions of the input images, are depicted in Figure 4.5.

4.4 Post-processing

After the application of the Canny algorithm, all of the the detected edges are one pixel thick. Since the Radon transform will be employed for the purpose of extracting suitable features from an ROI-masked edge image (as will be explained in the next chapter), it is advantageous that the edges in question are sufficiently thick so as to ensure easily perceptible peaks in projection profiles. Morphological dilation with a disk-shaped structuring element of radius 2 pixels are therefore applied to the original edge images. This ensures that the edges are three pixels thick and serves the additional purpose of connecting broken edges (see Figure 4.6).



(a)

(b)

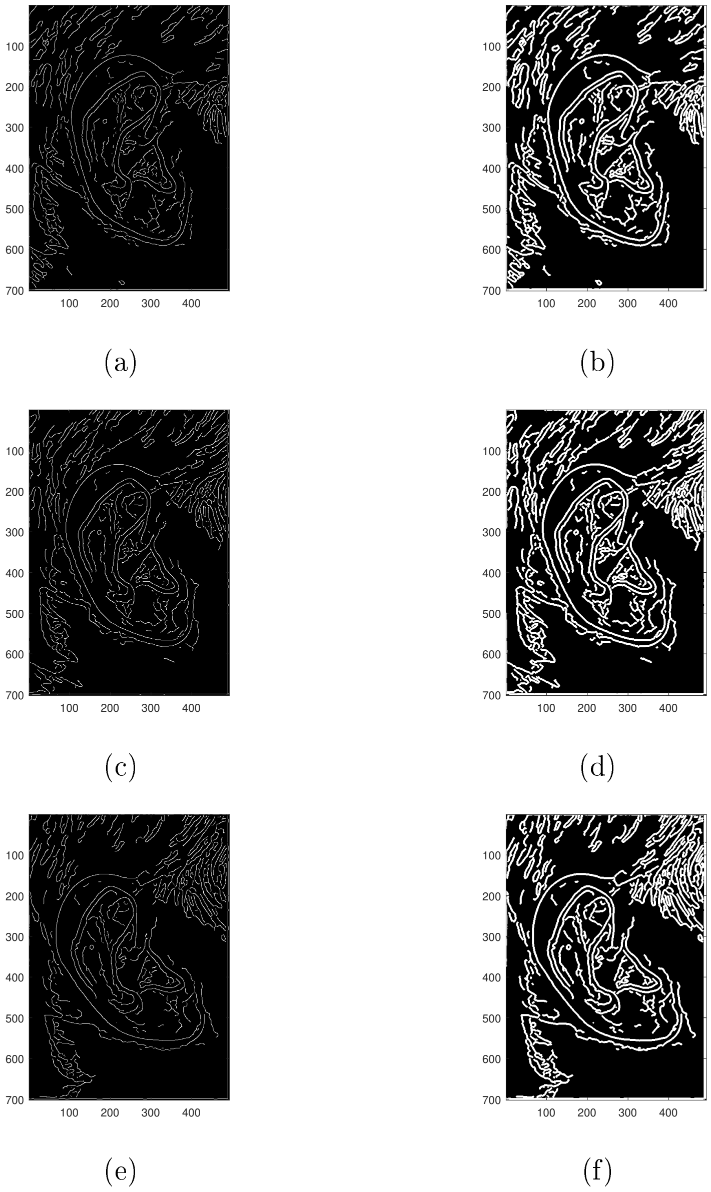
(c)

(d)

(e)

(f)

Figure 4.5: Edge detection. (**Left**) Preprocessed (smoothed) ear images from the AMI ear database. The images depicted in Figure 4.3 (right) have been reproduced here. (**Right**) Detected edges corresponding to the images on the left.



(a) (b) (c) (d) (e) (f)

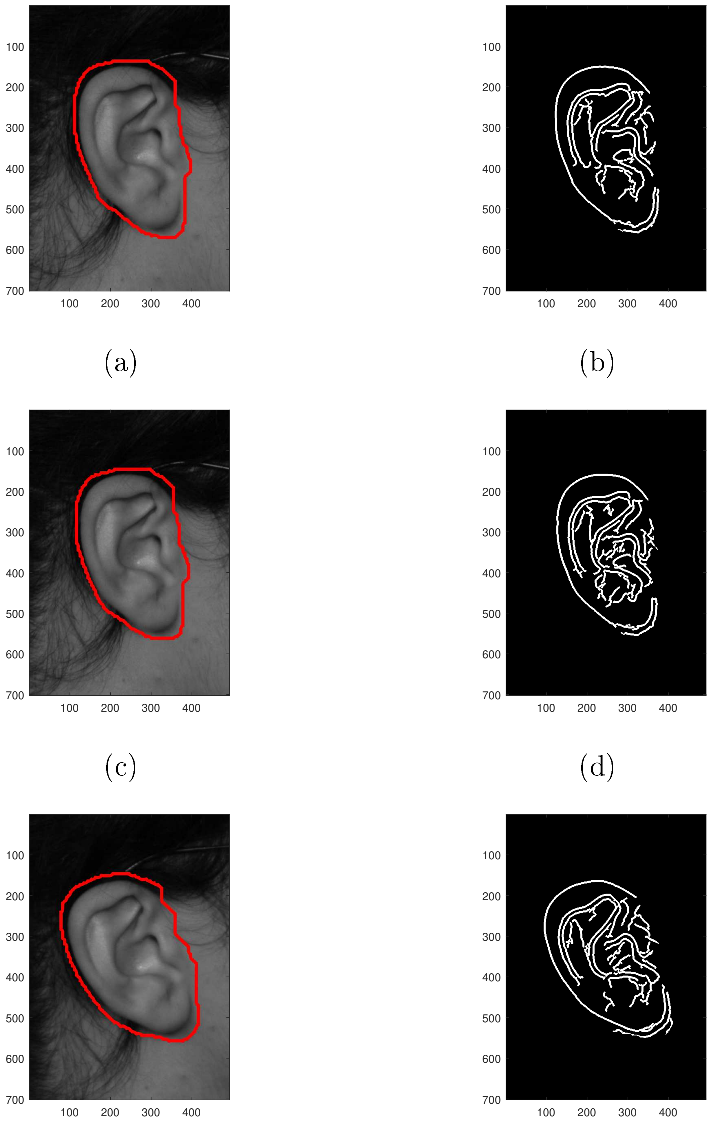
Figure 4.6: Post-processing. (**Left**) Original edge maps within the context of the AMI ear database. The images depicted in Figure 4.5 (right) have been reproduced here. (**Right**) Dilated edge images corresponding to the maps on the left.

4.5 ROI masking

During the final stage of contour detection the binary image constituting a manually specified or automatically detected ROI is employed as a mask in order to remove all of the edges *not* associated with ear contours. This is followed by the removal of all connected components with a length smaller than a predetermined threshold. In the case of the IIT Delhi ear database the image borders are also cleared, that is, all connected components that touch the border are removed.

The results for the three samples from the AMI ear database, which have been considered for illustrational purposes throughout this chapter, are presented in Figure 4.7.

The results for three samples from the IIT Delhi ear database are presented in Figure 4.8. The same protocol (as the one for the AMI ear database) has been followed, except for the fact that border clearing was required (see Figure 4.9)

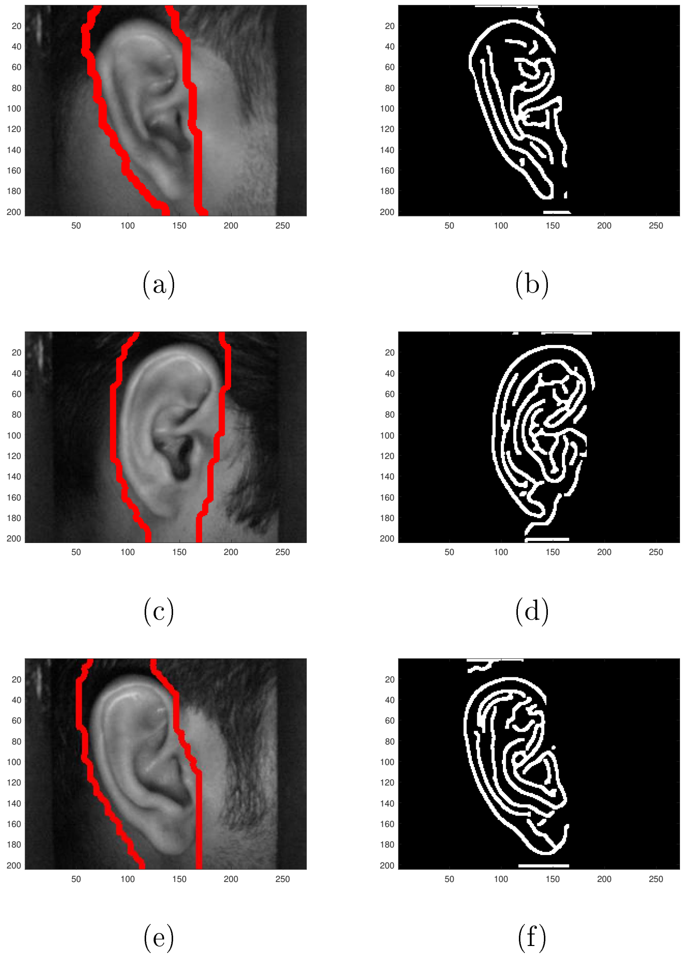


(a) (b)

(c) (d)

(e) (f)

Figure 4.7: ROI-masking. (**Left**) Original grey-scale versions of ear images from the AMI ear database. The images depicted in Figure 4.3 (left) have been reproduced here. The boundaries of the respective automatically detected ROIs are indicated in red. (**Right**) Detected prominent contours within the corresponding images on the left after ROI-masking and the removal of small connected components.

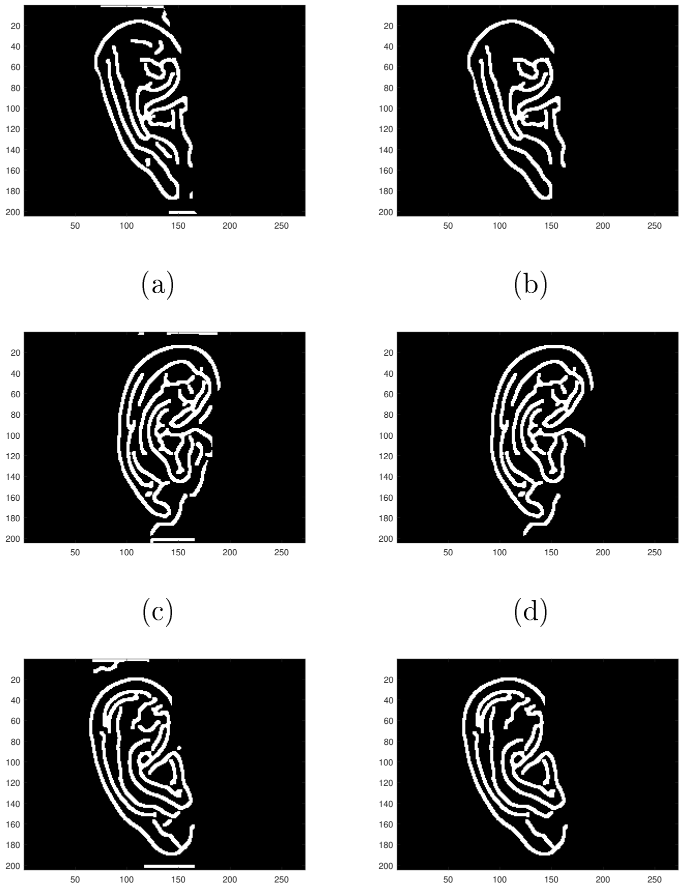


(a) (b)

(c) (d)

(e) (f)

Figure 4.8: ROI-masking. (**Left**) Original versions of ear images from the IIT Delhi ear database. These images are associated with three different individuals. The boundaries of the respective automatically detected ROIs are indicated in red. (**Right**) Detected prominent contours within the corresponding images on the left after ROI-masking.



(a) (b)

(c) (d)

(e) (f)

Figure 4.9: Border clearing. (**Left**) Detected prominent contours within the context of the IIT Delhi ear database after ROI-masking. The images depicted in Figure 4.8 (right) have been reproduced here. (**Right**) Detected prominent contours associated with the images on the left after the border has been cleared and small connected components have been removed.

4.6 Concluding remarks

In this chapter a protocol for detecting prominent contours associated with the shell of a human ear was proposed. This protocol is based on Canny edge detection, while ROI-based masking is employed in order to ensure that contours associated with hair and jewellery are discarded. An algorithm capable of automatically detecting such an ROI was proposed in Chapter 3.

Preprocessing and post-processing is conducted in order to ensure that noise and short edge segments are removed, and that the edges are sufficiently thick to facilitate feature extraction via the Radon transform. The proposed strategy that facilitates feature extraction and feature matching, as well as the proposed verification protocol, is discussed in detail in the next chapter.

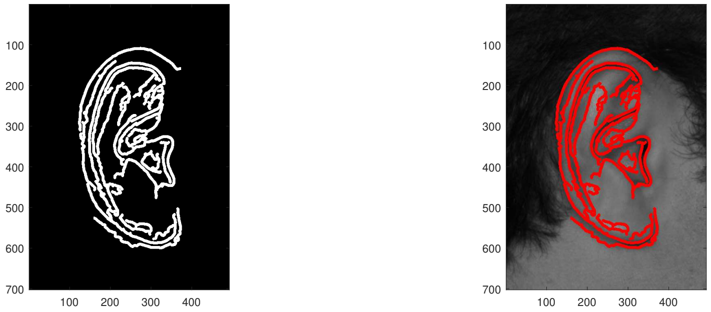
Chapter 5

Feature extraction, feature matching, and verification

5.1 Introduction

In this chapter a protocol is proposed for extracting suitable features from a binary image that contains prominent contours associated with the shell of a human ear (see Figure 5.1). The aforementioned protocol is based on the calculation of the discrete Radon transform (DRT) of the input image and is geared towards the detection of prominent *straight* lines within the image (see Section 5.2). This is followed by feature normalisation (see Section 5.3) and template matching (see Section 5.4). The proposed verification protocol, which is based on the construction of a so-called ranking verifier, is finally discussed in Section 5.5. The protocol for detecting prominent contours associated with the ear shell through the application of the Canny edge detector, followed by suitable post-processing, was discussed in the previous chapter.

The rationale behind the extraction of *manually tailored* features within the current context, followed by template matching, as opposed to the extraction of *learned* features by (for example) employing a neural network, lies in the fact that DRT-based features have proved to be very reliable in proficiently describing binary images that contain curved lines on a number of previous occasions, which include offline handwritten signature verification (Coetzer *et al.*, 2004) and hand vein-based biometric authentication (Beukes, 2018). The aforementioned features are therefore reliable within the context of both behavioural and physiological biometrics. It is also relatively simple



(a)

(b)

Figure 5.1: Prominent contours. (**Left**) A binary image that contains the detected prominent contours associated with the shell of a human ear. (**Right**) The detected prominent contours superimposed onto the original ear image.

to render the aforementioned features translation, scale and rotation invariant (as will be explained in Section 5.3).

In order to obtain reliable *learned* features through a process like deep learning, a large amount of training data is typically required for each individual enrolled into the system, which is *not* the case for the application being investigated in this thesis.

Feature extraction is a crucial step within the context of *pattern recognition.* A good feature set should satisfy the following requirements: Firstly, feature vectors should be highly *discriminative* in the sense that those features extracted from different samples of the same class should be relatively similar (associated with a small intraclass variance), while feature vectors extracted from different classes should differ substantially (associated with a large interclass variance). Secondly, feature vectors should be robust with respect to variations in scale, translation and orientation.

The proposed feature extraction protocol can be summarised as follows: Feature vectors are extracted from an image that contains prominent contours associated with the ear shell by applying the DRT to the image (see Section 5.2). The aforementioned feature vectors are subsequently normalised in such a way that they constitute scale, translation and rotation invariant representations of the original image (see Section 5.3). After appropriate normalisation, each *projection profile* (obtained from a different angle via the DRT) constitutes a *feature vector.* The Euclidean distance is finally employed for the purpose of feature matching (see Section 5.4). The distance between two ears is therefore quantified by the average Euclidean distance between the corresponding *normalised* feature vectors.

The proposed feature extraction protocol is conceptualised in Figure 5.2.

5.2 Feature extraction

The proposed feature extraction protocol is based on the calculation of the discrete Radon transform (DRT) of the binary contour image. These *global* features are able to describe distinguishable line segments that form part of the prominent contours associated with the shell of a human ear. The DRT of an image is obtained when multiple, parallel-beam projections of the image are calculated from *equally* distributed angles within an interval $\theta \in [0°, 180°)$ (Coetzer, 2005). Let $\mathbf{I}(m, n)$ denote the binary image of size $M \times N$ pixels containing the prominent contours associated with the shell of a human ear (see Figure 5.1 (a)), where the intensity of the i-th pixel is denoted by $I_i, i = 1, ..., MN$. The DRT of the image can be expressed as follows

$$R_j = \sum_{i=1}^{MN} \delta_{ij} I_i, \quad \text{for } j = 1, 2, \ldots, \beta\Theta, \tag{5.1}$$

where R_j denotes the jth beam-sum which constitutes the cumulative intensity of the pixels that overlap with the jth beam, β denotes the number of non-overlapping beams per angle, Θ represents the total number of angles and δ_{ij} denotes the weight indicative of the contribution of the i-th pixel towards the jth beam-sum. The aforementioned weight is typically proportional to the fraction of the surface area of the pixel that overlaps with the beam in question and equals one if the entire pixel falls within the beam. A detailed description of the theory and implementation of the DRT can be found in (Coetzer, 2005).

The aforementioned DRT of an input image therefore constitutes a matrix where each column of the DRT represents a *projection profile* of the input image acquired from a specified angle θ. The acquisition of a single projection profile from a specific angle θ is conceptualised in Figure 5.3. When converted into a grey-scale image, the DRT is often referred to as a *sinogram*.

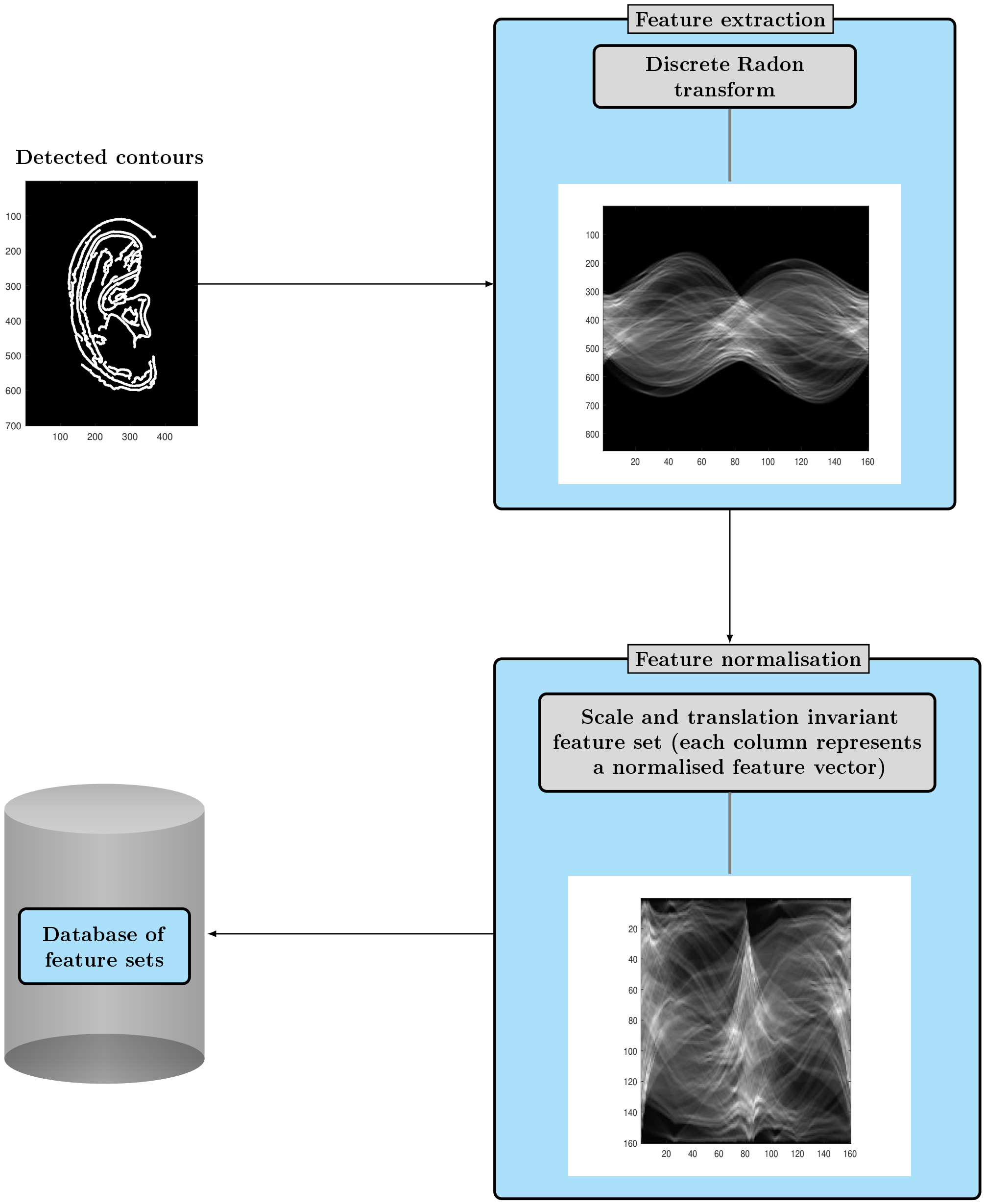


Figure 5.2: Schematic representation of the proposed feature extraction protocol. Rotational invariance is ensured by iteratively shifting the columns of two feature sets with respect to each other (with wrap-around) before they are matched, after which the alignment that results in the smallest average Euclidean distance between the corresponding *normalised* feature vectors is deemed optimal (see Section 5.3).

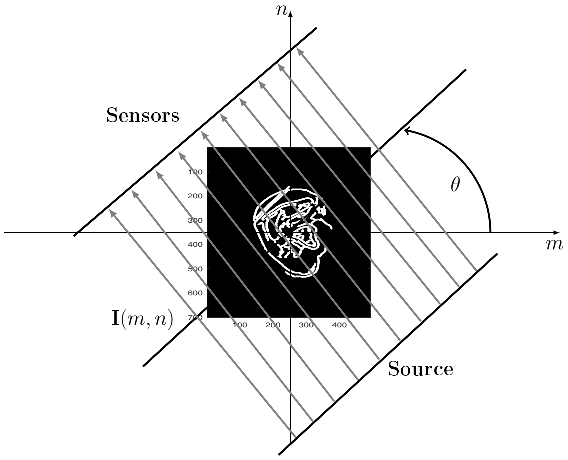
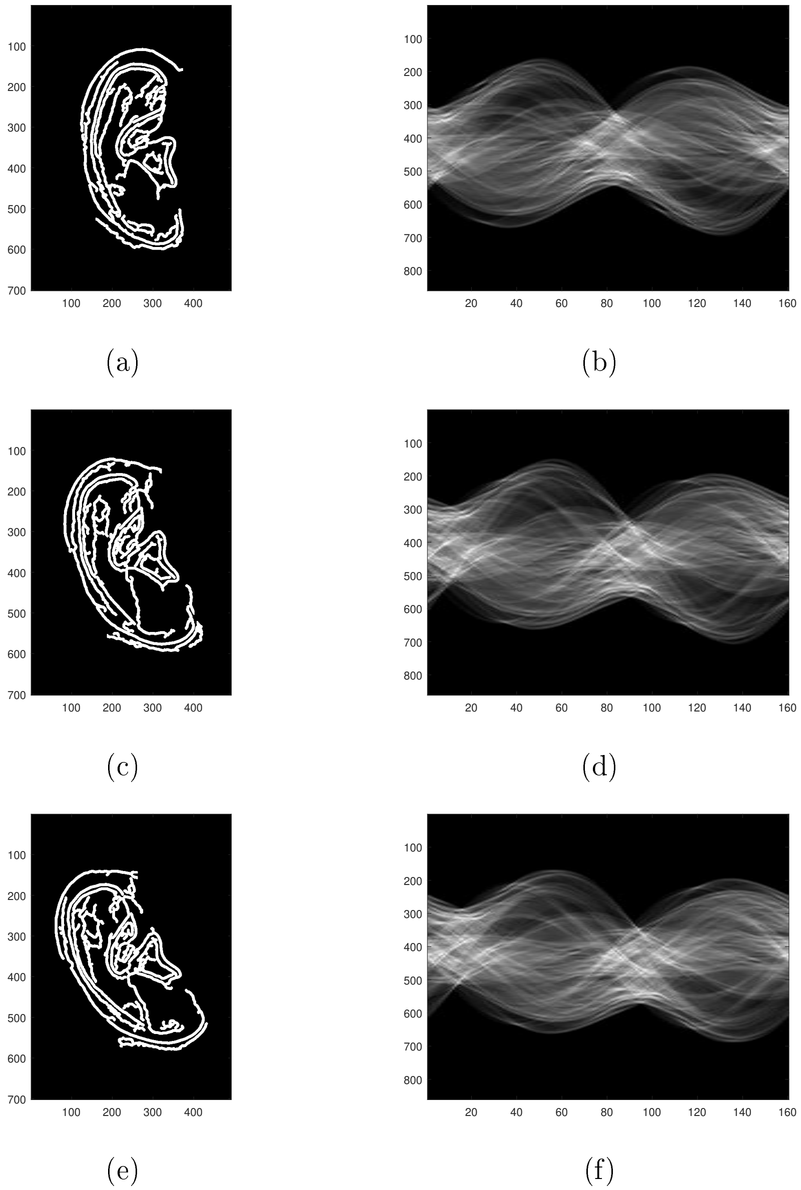


Figure 5.3: Conceptualisation of the acquisition of a single parallel-beam projection profile of a typical contour image from an angle θ. Although the terms "source" and "sensors" are applicable to computer-aided tomography (CAT) scans, the pixels that overlap with a specific beam are simply summed within the current context. An appropriate weight is assigned to pixels that only partially overlap with the beam in question.

In this thesis it is proposed that projection profiles are calculated for a full revolution, that is from angles in the interval $\theta \in [0°, 360°)$, instead of the required $\theta \in [0°, 180°)$. The projection profiles calculated from angles in the interval $\theta = [180, 360)$ simply constitute reflections of those already calculated in the interval $\theta \in [0°, 180°)$. The aforementioned (seemingly redundant) protocol is followed in order to facilitate *rotational invariance* (as explained in more detail in Section 5.3).

For illustrational purposes three samples of contour images from the Mathematical Analysis of Images (AMI) ear database are considered. These contour images were constructed from the originally acquired ear images belonging to the same individual, but with the head tilted in three different ways, that is downwards, towards the front and upwards, respectively. The respective DRTs (sinograms) of the aforementioned contour images are depicted in Figure 5.4. Each sinogram has 160 columns (the number of equally distributed angles in the interval $\theta \in [0°, 360°)$), where each column represents a projection profile.



(a)

(b)

(c)

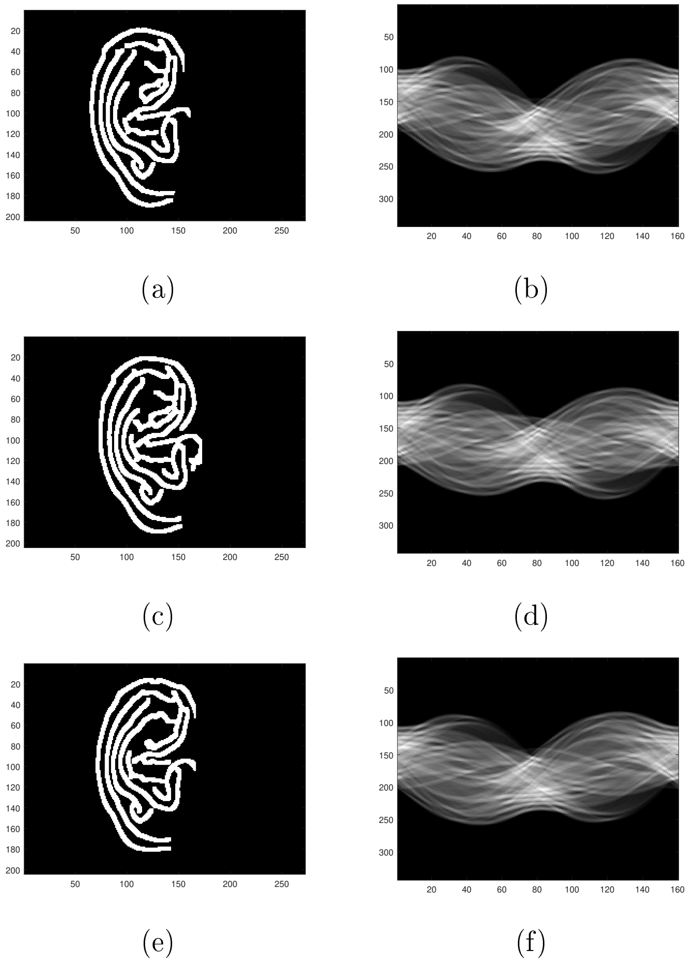
(d)

(e)

(f)

Figure 5.4: Contour images and their respective sinograms. (**Left**) Contour images within the context of the AMI ear database. These images belong to the same individual with the head tilted downwards, towards the front, and upwards, respectively. (**Right**) Sinograms corresponding to the contour images on the left. Each sinogram has $\Theta = 160$ columns and therefore represents $\Theta = 160$ projection profiles.

Within the context of the IIT Delhi ear database, three samples of contour images belonging to the same individual and their respective sinograms are presented in Figure 5.5 for illustrational purposes.



(a) (b)

(c) (d)

(e) (f)

Figure 5.5: Contour images and their respective sinograms. (**Left**) Contour images within the context of the IIT Delhi ear database. These images belong to the same individual. (**Right**) Sinograms corresponding to the contour images on the left. Each sinogram has $\Theta = 160$ columns and therefore represents $\Theta = 160$ projection profiles.

5.3 Feature normalisation

As previously mentioned, a good feature set within the current context should be unaffected by (that is remain unchanged despite) variations in scale, translation and orientation. Feature *normalisation* is therefore required and achieved by implementing a protocol that involves the following steps:

Firstly, in order to ensure scale and translational invariance, all of the zero-valued components are removed (decimated) from each projection profile (only the non-zero components are retained), after which the dimension of each projection profile is adjusted to a predefined value. This is achieved by compressing or expanding all of these decimated vectors to a fixed length of $\ell = 160$ through linear interpolation.

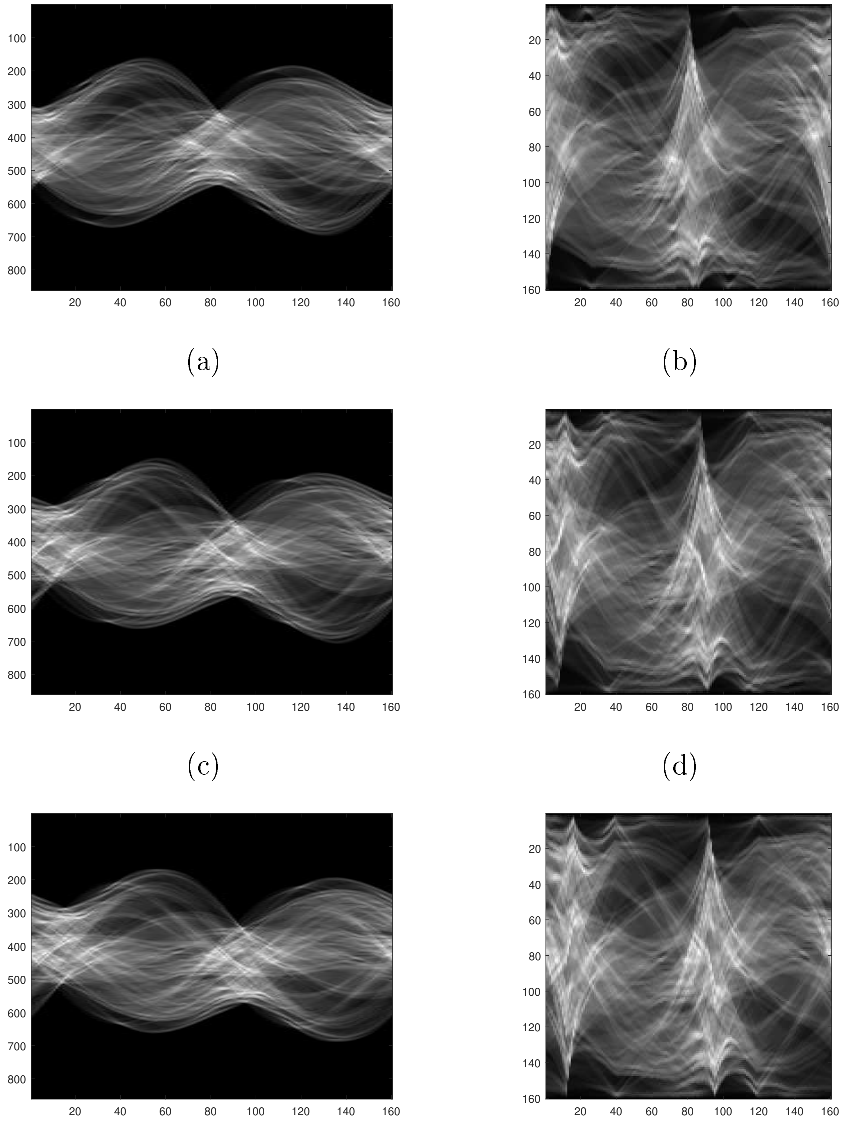
A new matrix (that replaces the original sinogram) is therefore constructed by packing the adjusted projection profiles as columns into the matrix. The intensities of the aforementioned new matrix are subsequently adjusted by dividing each matrix entry by the standard deviation across all of the entries. The columns of the resulting matrix therefore constitute *normalised* scale and translation invariant feature vectors that will subsequently be employed for the purpose of feature matching.

By considering the same samples as those associated with Figures 5.4 and 5.5, the respective results after the implementation of the proposed feature normalisation protocol are presented in Figures 5.6 and 5.7.

As previously mentioned, the DRT is calculated from angles that constitute a full revolution, that is angles within the interval $\theta \in [0°, 360°)$. This (seemingly redundant) strategy is followed in order to render the DRT periodic which allows for an elegant protocol for ensuring *rotational invariance*. This protocol is discussed within the context of feature matching in the following section.

5.4 Feature matching

In this section a feature matching protocol is proposed for quantifying the difference between two *feature sets* after appropriate *feature normalisation*. This protocol simply involves the calculation of the average Euclidean distance between the corresponding feature vectors that belong to the feature sets in question.



(a) (b)

(c) (d)

(e) (f)

Figure 5.6: Sinograms before and after normalisation. (**Left**) Sinograms for ear images from the AMI ear database. The images depicted in Figure 5.4 (right) have been reproduced here. (**Right**) Scale and translation invariant feature sets that correspond to the sinograms on the left. The columns of these feature sets (matrices) constitute normalised feature vectors.

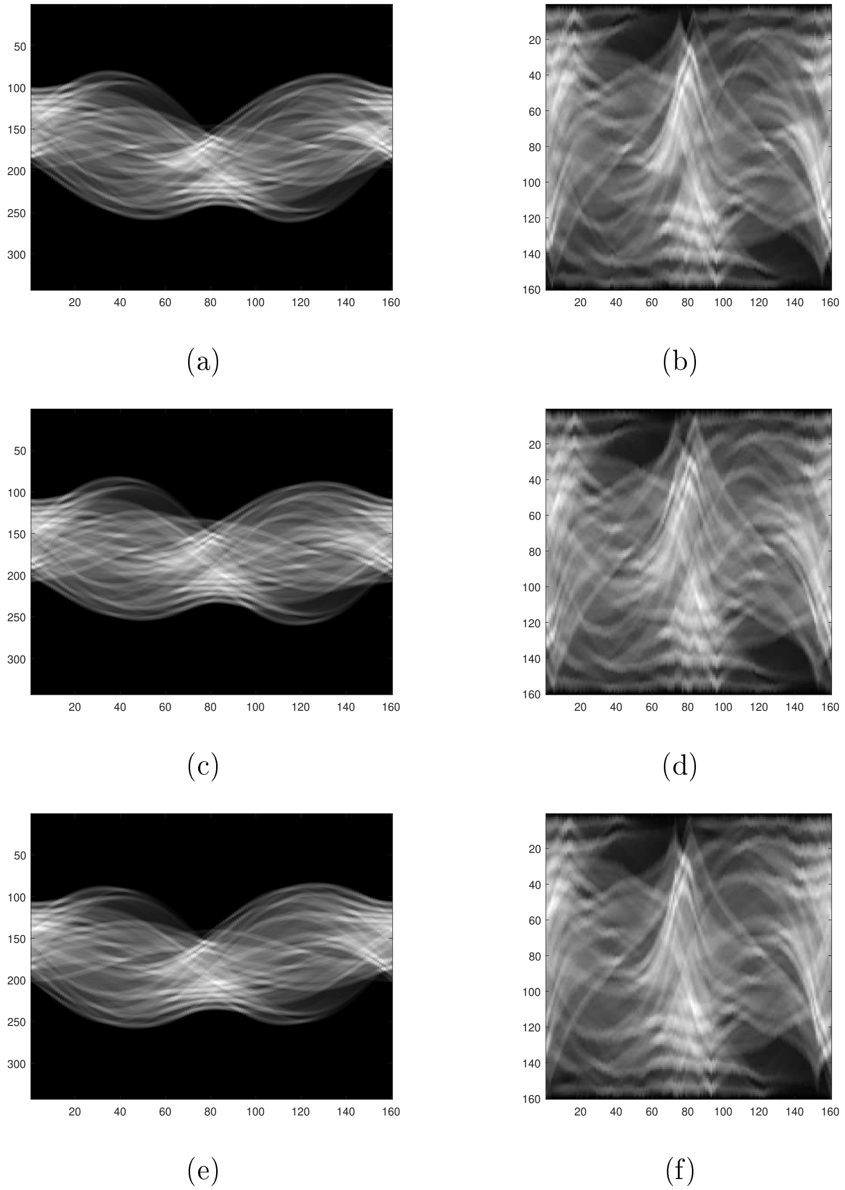


Figure 5.7: Sinograms before and after normalisation. (**Left**) Sinograms for ear images from the IIT Delhi ear database. The images depicted in Figure 5.5 (right) have been reproduced here. (**Right**) Scale and translation invariant feature sets that correspond to the sinograms on the left. The columns of these feature sets (matrices) constitute normalised feature vectors.

Rotational invariance is ensured by iteratively shifting (with wrap-around) the positions of the normalised feature vectors (columns) associated with a questioned feature set (matrix) with respect to those belonging to a reference (template) feature set. The average Euclidean distance between the corresponding feature vectors associated with the newly *aligned* feature sets are calculated after each iteration. During re-alignment, the position of (for example) the first feature vector associated with a questioned feature set is shifted in such a way that it corresponds to the position of every feature vector associated with the reference (template) feature set, while the sequence of the feature vectors within the respective feature sets is maintained. This involves a wrap-around strategy, which is made possible by the periodic nature of the DRT. When a feature vector occupies the last column of the questioned (shifting) feature set during the current iteration, it will occupy the first column of the feature set during the subsequent iteration. Every possible difference in orientation between the ear images being matched, which may theoretically encompass an entire revolution, is therefore permissible. The feature matching protocol, together with the protocol for ensuring rotational invariance, are conceptualised in Figure 5.8.

Selected illustrational results for feature matching

In order to provide additional perspective, some selected results that illustrate the usability of the feature matching technique proposed in this section are presented here within the context of the AMI ear database. A "zoomed" image of an ear that belongs to the claimed individual (who is typically facing front) is employed for reference purposes (see Figure 5.9). The respective dissimilarities (average Euclidean distances) between the aforementioned reference ear and three ears that also belong to the claimed individual, but with the head tilted in a three different ways (see Figure 5.10), are listed in Table 5.1. The respective dissimilarities (average Euclidean distances) between the above-mentioned reference ear (see Figure 5.9) and three ears that do not belong to the claimed individual (see Figure 5.11), are listed in Table 5.2.

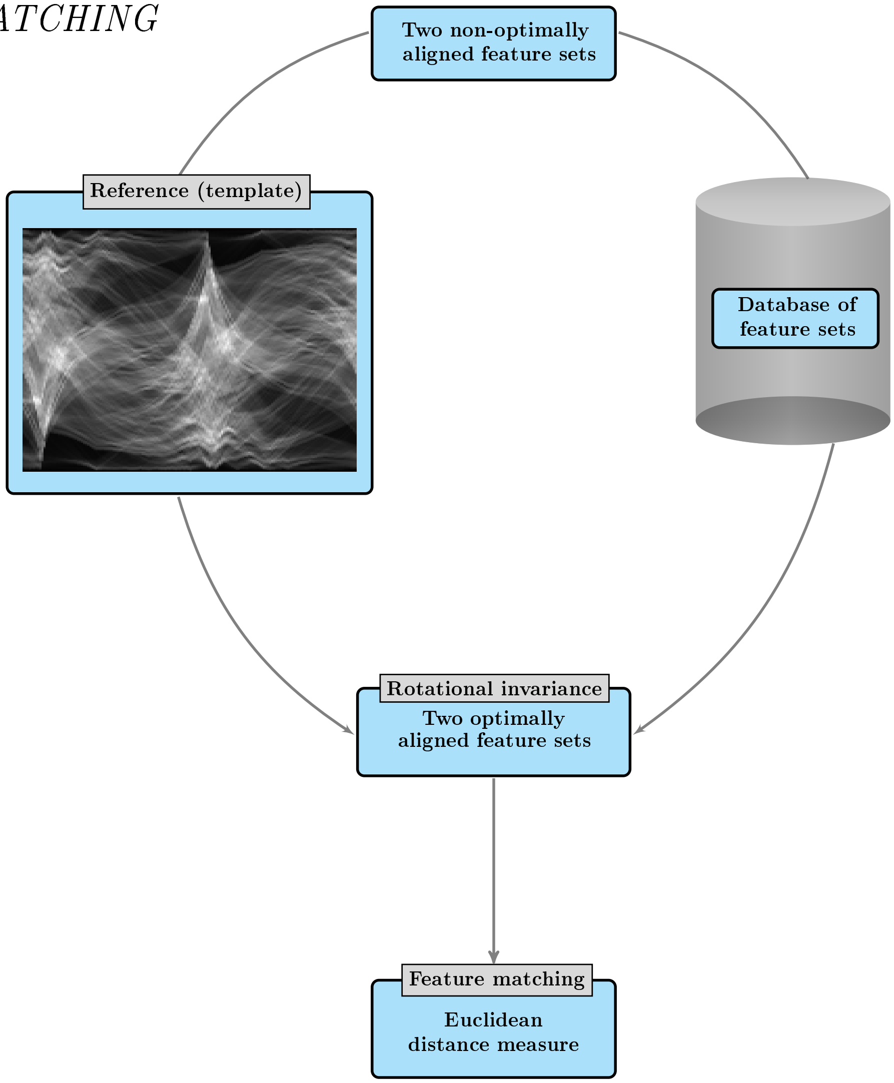

Two non-optimally aligned feature sets
Reference (template)
Database of feature sets
Rotational invariance
Two optimally aligned feature sets
Feature matching
Euclidean distance measure

Figure 5.8: Schematic representation of the proposed feature matching protocol that also ensures rotational invariance.

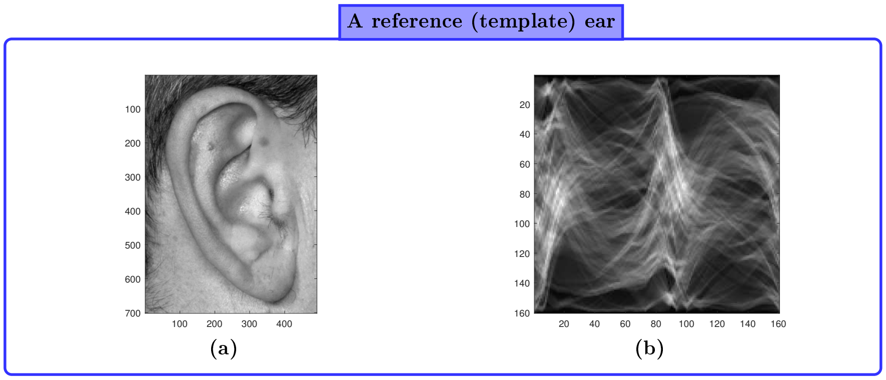


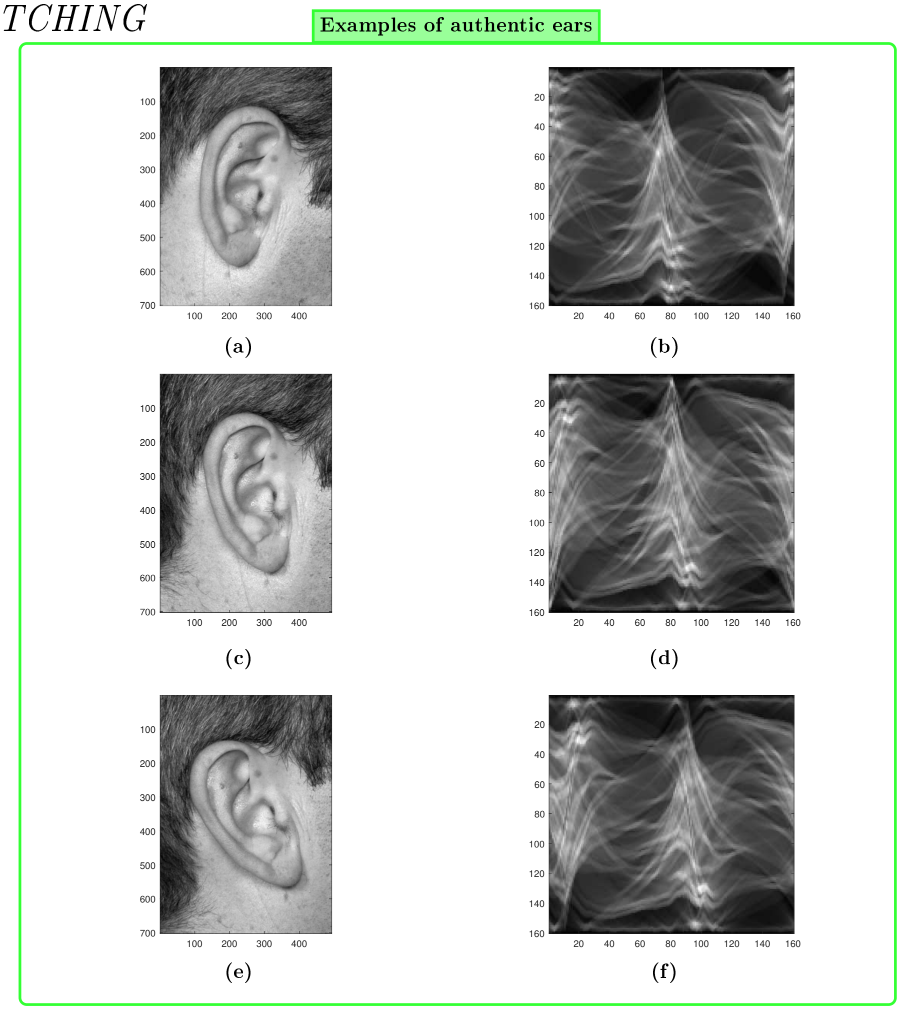
Figure 5.9: **(a)** An example of a reference (template) image from the AMI ear database. **(b)** The corresponding feature set.

Reference ear	Questioned authentic ear	Dissimilarity
Figure 5.9 (a)	Figure 5.10 (a) (down)	10779.696
Figure 5.9 (a)	Figure 5.10 (b) (front)	11077.026
Figure 5.9 (a)	Figure 5.10 (c) (up)	11758.422

Table 5.1: The respective dissimilarities (average Euclidean distances) between the reference (template) ear depicted in Figure 5.9 and the questioned authentic ears depicted in Figure 5.10.

Reference ear	Questioned imposter ear	Dissimilarity
Figure 5.9 (a)	Figure 5.11 (a) (down)	21102.056
Figure 5.9 (a)	Figure 5.11 (b) (down)	19532.987
Figure 5.9 (a)	Figure 5.11 (c) (down)	24087.153

Table 5.2: The respective dissimilarities (average Euclidean distances) between the reference (template) ear depicted in Figure 5.9 and the imposter ears depicted in Figure 5.11.

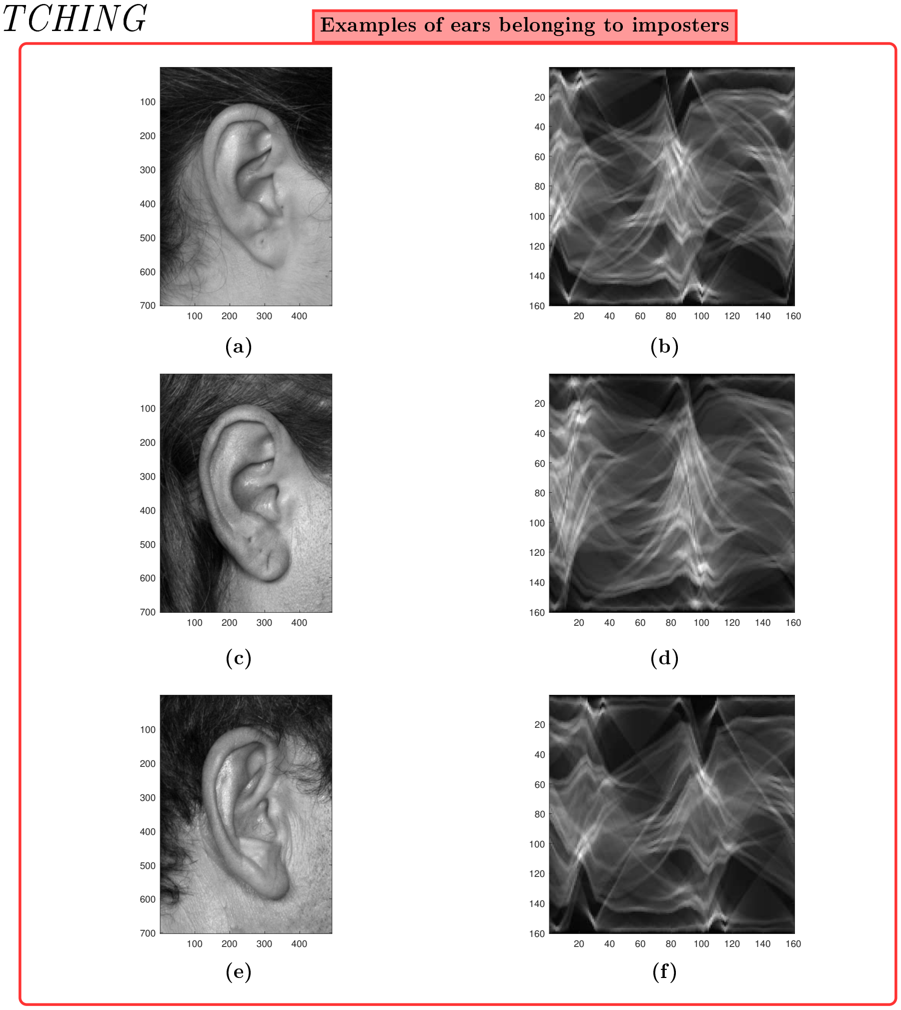


(a) (b)

(c) (d)

(e) (f)

Figure 5.10: Feature matching. (**Left**) Original grey-scale ear images associated with the same individual as the one referred to in Figure 5.9, but with the head tilted in three different ways, that is downwards, towards the front and upwards, respectively. (**Right**) Scale and translation invariant feature sets corresponding to the images on the left. Rotational invariance is achieved by iteratively shifting the positions of the columns of these questioned matrices one pixel towards the right (with wraparound), while the positions of the columns of the reference (template) matrix, depicted in Figure 5.9 (b), remain unchanged. The dissimilarity between the ears in question constitutes the average Euclidean distance between the corresponding feature vectors associated with the optimally aligned feature sets.



(a)

(b)

(c)

(d)

(e)

(f)

Figure 5.11: Feature matching. (**Left**) Original grey-scale ear images associated with three different individuals than the one referred to in Figure 5.9. (**Right**) Scale and translation invariant feature sets corresponding to the images on the left. Rotational invariance is achieved by iteratively shifting the positions of the columns of these questioned matrices one pixel towards the right (with wraparound), while the positions of the columns of the reference (template) matrix, depicted in Figure 5.9 (b), remain unchanged. The dissimilarity between the ears in question constitutes the average Euclidean distance between the corresponding feature vectors associated with the optimally aligned feature sets.

It is clear from Tables 5.1 and 5.2 that the dissimilarities (average Euclidean distances) associated with questioned ears belonging to imposters are approximately twice as large as those associated with questioned ears belonging to the claimed individual.

5.5 Verification

In this section a protocol is proposed for the purpose of establishing whether a claim that a questioned ear belongs to a certain individual is authentic or fraudulent. This is achieved by comparing the questioned ear to a number of ears in the database. The aforementioned ears include a *reference* ear that is known to belong to the *claimed* individual, as well as ears that belong to other so-called *ranking* individuals.

The dissimilarity between the questioned ear and the reference ear, as well as the respective dissimilarities between the questioned ear and those belonging to the ranking individuals are placed in a list, with the smallest dissimilarity at the top of the list and the largest dissimilarity at the bottom of the list. Verification is subsequently based on the relative position (ranking) of the dissimilarity associated with the reference ear in the aforementioned list. The claim may for example be deemed valid if and only if the reference ear has a ranking of one, in which case the questioned ear is accepted as authentic. Alternatively, the system may be rendered more lenient by requiring that the reference ear be ranked higher than or equal to some threshold value, where the threshold value is greater than one. An optimal threshold value may also be determined empirically by employing a suitable data partitioning protocol as will be explained in the next chapter.

5.6 Concluding remarks

In this chapter the DRT was proposed for the purpose of extracting features from a contour image associated with the shell of a human ear. The resulting feature set was normalised in such a way that it constitutes a translational, rotational and scale invariant representation of the contours in question. Feature matching is achieved by calculating the average Euclidean distance between the corresponding feature vectors associated with the respective feature sets. Verification is finally achieved by constructing a ranking verifier.

In the next chapter the two datasets considered in this research are discussed in detail. This is followed by an outline of the experimental protocol. Exhaustive experiments are also conducted in order to evaluate the automated region of interest (ROI)-detection strategy, as well as the semi-automated and fully automated ear-based authentication systems, proposed in this thesis.

Chapter 6

Experiments

6.1 Introduction

In this chapter exhaustive experiments are conducted in order to gauge the proficiency of the proposed automated region of interest (ROI) detection algorithm, as well as the respective proficiencies of the semi-automated and fully automated ear-based biometric authentication systems developed in this thesis. Recall that the ROI is manually specified in the case of the semi-automated system, while the ROI is automatically detected in the case of the fully automated system. The aforementioned experiments are conducted on two independent datasets. These datasets are described in detail in Section 6.2. The experimental protocol that is followed for each of the individual experiments is outlined in Section 6.3. The experimental results are presented and quantitatively analysed in Section 6.4. Finally, an overview of the software developed and hardware utilised in this thesis is presented in Section 6.5.

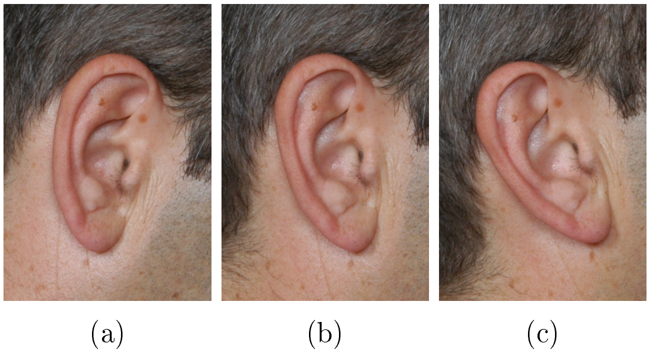
6.2 Data

The experiments are conducted on (1) the Mathematical Analysis of Images (AMI) ear dataset and (2) the Indian Institute of Technology (IIT) Delhi ear dataset. The aforementioned two independent datasets are discussed in the following subsections.

6.2.1 AMI ear dataset

The AMI ear database was acquired by Esther Gonzalez for her PhD thesis in Computer Science (Gonzalez *et al.*, 2012). This dataset consists of RGB images that were captured under the same illumination conditions from a fixed camera position and involves 100 different individuals aged between 19 and 65 years. For each individual, *seven* images (six images of the right ear and one image of the left ear) were captured at a resolution of 702×492 pixels, while the head is tilted in a specific way:

- Three of these images were captured from the right and therefore contain the right ear (see Figure 6.1). The first image, which is referred to as DOWN, was captured while the head is tilted downwards. In the case of the second and third images, which are referred to as FRONT and UP, the head is tilted towards the front and upwards respectively.



(a) (b) (c)

Figure 6.1: Examples of images from the AMI ear database. These images contain the right ear of the same individual, while the head is tilted in three different ways, that is downwards, towards the front and upwards, respectively.

- A further two images were also captured from the right and therefore contain the right ear (see Figure 6.2). The first of these, which is referred to as RIGHT, was captured while the head is tilted towards the camera. In the case of the second image, which are referred to as LEFT, the head is tilted away from the camera,

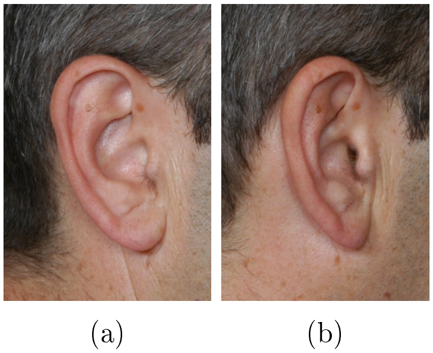


Figure 6.2: Examples of images from the AMI ear database. These images contain the right ear of the same individual, while the head is tilted in two different ways, that is towards the right and towards the left, respectively.

- Another image, which is referred to as BACK, was captured from the left while the head is tilted towards the front and therefore contains the left ear (see Figure 6.3 (a)),

- The final image was captured from the right by employing a different focal length and therefore contains the right ear (see Figure 6.3 (b)). This image (referred to as ZOOM) constitutes a zoomed in version of image referred to as FRONT.

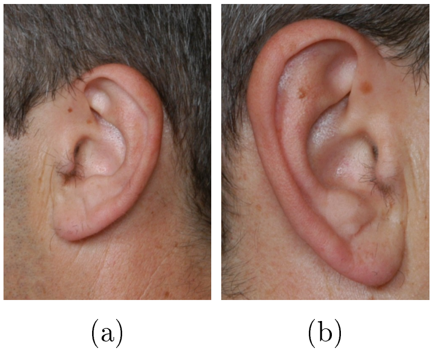


Figure 6.3: Examples of images from the AMI ear database that belong to the same individual. **(a)** This image was captured from the left while the head is tilted towards the front and therefore contains the left ear. **(b)** This image contains the right ear and constitutes a zoomed in version of the image depicted in Figure 6.1 (b).

6.2.2 IIT Delhi ear dataset

The IIT Delhi ear database contains touchless images provided by the Hong Kong Polytechnic University (Kumar, 2007). This dataset consists of ear images that were acquired over a period of nine months from October 2006 to June 2007 at IIT Delhi under different indoor lighting conditions. These ear images belong to students and staff members aged between 14 and 58 years. The dataset consists of 375 grey-scale images that belong to 125 different individuals. *Three* ear images were therefore captured for each individual while the head is tilted in a different way. Each of these images has a resolution of 272×204 pixels. Figure 6.4 depicts samples of ear images from the IIT Delhi ear database.

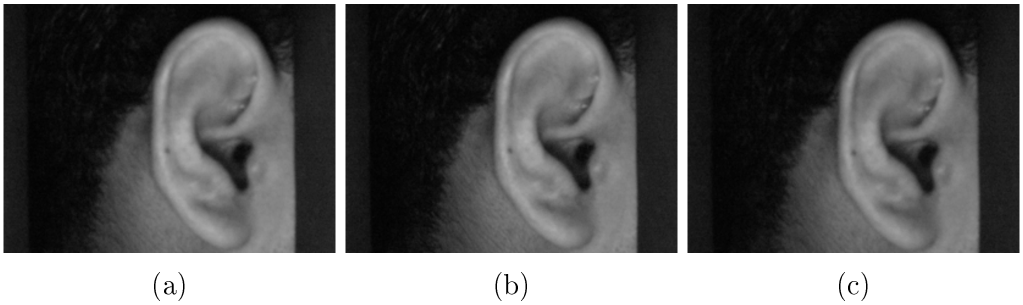


Figure 6.4: Examples of images from the IIT Delhi ear database. These images are associated with the same individual, but the head is tilted in three different ways.

6.3 Protocol

Recall that a ranking verifier is employed in this thesis for the purpose of establishing the authenticity of a questioned ear. All of the subsequent experiments are conducted independently on both of the datasets introduced in the previous section. The experimental protocol for three main (independent) experiments are discussed in the remainder of this section. The aforementioned experiments are dichotomized as follows:

(1) **Experiment 1.** This experiment investigates the proficiency of the proposed *semi-automated* ear-based authentication system. The aforementioned system employs a protocol in which the ROI is *manually* specified,

followed by feature extraction, template matching and verification. Recall that a questioned ear is matched to a sample belonging to the claimed individual, as well as to ears that belong to a number of ranking individuals. The resulting distances are subsequently ranked from small to large. Two sub-experiments are therefore conducted:

(a) **Experiment 1A.** In this sub-experiment a questioned ear is only accepted when the distance associated with the reference sample belonging to the claimed individual is the smallest, in which case the questioned ear has a ranking of one. This is referred to as the *rank-1 scenario*.

(b) **Experiment 1B.** In this sub-experiment a questioned ear is accepted when it has a ranking that is better than or equal to a specific optimal ranking. The optimal ranking (which may be greater than one) is estimated by considering a separate set of optimisation individuals. This is referred to as the *optimal ranking scenario*.

(2) **Experiment 2.** This experiment investigates the proficiency of the proposed *automated* ROI detection algorithm.

(3) **Experiment 3.** This experiment investigates the proficiency of the proposed *fully automated* ear-based authentication system. The aforementioned system employs a protocol in which the ROI is *automatically* detected through deep learning, followed by feature extraction, template matching and verification. Due to time constraints, only the rank-1 scenario is investigated.

It is important to note that this investigation is based on the assumption that only *one* positive sample is available for each individual enrolled into the system. The aforementioned single sample therefore serves as a reference sample for the corresponding individual during template matching. Each of the experiments employs k-fold cross-validation in order to gauge the proficiency of the proposed systems in an unbiased way.

6.3.1 Experiment 1: Semi-automated ear-based authentication

In this section an experimental protocol is outlined for the purpose of evaluating the proficiency of the proposed semi-automated ear-based authentication system, after the ROI has been manually specified.

Recall that the AMI ear dataset contains ears that belong to 100 different individuals. For each individual in the AMI ear dataset, only *four* (of the available *seven*) images are considered for the experiments outlined in this chapter. For each individual, three ear images, that is the images referred to as FRONT, UP and DOWN, are employed for evaluation purposes, while *one* ear image, that is the image referred to as ZOOM, is used as a template or reference sample for ranking purposes. A total number of 300 images (75%) are therefore used for evaluation purposes within the context of the current experiment, while 100 images (25%) are used for ranking purposes. The ears referred to as BACK, RIGHT and LEFT are not considered for experimental purposes in this thesis since it is presumed that the opposite side of the head is supported by (for example) a wall, which does not allow for the head to be tilted towards or away from the camera. The scope of this thesis is also restricted to the evaluation of right ears as outlined in Section 1.2.

Recall that the IIT Delhi ear dataset contains ears that belong to 125 different individuals. For each individual in the IIT Delhi ear dataset, *two* ear images are used for evaluation purposes within the context of the current experiment, while *one* ear image is used as a template or reference sample for ranking purposes.

As mentioned previously, the current experiment is dichotomized into Experiment 1A and Experiment 1B.

Experiment 1A

In this sub-experiment a *questioned* ear is matched to a *reference* sample that is *known* to belong to the *claimed* individual, as well as to templates that belong to other so-called *ranking* individuals. The *questioned* ear is accepted as authentic if and *only* if the *reference* ear has a *ranking* of one. Within the context of this sub-experiment both of the ear datasets considered in this study are partitioned into two sets based on the individuals that are employed for *evaluation* and *ranking* purposes.

Within the context of the AMI ear database, a 100-fold cross-validation procedure is conducted as conceptualised in Figure 6.5. For each fold, three positive samples ($\oplus$) belonging to a single claimed individual and 150 negative samples ($\ominus$) belonging to 50 other individuals are employed for evaluation purposes, while 49 templates belonging to 49 ranking individuals and one reference sample belonging to the claimed individual are used for ranking purposes.

For a specific fold, 2550-fold cross-validation is also conducted across the evaluation individuals. For the first 50 of these sub-folds, the claimed individual (⊕) is kept fixed, while the 50 other individuals (⊖) iteratively provide negative evaluation samples (see Figure 6.6). For the next 50 sub-folds, the claimed individual (⊕) is moved one place towards the left and again kept fixed, after which the process is repeated. In this way it is ensured that the data is *balanced*, in the sense that an *equal* number of positive and negative samples are evaluated.

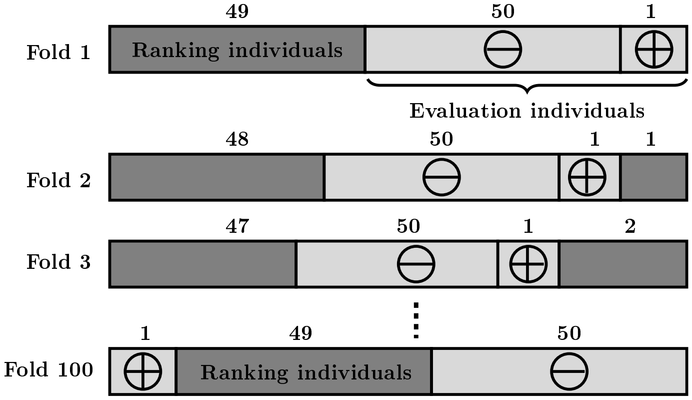


Figure 6.5: Conceptualisation of the proposed data partitioning protocol for the AMI ear dataset within the context of Experiment 1A. Within each fold, 49 templates (that is the images referred to as ZOOM) associated with 49 ranking individuals constitute the ranking set (dark gray), while three images (that is the images referred to as FRONT, UP and DOWN) associated with each of the respective 51 evaluation individuals constitute the evaluation set (light gray). One of the aforementioned evaluation individuals (⊕) constitutes the claimed individual. Technically, one image (that is the image referred to as ZOOM) associated with the claimed individual is also employed for ranking purposes.

A similar 125-fold cross-validation protocol is employed within the context of the IIT Delhi ear dataset as conceptualised in Figure 6.7. For each fold, the sub-folds for the evaluation individuals are specified in a similar way than is the case for Figure 6.6, except for the fact that the total number of sub-folds is 5700.

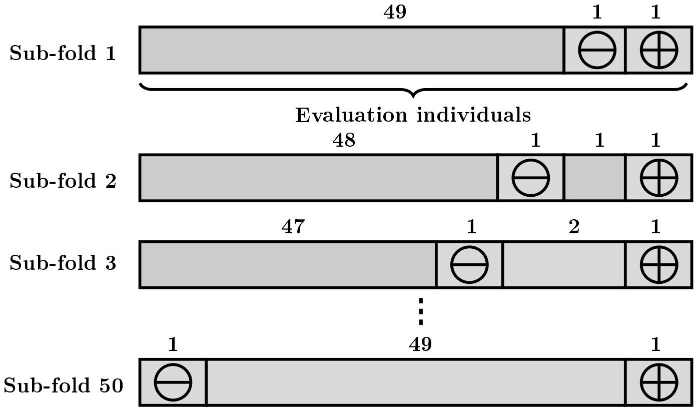


Figure 6.6: Conceptualisation of the proposed data partitioning protocol for the evaluation individuals, within the context of Experiment 1A and the AMI ear dataset. For the subsequent sub-folds (not shown), the claimed individual (⊕) occupy other positions.

Experiment 1B:

In this sub-experiment, the system is rendered more flexible such that a questioned ear is accepted when it has a ranking that is better than or equal to a very specific optimal ranking, which may be greater than one. This optimal ranking is empirically determined by employing a suitable data partitioning protocol. Within the context of this sub-experiment both of the ear datasets considered in this study are partitioned into a *ranking* set, an *optimisation* set and an *evaluation* set. The aforementioned data partitioning and cross validation protocol is conceptualised in Figures 6.8 and 6.9 for the AMI ear dataset and IIT Delhi ear dataset respectively. For each fold, the sub-folds within the context of the optimisation *and* evaluation individuals are defined according to a similar protocol than the one conceptualised in Figure 6.6.

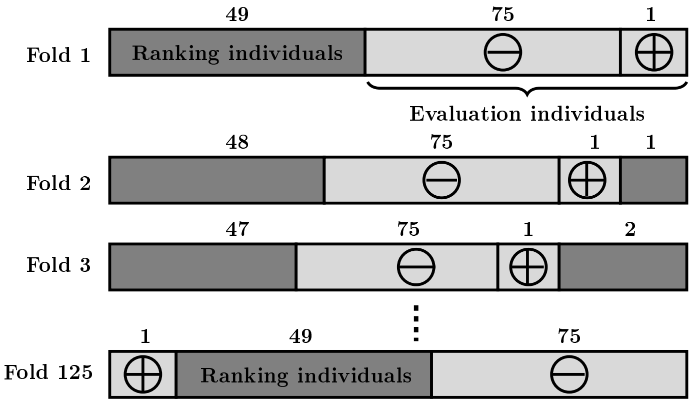


Figure 6.7: Conceptualisation of the proposed data partitioning protocol for the IIT Delhi ear dataset within the context of Experiment 1A. Within each fold, templates (that is the first ear for each individual) associated with 49 ranking individuals constitute the ranking set (dark gray), while two images (that is the second and the third ears) associated with each of the 76 respective evaluation individuals constitute the evaluation set (light gray). One of the aforementioned evaluation individuals constitutes the claimed individual. Technically, one image (that is the first image) associated with the claimed individual is also employed for ranking purposes.

The proposed protocol for Experiment 1B is now described in more detail:
As is the case for Experiment 1A, a 100-fold and 125-fold cross validation procedure are conducted for the AMI and IIT ear databases respectively. For a specific fold, cross-validation is conducted across the respective optimisation individuals according to the protocol conceptualised in Figure 6.6. The estimated optimal ranking based on *both* the average error rate (AER) *and* the equal error rate (EER), as defined in Table 6.1, is then employed to authenticate the ears associated with the evaluation individuals. For a specific fold, cross-validation is *again* conducted across the respective evaluation individuals using the protocol conceptualised in Figure 6.6.

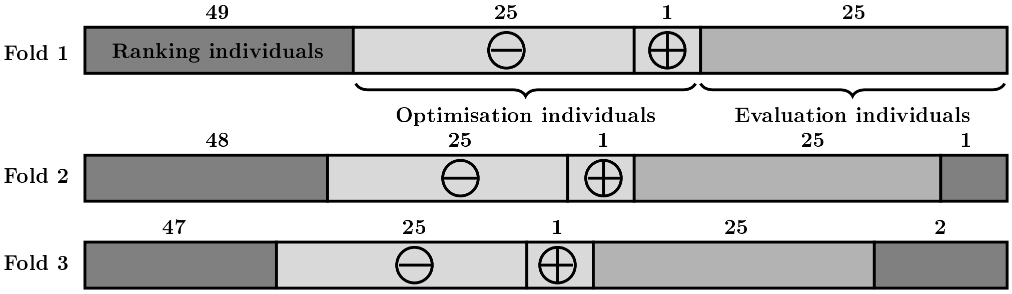


Figure 6.8: Conceptualisation of the first three (out of a total of 100) folds of the proposed data partitioning protocol for the AMI ear dataset within the context of Experiment 1B.

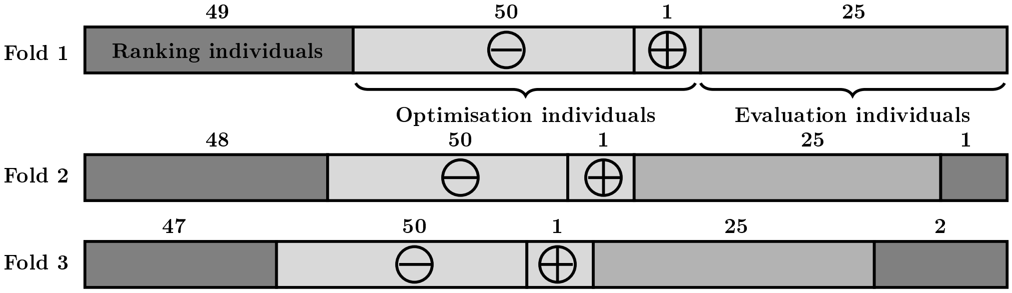


Figure 6.9: Conceptualisation of the first three (out of a total of 125) folds of the proposed data partitioning protocol for the IIT Delhi ear dataset within the context of Experiment 1B.

6.3.2 Experiment 2: Automated ROI detection

In this section an experimental protocol is proposed to gauge the proficiency of the proposed convolutional neural network (CNN)-based *automatic* ROI detection algorithm. The *manually* selected (specified) ROI serves as a ground truth for evaluating the proposed automated CNN-based ROI detection protocol. For each dataset, ear images from different individuals are used for *training*, *validation* and *testing* purposes. The output of the CNN-based algorithm is compared to the *manually* selected ROIs for the purpose of evaluating the proposed segmentation protocol. The pixels that are correctly classified as part of the earlobe are referred to as true positives, while those correctly classified as part of the background are referred to as true negatives. False positives constitute those pixels that are erroneously classified as part of the earlobe, while the pixels that are erroneously classified as part of the background are referred to as false negatives. For both of the datasets, 50% of the data is assigned to the *training* set, while 25% is assigned to the *validation* set and 25% to the *test* set (see Figure 6.10). A 4-fold cross validation experimental protocol is conducted on both of the AMI and IIT Delhi ear datasets.

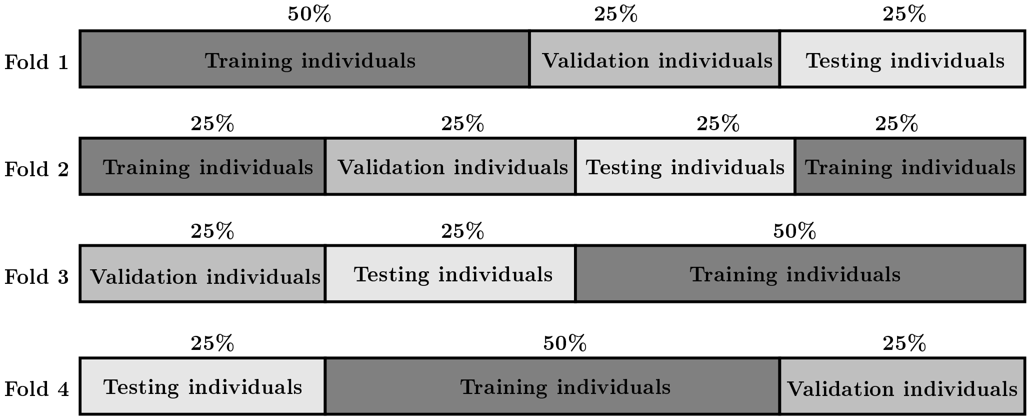


Figure 6.10: Conceptualisation of the proposed data partitioning protocol implemented for Experiment 2.

6.3.3 Experiment 3: Fully automated ear-based authentication

In this section an experimental protocol is proposed to evaluate the proficiency of the proposed *fully automated* ear-based biometric authentication system, where a suitable ROI is automatically detected using an appropriate CNN. For this experiment a questioned ear is accepted as authentic if and only if the reference ear has a ranking of *one*. Within the context of this experiment both of the ear datasets considered in this study are partitioned into four subsets, where each subset contains images associated with different individuals that is a *training* set, a *validation* set, a *ranking* set and an *evaluation* set.

6.4 Results

In this section a comprehensive analysis of the results for the conducted experiments relating to the *semi-automated* system in which the ROI is *manually* selected for each questioned ear, the *automated* ROI detection algorithm, and the *fully automated* end-to-end ear-based biometric authentication system are discussed. These results are categorised according to the aforementioned experimental protocols. The quantifiable assessment of the proficiency of the proposed systems are based on the following:

- The number of true positives (TP), that is the number of positive samples correctly accepted;
- The number of false positives (FP), that is the number of negative samples incorrectly accepted;
- The number of false negatives (FN), that is the number of positive samples incorrectly rejected;
- The number of true negatives (TN), that is the number of negative samples correctly rejected.

The relevant statistical performance measures employed in this thesis for the purpose of quantifying the proficiency of the proposed systems are listed and defined in Table 6.1

Performance measure	Definition
False acceptance rate (FAR)	FP/(FP+TN)
False rejection rate (FRR)	FN/(FN+TP)
Average error rate (AER)	(FAR+FRR)/2
Equal error rate (ERR)	FAR $\approx$ FRR
Precision (PRE)	TP/(TP+FP)
Recall (REC)	TP/(TP+FN)
Accuracy (ACC)	(TP+TN)/(TP+FN+FP+TN)
F_1 score	2 * PRE * REC/(PRE+REC)

Table 6.1: The statistical performance measures employed in this thesis. These performance measures are often expressed as percentages.

6.4.1 Semi-automated ear-based authentication

Rank-1 scenario: In this subsection the experimental protocol for Experiment 1A, as outlined in Section 6.3.1, is implemented and analysed. Recall that according to this protocol the ROI is manually specified and a questioned ear is only accepted when it has a ranking of one. The results are presented in Table 6.2.

Performance measure	AMI ear dataset (%)	IIT Delhi ear dataset (%)
FAR	0.04	0.18
FRR	4.76	13.00
AER	2.40	6.59
PRE	94.65	87.58
REC	95.24	87.00

Table 6.2: The results for the proposed semi-automated ear-based authentication system within the context of the rank-1 scenario for the AMI and IIT Delhi ear datasets. These results constitute averages across the relevant folds according to the protocol outlined for Experiment 1A.

From the results presented in Table 6.2 it is clear that the proposed systems are more proficient in the case of the AMI ear database, presumably due to the fact that these images have a higher resolution than those in the IIT Delhi ear database. It is furthermore evident that the rank-1 scenario, for which the FARs are relatively low and the FRRs are relatively high, renders the system very strict. This can be remedied by empirically determining an optimal (more lenient) ranking criterion as will be investigated and analysed in the next paragraph.

Optimal ranking scenario: In this subsection the experimental protocol for Experiment 1B, as outlined in Section 6.3.1, is implemented and analysed. Recall that according to this protocol the ROI is manually specified and a questioned ear is accepted when it has a ranking that is better than or equal to an optimal ranking. The AER and EER were investigated as optimisation criteria for selecting the optimal ranking, and its was found that (in general) the *same* optimal ranking is inferred irrespective of the criterion employed. The optimisation protocol for the AMI and IIT Delhi ear datasets is quantitatively illustrated in Figures 6.11 and 6.12 respectively.

In the aforementioned figures the average FAR and FRR are plotted as a function of the ranking across all folds (and sub-folds) by only considering the optimisation individuals.

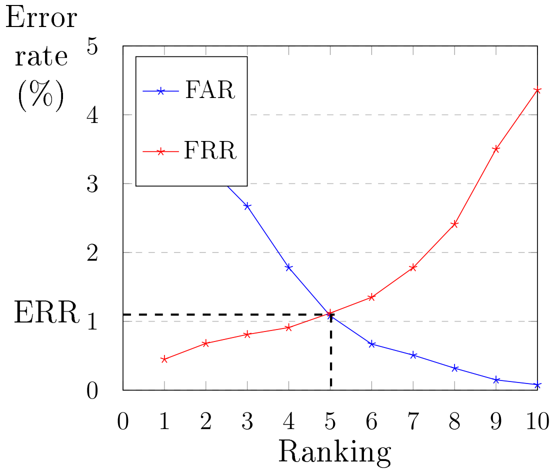


Figure 6.11: The average FAR and FRR as functions of the ranking across all folds (and sub-folds) by *only* considering the *optimisation* individuals within the context of the *AMI* ear dataset. The EER (and AER) correspond to an optimal ranking of 5. All of the ear images in the evaluation sets that has a ranking of 5 or better will therefore be accepted.

From Figures 6.11 and 6.12 it is therefore clear that based on the optimisation sets across all folds (on average) questioned ear images with rankings of 5 (or better) and 7 (or better) should be accepted within the context of the AMI and IIT Delhi ear datasets respectively in order to expect a low AER, as well as a similar FAR and FRR. When the aforementioned optimal rankings are imposed on the respective evaluation sets for the AMI and IIT Delhi ear datasets respectively the average performance metrics (across all folds) listed in Tables 6.3 and 6.4 are obtained. It should be noted that *different* optimal rankings may be imposed for *different* folds, but this was not investigated in this thesis due to time constraints. This being said, it is however important to note that the individual optimal rankings for different folds are very consistent.

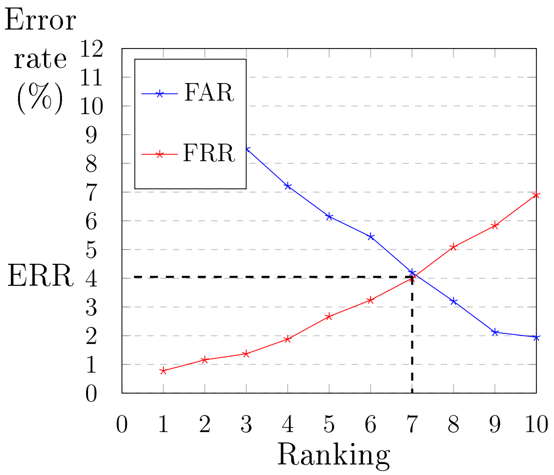


Figure 6.12: The average FAR and FRR as functions of the ranking across all folds (and sub-folds) by *only* considering the *optimisation* individuals within the context of the *IIT Delhi* ear dataset. The EER (and AER) correspond to an optimal ranking of 7. All of the ear images in the evaluation sets that has a ranking of 7 or better will therefore be accepted.

Performance measure	Rank-5 (%)
FAR	2.74
FRR	1.08
AER	1.91
PRE	70.24
REC	98.92

Table 6.3: The results for the proposed semi-automated ear-based authentication system within the context of the AMI ear database and an optimal ranking of 5. These performance evaluation measures constitute average percentages across all of the folds and *only* involve *evaluation* individuals. Only questioned images with a ranking of 5 or better are accepted.

When the results in Tables 6.3 and 6.4 are compared to those in Table 6.2 it is clear that the AER can be decreased by also allowing for ranking criteria other than rank-1.

Performance measure	Rank-7 (%)
FAR	6.94
FRR	4.20
AER	5.07
PRE	60.99
REC	94.06

Table 6.4: The results for proposed automated ear authentication system within the context of the IIT Delhi ear database and an optimal ranking of 7. These performance evaluation measures constitute average percentages across all of the folds and *only* involve *evaluation* individuals. Only questioned images with a ranking of 7 or better are accepted.

6.4.2 Automated ROI detection system

In this subsection the proficiency of the proposed automatic ROI detection algorithm is investigated and analysed. The proficiency of the proposed system is analysed by comparing the manually specified ROIs (which serve as a ground truth for the proposed system) and the automatically detected ROIs (the output of the proposed CNN). In Tables 6.5 and 6.6 the ROI detection results are summarised for the AMI and IIT Delhi ear datasets respectively. The precision, recall, accuracy and F_1 score are employed as performance evaluation measures.

From the results presented in Tables 6.5 and 6.6 it is clear that the proposed ROI detection protocol is more proficient in the case of the AMI ear database, again presumably due to the fact that these images have a higher resolution than those in the IIT Delhi ear database.

In order to visually compare the manually selected and automatically detected ROIs, a few examples within the context of the AMI and IIT Delhi ear databases are presented in Figures 6.13 and 6.14 respectively. The true positive, true negative, false positive and false negative pixels are depicted in white, black, green and pink respectively.

Performance measure	ROI detection (%)
PRE	80.30
REC	90.88
ACC	91.01
F_1	87.66

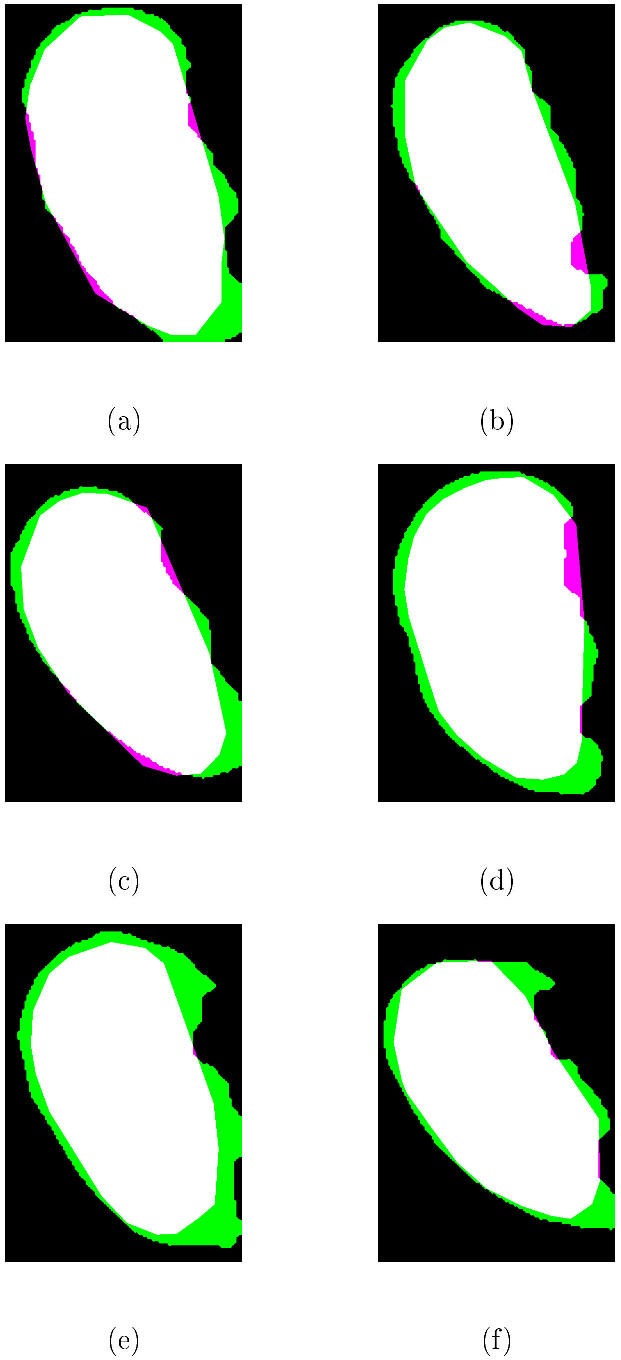
Table 6.5: The results for the proposed automatic ROI detection protocol within the context of the *AMI* ear dataset. The tabulated results constitute average percentages (across all folds) of the employed performance evaluation measures.

Performance measure	ROI detection (%)
PRE	70.26
REC	81.86
ACC	87.93
F_1	73.40

Table 6.6: The results for the proposed automatic ROI detection protocol within the context of the *IIT Delhi* ear dataset. The tabulated results constitute average percentages (across all folds) of the employed performance evaluation measures.

6.4.3 Fully automated ear-based authentication

In this subsection the results for the proposed fully automated ear-based authentication system are presented. This system employs a protocol in which the ROI is *automatically* detected through deep learning, followed by feature extraction, feature matching, and verification. Note that only the rank-1 scenario was investigated within this context. The results are presented in Table 6.7.



(a)

(b)

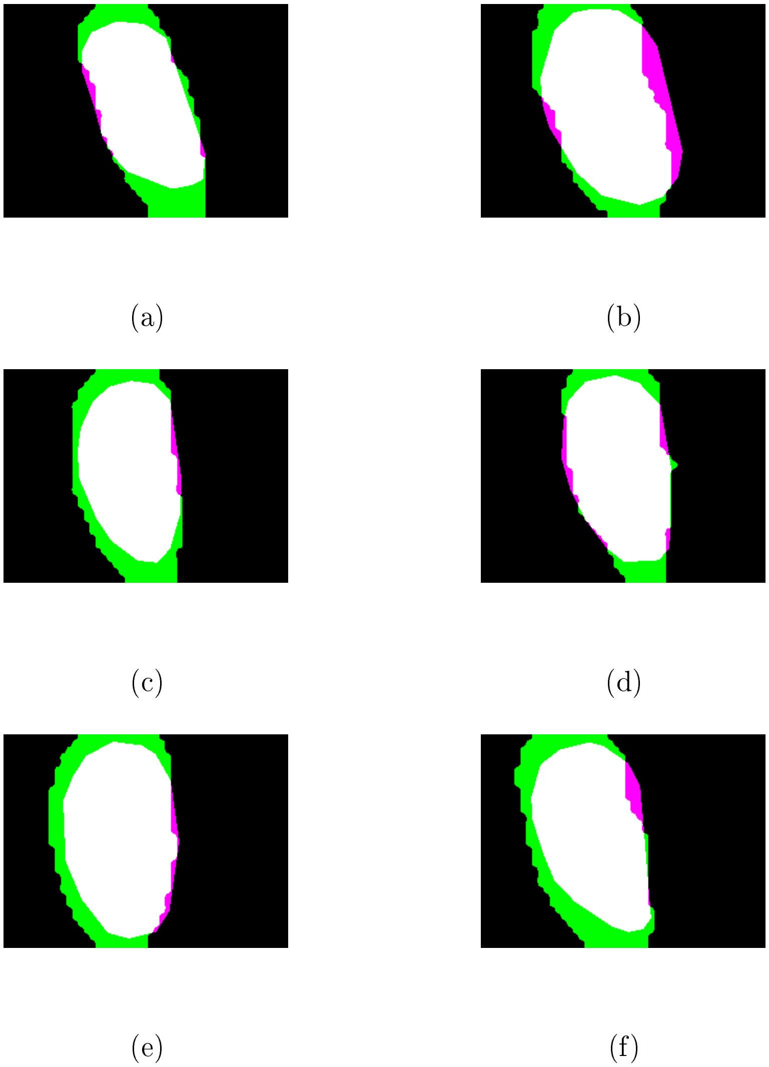
(c)

(d)

(e)

(f)

Figure 6.13: Examples of ear images from the *AMI* ear dataset for the purpose of comparing the manually selected (ground truth) and (CNN-based) automatically detected ROIs. The true positive, true negative, false positive and false negative pixels are depicted in white, black, green and pink respectively.



(a)

(b)

(c)

(d)

(e)

(f)

Figure 6.14: Examples of ear images from the *IIT Delhi* ear dataset for the purpose of comparing the manually selected (ground truth) and (CNN-based) automatically detected ROIs. The true positive, true negative, false positive and false negative pixels are depicted in white, black, green and pink respectively.

Performance measure	AMI ear dataset (%)	IIT Delhi ear dataset (%)
FAR	1.74	3.45
FRR	23.18	42.65
AER	12.46	23.05
PRE	76.52	56.77
REC	76.82	57.35

Table 6.7: The results for the proposed fully automated ear-based authentication system within the context of the rank-1 scenario for the AMI and IIT Delhi ear datasets. These results constitute average percentages (across all folds) of the employed performance evaluation measures.

Since the rank-1 scenario within the context of the proposed fully automated system is very strict and is therefore characterised by a low FAR and a high FRR, it is reasonable to expect similar improvements to those reported for the semi-automated system discussed in Section 6.3.1, when an optimal (more lenient) ranking is estimated and imposed within the current context. This was not investigated in this thesis due to time constraints.

6.5 Software and hardware employed

The systems proposed in this thesis were implemented in MATLAB$^{\text{TM}}$ (versions R2017b and R2018a). The following toolboxes were employed:

- Image Processing Toolbox$^{\text{TM}}$ (version R2017b);
- Neural Network Toolbox$^{\text{TM}}$ (version R2018a); and
- Statistics and Machine Learning Toolbox$^{\text{TM}}$ (version R2018a).

The algorithms were implemented on an 8*th* Generation Intel® Core$^{\text{TM}}$ i5 workstation with 8 GB RAM.

6.6 Discussion

It was demonstrated that in scenarios where the ROI is manually specified and a questioned ear is only accepted when it has a ranking of one, AERs of 2.4% and 6.59% is achievable within the context of the AMI and IIT Delhi ear databases. When all questioned ears with a ranking equal to or better than an estimated optimal ranking are accepted, the aforementioned AERs can be reduced to 1.91% and 5.07% respectively. The discrepancy in the proficiency for the two datasets in question may be attributed to the quality (resolution) of the images.

Accuracies of 91% and 88% are reported for the proposed CNN-based ROI detection protocol within the context of the AMI and IIT Delhi ear databases respectively.

As expected, the proficiency of the proposed fully automated end-to-end system, in which the ROI is automatically detected, followed by feature extraction, feature matching, and rank-1-based verification, is significantly lower than that of the corresponding rank-1-based semi-automated system, in which the ROI is manually specified. For the fully automated system, AERs of 12.8% and 23.05% are reported within the context of the AMI and IIT Delhi ear databases respectively. These results may be improved upon by also considering optimal rankings, which do not necessarily coincide with a ranking of one.

This research provided valuable insight into the problem of ear-based biometric authentication and opened up various avenues for further research. The above-mentioned topics will be discussed in more detail in the final chapter.

Chapter 7

Conclusion and future work

7.1 Conclusion

In this thesis novel ear-based authentication systems were proposed. Firstly, a segmentation protocol which facilitates the automatic detection of a region of interest (ROI) that encloses the entire ear shell was developed. The aforementioned protocol, which employs a convolutional neural network, is followed by morphological post-processing. The Canny edge detector was subsequently applied to find prominent contours associated with the ear shell. This was followed by the extraction of features from the aforementioned contours through the application of the discrete Radon transform (DRT). Appropriate feature normalisation techniques were applied to ensure translation, scale and rotation invariance across all feature sets. The difference between (dissimilarity of) two feature sets was quantified by the average Euclidean distance between the corresponding feature vectors. A ranking verifier was constructed for verification purposes by computing the dissimilarity between the questioned ear and a reference ear, as well as the respective dissimilarities between the questioned ear and those belonging to ranking individuals.

The proficiency of the proposed systems was estimated by considering two datasets that is (1) the Mathematical Analysis of Images (AMI) ear dataset and (2) the Indian Institute of Technology (IIT) Delhi ear dataset. Within the context of the proposed semi-automated ear-based authentication system in which the ROI is manually specified, the performance of the aforementioned system was demonstrated to be comparable to those of existing systems for a rank-1 scenario. It was furthermore demonstrated that the proficiency may be improved upon through the estimation of an optimal ranking by considering a separate set of so-called optimisation individuals. The proficiency of the proposed CNN-based ROI-detection protocol within the context of ear-based biometric authentication was also demonstrated to be comparable to those of existing systems.

7.2 Future work

Although the research conducted in this thesis provided valuable insight into numerous aspects relating to ear-based biometric authentication and deep learning, the following avenues have not been pursued due to time constraints and should therefore represent interesting future work:

(1) Within the context of the fully automated system developed in this thesis, only the ROI-detection protocol (that is the first part of the system) employs a deep learning-based approach. The remainder (second part) of the aforementioned system relies on the extraction of manually tailored features, template matching, and a ranking verifier. An investigation into the development of an end-to-end deep learning-based approach, or the utilisation of another machine learning-based approach, like a support vector machine, for the second part of the fully automated system developed in this thesis should be very interesting.

(2) The research conducted in this thesis was restricted to the authentication of ears that are only allowed to rotate within a plane parallel to the plane of the camera. An investigation into the feasibility of appropriate affine transformations, amongst other things, for the purpose of authenticating ears that are also allowed to rotate out of the aforementioned plane, should be of value. Ways for dealing with the inevitable resulting occlusions may also be investigated.

(3) Only two datasets were considered in this research. As discussed in Chapter 2, many other datasets may be publicly available and used for experimental purposes.

(4) A more in-depth investigation into the very specific problem cases that negatively impacted the reported proficiency of the systems developed in this thesis should be conducted.

(5) Within the context of the fully automated system developed in this thesis, only the rank-1 scenario was investigated. This (very strict) scenario resulted in a low false acceptance rate (FAR) and a high false rejection rate (FRR). Within the context of the semi-automated system developed in this thesis, it was demonstrated that the average error rate (AER) can be decreased significantly by considering an optimal ranking scenario based on either the equal error rate (EER) or the AER. This was not investigated for the fully automated system developed in this thesis and therefore constitutes viable future research.

The objectives of this research, as outlined in Section 1.3, have therefore been achieved.

Bibliography

Abaza, A., Hebert, C., & Harrison, M. A. F. (2010). Fast learning ear detection for real-time surveillance. In *Biometrics: Theory applications and systems (btas), 2010 fourth ieee international conference on* (pp. 1–6).

Abraham, A., & Sathya, R. (2013). Comparison of supervised and unsupervised learning algorithms for pattern classification. *International Journal of Advanced Research in Artificial Intelligence*, *2*(2), 34–38.

Annapurani, K., Sadiq, M., & Malathy, C. (2015). Fusion of shape of the ear and tragus–a unique feature extraction method for ear authentication system. *Expert Systems with Applications*, *42*(1), 649–656.

Beukes, E. (2018). *Hand vein-based biometric authentication with limited training samples.* MSc, Stellenbosch University.

Bishop, C. M. (2006). *Pattern recognition and machine learning.* Springer.

Burge, M., & Burger, W. (1996). Ear biometrics. In *Biometrics* (pp. 273–285). Springer.

Canny, J. (1986). A computational approach to edge detection. *IEEE Transactions on pattern analysis and machine intelligence*(6), 679–698.

Coetzer, J. (2005). *Off-line signature verification.* PhD, Stellenbosch University.

Coetzer, J., Herbst, B. M., & du Preez, J. A. (2004). Offline signature verification using the discrete radon transform and a hidden markov model. *EURASIP Journal on applied signal processing*, *2004*, 559–571.

Galdámez, P. L., Raveane, W., & Arrieta, A. G. (2017). A brief review of the ear recognition process using deep neural networks. *Journal of Applied Logic*, *24*, 62–70.

Glorot, X., & Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth interna-*

tional conference on artificial intelligence and statistics (pp. 249–256).

Gonzalez, E., Alvarez, L., & Morazza, L. (2012). *Ami ear database.* http://ctim.ulpgc.es/research_works/ami_ear_database/.

Gonzalez, R. C., & Woods, R. E. (2010). *Digital image processing.* Pearson-Prentice-Hall.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning* (Vol. 1). MIT press Cambridge.

Han, X., & Dai, Q. (2018). Batch-normalized mlpconv-wise supervised pre-training network in network. *Applied Intelligence*, *48*(1), 142–155.

Iannerelli, A. (1989). *Ear identification, forensic identification series.* California: Paramount Publishing Company, Fremont.

Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*.

Karpathy, A., Li, F., & Johnson, J. (2017). Cs231n: Convolutional neural networks for visual recognition. http://cs231n.github.io.

Kauderer-Abrams, E. (2017). Quantifying translation-invariance in convolutional neural networks. *arXiv preprint arXiv:1801.01450*.

Kotsiantis, S. B., Zaharakis, I., & Pintelas, P. (2007). Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering*, *160*, 3–24.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097–1105).

Kumar, A. (2007). Iit delhi ear database version 1.0. http://webold.iitd.ac.in/~biometrics/Database_Ear.htm.

Kumar, A., & Wu, C. (2012). Automated human identification using ear imaging. *Pattern Recognition*, *45*(3), 956–968.

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278–2324.

Lee, S. (2008). Convolutional neural networks (cnn). http://i-systems.github.io/HSE545/machine%20learning%20all/Workshop/180208_COSEIK/06_CNN.html.

Li, J., Cheng, J.-h., Shi, J.-y., & Huang, F. (2012). Brief introduction of back propagation (bp) neural network algorithm and its improvement. In *Advances in computer science and information engineering* (pp. 553–558). Springer.

Maas, A. L., Hannun, A. Y., & Ng, A. Y. (2013). Rectifier nonlinearities improve neural network acoustic models. In *Proc. icml* (Vol. 30, p. 3).

Moreno, B., Sanchez, A., & Vélez, J. F. (1999). On the use of outer ear images for personal identification in security applications. In *Security technology, 1999. proceedings. ieee 33rd annual 1999 international carnahan conference on* (pp. 469–476).

Neter, J., Kutner, M. H., Nachtsheim, C. J., & Wasserman, W. (1996). *Applied linear statistical models* (Vol. 4). Irwin Chicago.

Omara, I., Li, F., Hagag, A., Chaib, S., & Zuo, W. (2016). Ear recognition using a novel feature extraction approach. *International Journal of Computer Science Issues (IJCSI)*, *13*(6), 46.

Omara, I., Li, F., Zhang, H., & Zuo, W. (2016). A novel geometric feature extraction method for ear recognition. *Expert Systems with Applications*, *65*, 127–135.

Otsu, N. (1979). A threshold selection method from gray-level histograms. *IEEE transactions on systems, man, and cybernetics*, *9*(1), 62–66.

Rahman, M., Sadi, M. S., & Islam, M. R. (2014). Human ear recognition using geometric features. In *Electrical information and communication technology (eict), 2013 international conference on* (pp. 1–4).

Ren, X., & Malik, J. (2003). Learning a classification model for segmentation. In *Proceedings of the ninth ieee international conference on computer vision (iccv 2003)* (p. 10).

Shu-zhong, W. (2013). An improved normalization method for ear feature extraction. *Shandong College of Information Technology, Weifang, China*.

Simonyan, K., Vedaldi, A., & Zisserman, A. (2013). Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, *15*(1), 1929–1958.

Sutskever, I., Martens, J., Dahl, G., & Hinton, G. (2013). On the importance of initialization and momentum in deep learning. In *International conference on machine learning* (pp. 1139–1147).

Tharwat, A., Ibrahim, A., & Ali, H. A. (2012). Personal identification using ear images based on fast and accurate principal component analysis. In *Informatics and systems (infos), 2012 8th international conference* *on* (pp. MM–56).

Van Ooyen, A., Nienhuis, B., et al. (1992). Improving the convergence of the back-propagation algorithm. *Neural Networks*, *5* (3), 465–471.

Vélez, J. F., Sánchez, Á., Moreno, B., & Sural, S. (2013). Robust ear detection for biometric verification. *IADIS International Journal on Computer Science and Information Systems*, *8* (1), 31–46.

Yuan, L., & Mu, Z. (2014). Ear recognition based on gabor features and kfda. *The Scientific World Journal*, *2014*.

Zhang, Y., & Mu, Z. (2017). Ear detection under uncontrolled conditions with multiple scale faster region-based convolutional neural networks. *Symmetry*, *9* (4), 53.