TOWARDS LEAST-CONSTRAINED HUMAN IDENTIFICATION BY RECOGNIZING IRIS AND PERIOCULAR AT-A-DISTANCE

ZHAO ZIJING

PhD

The Hong Kong Polytechnic University

2018

THE HONG KONG POLYTECHNIC UNIVERSITY

DEPARTMENT OF COMPUTING

Towards Least-Constrained Human Identification by Recognizing Iris and Periocular At-a-distance

ZHAO ZIJING

A thesis submitted in partial fulfilment
of the requirements for the degree of
Doctor of Philosophy

March 2018

CERTIFICATE OF ORIGINALITY

I hereby declare that this thesis is my own work and that, to the best of my knowledge and belief, it reproduces no material previously published or written, nor material that has been accepted for the award of any other degree or diploma, except where due acknowledgement has been made in the text.

______________________ (Signed)

ZHAO ZIJING (Name of student)

ABSTRACT

Recognizing humans in least-constrained environments is one of the key research goals for academia and industry. At-a-distance eye region based human recognition using iris and periocular information has emerged as a promising approach for addressing this problem due to the high level of uniqueness and stability of eye regions under less-constrained environments. However, image samples acquired under less-constrained conditions usually suffer from degrading factors such as noise, occlusion and lower resolution. Therefore, advanced algorithms beyond traditional methods are required to fully exploit useful iris and periocular information from degraded images. This thesis focuses on developing effective and reliable algorithms for at-a-distance iris and periocular recognition under such conditions.

The first stage of this thesis investigates accurate iris segmentation under less constrained environments, which is a key prerequisite for the iris recognition process. The key challenge comes from undesired factors such as noise, occlusion and light source reflection in degraded eye images. We built a novel relative total variation model with l^1-norm regularization, referred to as RTV-L^1, to deal with the aforementioned obstacles. With this new model, noise and texture can be suppressed from the acquired eye images while structures are soundly preserved, which provides ideal conditions for preliminary segmentation. We then applied a series of robust post-processing to refine the segmentation contours. The proposed approach significantly outperforms other state-of-the-art iris segmentation methods, especially for degraded eye images acquired under less constrained environments.

Followed by the RTV-L^1 based iris segmentation framework, we developed a novel deep learning based approach for extracting spatially corresponding features from iris images for more accurate and reliable matching. This approach is based on fully convolutional network (FCN) which can retain critical locality of the deep iris features, and a newly designed extended triplet loss (ETL) function is able to accommodate non-iris occlusion and spatial translation during the learning process. The learned features are shown to offer superior matching accuracy and outstanding generalizability to different imaging environments, compared with traditional hand-crafted iris features as well as convolutional neural network (CNN) based deep features.

Another important contribution of this thesis is the development of deep learning based periocular recognition algorithms for improved accuracy and adaptiveness. Inspired by human inference mechanism, we firstly investigated combining high-level semantic information in the periocular images (*e.g.*, gender, left/right) into deep features learned by CNN. Supplement of such semantic information can help to recover more comprehensive and discriminative features and reduce the over-fitting problem, and superior performances over state-of-the-art periocular recognition methods were obtained. Furthermore, we proposed an attention based deep architecture for periocular recognition to further simulate the visual classification system of human. In this part, we inferred that regions of eye and eyebrow are of critical importance for identifying periocular and deserve more attention during visual feature extraction. We therefore incorporated such visual attention by emphasizing convolutional responses within detected eye and eyebrow regions in CNNs to enhance the feature discriminability. This approach further boosted state-of-the-art performance dramatically for periocular recognition under varying less constrained situations.

PUBLICATIONS

- Zijing Zhao and Ajay Kumar, "A Deep Learning based Unified Framework to Detect, Segment and Recognize Irises Using Spatially Corresponding Features", *Pattern Recognition*, Under Review, 2018.

- Zijing Zhao and Ajay Kumar, "Improving Periocular Recognition by Explicit Attention to Critical Regions in Deep Neural Network", *IEEE Transactions on Information Forensics and Security* (*T-IFS*), vol. 13, no. 12, pp. 2937-2952, 2018.

- Zijing Zhao and Ajay Kumar, "Towards More Accurate Iris Recognition Using Deeply Learned Spatially Corresponding Features", *Proceedings of the IEEE International Conference on Computer Vision* (*ICCV*) *2017*, **(Spotlight, Acceptance Rate = 4.70%)**, Venice, Italy, 2017.

- Zijing Zhao and Ajay Kumar, "Accurate Periocular Recognition under Less Constrained Environment Using Semantics-Assisted Convolutional Neural Network", *IEEE Transactions on Information Forensics and Security* (*T-IFS*), vol. 12, no.5, pp. 1017-1030, 2017.

- Zijing Zhao and Ajay Kumar, "An Accurate Iris Segmentation Framework under Relaxed Imaging Constraints using Total Variation Model", *Proceedings of the IEEE International Conference on Computer Vision* (*ICCV*) *2015*, Santiago, Chile, 2015.

ACKNOWLEDGEMENTS

This thesis could not have been completed without the kind assistance from many people around me during my Ph.D study. I would like to take this opportunity here to express my most sincere thankfulness to some of them.

First of all, I could not be more appreciative to my chief supervisor, Dr. Ajay Kumar, not only for his valuable supervision on my research as well as the kindest care on my daily life. I could always obtain most useful insights and feedbacks from him when I encounter difficulties in my own research work. His extensive knowledge, constructive suggestions and research passions have been continuously motivating me to pursue higher research achievements in the past, and would even be so in the future.

Next, I have to sincerely thank my colleagues, Chenhao Lin, Kelvin Chang and Kuo Wang, for keeping discussion with me and offering insightful suggestions on my work during my Ph.D study in the laboratory. Their questions and comments on my research problems have significantly helped me in improving the quality of my work. I also benefit much from understanding their research problems through discussion, which in turn offers me valuable inspiration on my own study.

Last, but not the least, I would like to convey my deepest thankfulness to my dear parents and sister, for supporting every decision of mine without any reason and backing up my life and study not only during the postgraduate period but all along my life. Their care and support are the motivation for my persistence in overcoming difficulties and striving for higher academic achievements in these years.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS

AUC	Area under the Curve
CASIA	Institute of Automation Chinese Academy of Sciences
CMC	Cumulative Match Characteristic
CNN	Convolutional Neural Network
DBN	Deep Belief Network
DCT	Discrete Cosine Transform
DFT	Discrete Fourier Transform
DSC	Distance-driven Sigmoid Cross-entropy Loss
EER	Equal Error Rate
ETL	Extend Triplet Loss
FAR/FMR	False Accept Rate / False Match Rate
FCN	Fully Convolutional Network
FNIR	False Negative Identification Rate
FOCS	Face and Ocular Challenge Series Database
FPIR	False Positive Identification Rate
FRGC	Face Recognition Grand Challenge Face Image Database
FRR/FNMR	False Rejection Rate / False Non-Match Rate
GAR/VR	Genuine Accept Rate / Verification Rate
GMM	Gaussian Mixture Model
IITD	Indian Institute of Technology Delhi
LFW	Labeled Faces in the Wild Database
NICE	Noisy Iris Challenge Evaluation
NIR	Near-Infrared
NIST	National Institute of Standards and Technology
ROC	Receiver Operating Characteristic
RTV	Relative Total Variation
SCNN	Semantics-Assisted Convolutional Neural Network
UBIRIS	University of Beira Interior Iris Image Database
WVU	West Virginia University
YTF	YouTube Face Database

CHAPTER 1
Introduction

1.1 Biometrics for Human Recognition

Human recognition or identification plays an important role in modern society for resident authentication, security control, forensics, searching for missing people, and so on. Biometric recognition, or biometrics, refers to techniques for identifying a human based on his/her biological or behavioural patterns, which can be measured, analysed and used to distinguish a person from another. A desired biometric modality for accurate and reliable identification should provide several properties, such as uniqueness, permanence, measurability, performance, *etc.* [1] [2]. Typical biological patterns that can be used as a biometric modality include but are not limited to face, iris, fingerprint and palmprint, while useful behavioural patterns can be attributed to voice, gait, signature, *etc.* Figure 1.1 illustrates several popular biometric modalities which are widely studied and used in both academia and industry.

Exploiting such biometric patterns for human identification can offer a number of benefits over traditional authentication mechanisms such as password. The usefulness of biometrics has led to wide development and deployment of automated systems for identifying persons in various scenarios such as resident registration, border crossing, surveillance and crime detection. The workflow of a typical biometric system is shown in Figure 1.2. As shown in the figure, biometric patterns are acquired from users/suspects via certain sensors, and then the acquired data will be processed with several steps which usually involve pre-processing, feature extraction/template generation, and matching with known subject(s) that were previously enrolled/registered in the gallery. The performance of a biometric system is of vital

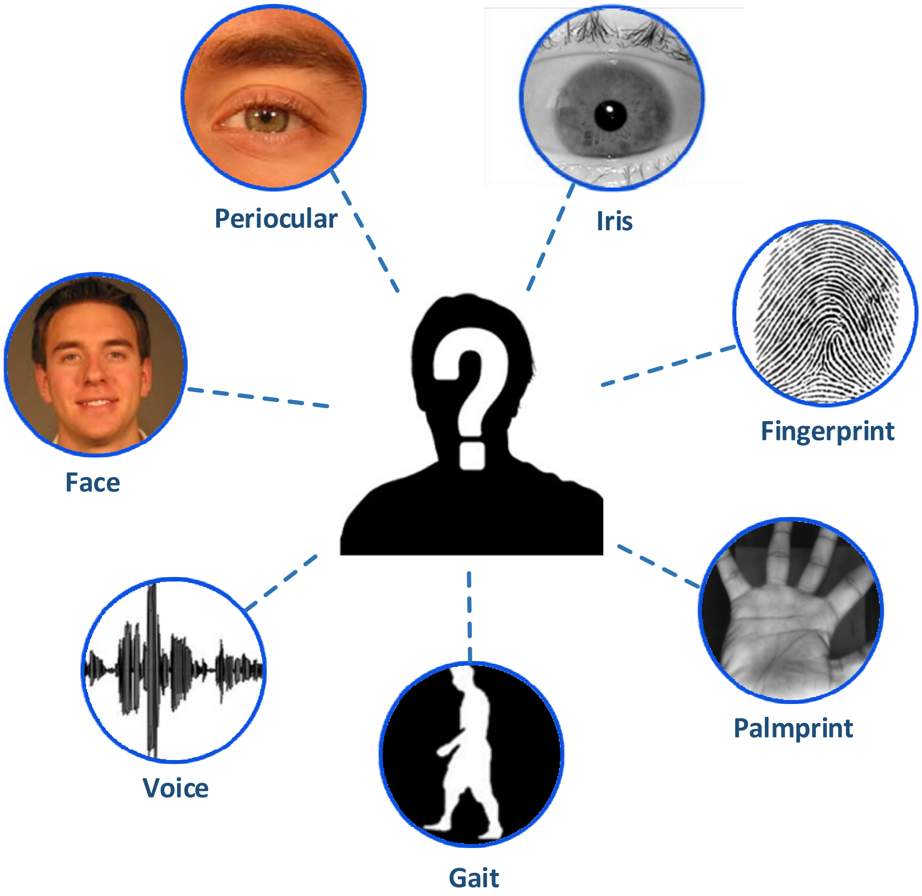


Figure 1.1: Some examples of widely studied biometric modalities that can be used to distinguish/identify people.

importance and will be affected by a variety of factors such as physical deployment environment, hardware configuration and matching algorithms. Although abundant approaches have been developed for accurately matching a wide range of biometric patterns, perfect solutions that can deal with every scenario do not exist and there is continuous demand for exploring more reliable methods on existing and new biometric modalities to deal with varying conditions.

1.2 Performance Evaluation for Biometric Systems

As discussed earlier, it is most unlikely for biometric recognition systems to identify

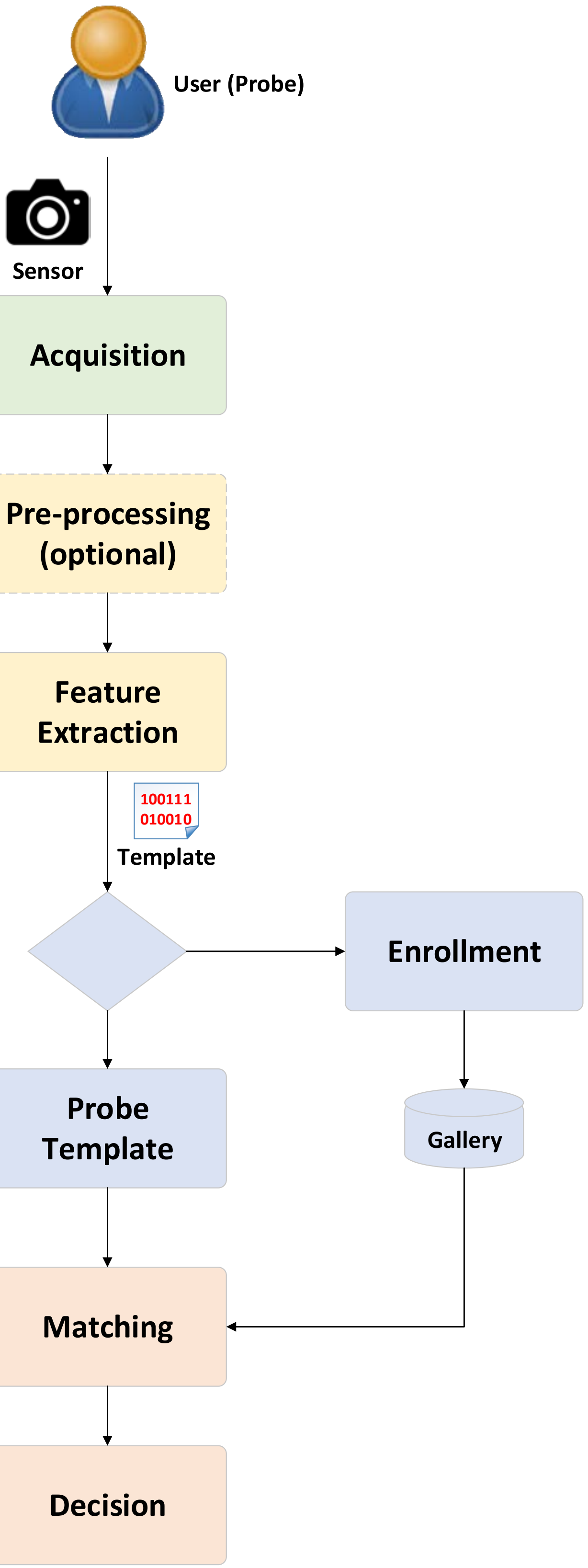


Figure 1.2: Workflow of a typical biometric recognition system.

every sample with perfect accuracy under various conditions, and it is required to quantitatively measure the performance for the biometrics systems in order to select most suitable approaches for the real applications. While there are many aspects to evaluate, within the scope of this thesis we will mainly focus on evaluating the accuracy of a specific biometric system or approach, *i.e.*, the correct rate of made decisions under certain experimental configurations. The accuracy is also considered as one of the most important factors for a biometric system. In some parts of our work which will be detailed in the following chapters, the time efficiency will also be evaluated in order to access the feasibility for real-time or online video stream-based deployment.

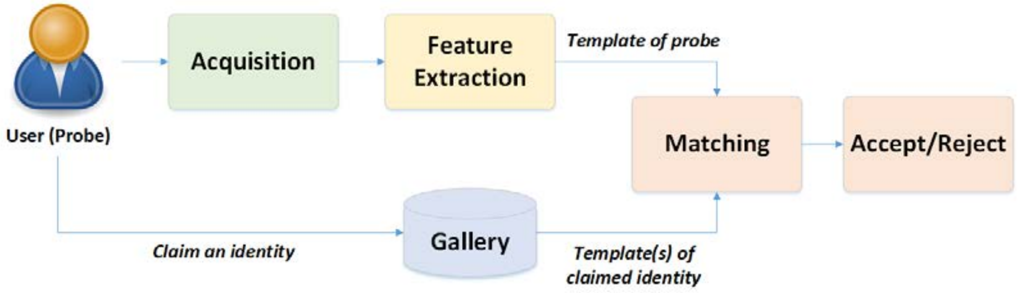
The accuracy of a biometric approach is related to the operation mode of the system, *i.e.*, verification and identification [1]. These two operation modes are most widely used and will be introduced in detail in the following, together with their corresponding metrics for measuring accuracy.

- *Verification mode (one-to-one)*

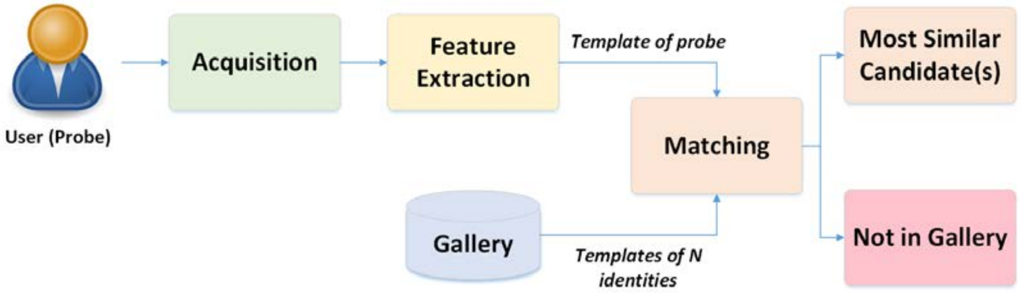
  In the verification mode, the user claims or is suspected to be a specific person who has been previously enrolled in the system, and the system is required to determine whether the present user is the same or a different person compared to the claimed one. Such operation mode is widely adopted for access control, resident administration, forensics, *etc.* Figure 1.3 (a) illustrates the workflow diagram of a verification system.

- *Identification Mode (one-to-many)*

  In this mode, the system is required to identify a presented person from a list of *N* known subjects, while the person does not need to claim an identity. This is often accomplished by iteratively matching (or sometimes distance-based searching) the template extracted from the presented subject with all previously extracted



(a) Verification system



(b) Identification system

Figure 1.3: Typical workflows for (a) a verification system and (b) an identification system.

templates from the list of subjects. The expected response from an identification system is whether the probe matches at least one gallery subjects, and if yes, returning the most similar candidates. A watchlist system is a typical application for this mode. Figure 1.3 (b) demonstrates the typical workflow of an identification system. The following explains commonly used metrics for evaluating biometric systems under these two modes.

a) Evaluating verification systems

When evaluating a verification system, we mainly consider two types of possible errors, which are: i) *False Accept / False Match*: Falsely accepting a person who is actually not the claimed identity; and ii) *False Reject / False Non-match*: Falsely rejecting a person who is indeed the claimed identity.

In order to measure these two types of errors, usually a certain number of pairs of biometrics samples will be matched to simulate matching the probe (presented user) with the stored template(s) for the claimed identity. The tested pairs will include some genuine pairs (samples from a same person) and some imposter pairs (samples from different persons), then the frequency of false decisions will be counted. More specifically, two kinds of corresponding error rates that are used to estimate the probabilities of errors are computed:

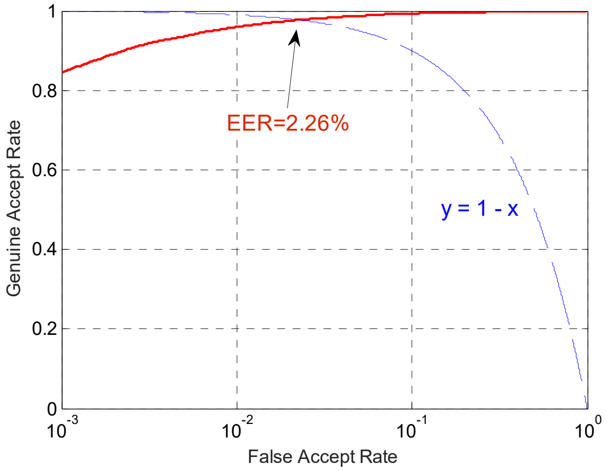
(i) *False Accept Rate (FAR) or False Match Rate (FMR)*:

The number of falsely accepted or approved imposter pairs over the total number of tested imposter pairs.

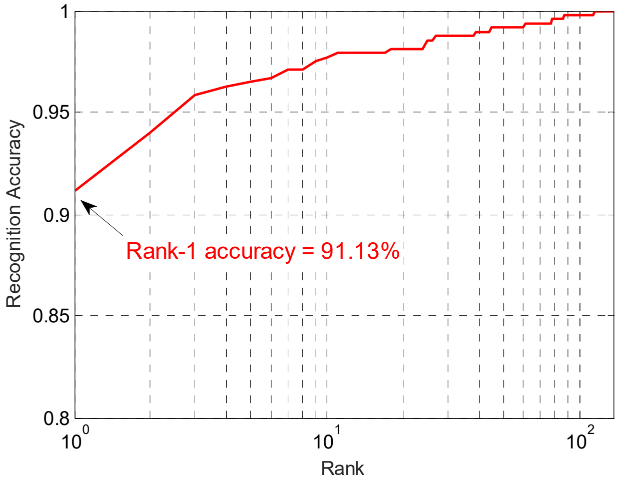
(ii) *False Reject Rate (FRR) or False Non-match Rate (FNMR)*:

The number of falsely rejected genuine pairs over the total number of tested genuine pairs.

Usually a trial of matching will generate a similarity (or dissimilarity) score for quantitatively describing how similar two samples or templates are. Given a score threshold *t*, if the score is larger than (sometimes smaller than, depending on the algorithm) *t*, the pair will be accepted and otherwise rejected. Apparently both the FAR and FRR are subjected to the given threshold *t*. When *t* varies within the possible range, FAR and FRR will also vary between zero and one accordingly, *i.e.*, one exact value of *t* corresponds to a pair of exact values of FAR and FRR. Therefore, all possible (FAR, FRR) can form a curve on the 2-D space $\{(x,y) \mid 0 \le x \le 1, 0 \le y \le 1\}$. Such a curve is referred to as *Receiver Operating Characteristic* (*ROC*) curve, which is actually a general performance metric for any binary classifiers. Sometimes the Genuine Accept Rate (GAR) or Verification Rate (VR), where GAR = 1 – FRR, is preferred when plotting ROCs, and in the following of this thesis we will adopt this style. Figure 1.4 (a) demonstrates a sample ROC for a specific biometric system/approach. Usually one biometric system or algorithm is considered superior than another one when its ROC is on top of that of the other under the same evaluation



(a) Example of ROC



(a) Example of CMC

Figure 1.4: Examples of (a) ROC curve, where the EER can be revealed by intersecting the curve with the line y = 1-x, and (b) CMC curve.

configuration. It can be inferred that the left hand side of ROC curve (lower FAR) represents the performance when the false accepts seldom happen, and therefore will be weighted more for applications that require higher level of security where FAR should be kept extremely low. From time to time, a specific value of error rate, *equal error rate* (*EER*), which satisfies FAR = FRR or FAR = 1 - GAR, will be considered apart from the complete ROC for roughly describing the overall accuracy of the system. However, whenever applicable, the complete ROC should be preferred since EER can be revealed by intersecting the ROC curve with the line $\{(x,y)\,|\,x+y=1\}$.

b) Evaluating identification systems

When evaluating an identification system, the primary factor to consider is whether the correct identity for the probe will appear in the most similar candidates after performing a search among the list of enrolled gallery subjects. This can be quantitatively measured with the rank-*k* accuracy (or top-*k* accuracy in some classification problems). Assume that a certain number of probes are iteratively input to an identification, while each probe will be matched with the templates from *K* known subjects that were previously enrolled in the system. The rank-*k* accuracy refers to the percentage of successful matches that the correct identity is given within the *k* most similar candidates for the probe. The discrete values of rank-*k* accuracies (*k* = 1, 2, …, *K*) can also form a 2-D curve which is called *cumulative match characteristic* (*CMC*) curve. Figure 1.4 (b) demonstrates a sample CMC curve from a specific identification system/algorithm under certain experimental setup. Apparently higher rank-k accuracies are preferred for an identification system, and the accuracies for smaller *k* values are usually weighted more than those for larger *k*.

More detailed explanation for the above metrics are available in a variety of references in the literature, such as [1] – [4]. In addition, several other performance metrics, such as false positive identified rate (FPIR) and false negative identification rate (FNIR) are adopted in some studies [4] [6] for more comprehensively evaluating biometric systems. Within the scope of this thesis, however, we will mainly use the metrics explained above for performance evaluation.

1.3 Towards Least-Constrained Recognition

By reviewing the research/engineering progress for biometric technologies over the past years, it can be concluded that the problem of accurate automated human recognition under relatively constant and controlled conditions has been largely addressed. For instance, advanced iris identification algorithms can reach rank-1 accuracy of over 95% for approximately four millions of identities [6], which is pretty satisfactory for such a large-scale evaluation. However, traditional approaches often require the users to be highly cooperative to provide good-quality biometric samples. The requirements of strict constraints and cooperative sampling from the users have greatly limited the usage and deployment of biometric systems. In some passive recognition scenarios, *e.g.,* surveillance and crime detection, cooperation from users/suspects cannot be expected and the systems should be able to identify subjects under less constrained environments. For some active recognition applications, such as border crossing, a more relaxing acquisition process is also desired to make the system more efficient and user-friendly. Due to such reasons, increasing research efforts have been devoted into more reliable person recognition under less constrained environments, and encouraging achievements have been made so far. To name a few, the Labeled Faces in the Wild (LFW) dataset [94] and YouTube Face (YTF) dataset [95], which were formed by collecting face images from Internet resource under unspecified/unconstrained conditions, have attracted a lot of research interest, and state-of-the-art approaches [57] [60] [62] have gained significant success in achieving high recognition accuracy on such datasets using deep learning models. Contactless fingerprint/palmprint recognition approaches [63] [64] have been proposed for relieving the constraints for the users and have obtained promising results. At-a-distance iris and periocular recognition also offers practical and effective solutions for less constrained person identification, and numerous methods have been developed to pursue continuous and solid progresses [10][47]. In this thesis, we will focus on less constrained iris and periocular recognition algorithms for more accurate and robust

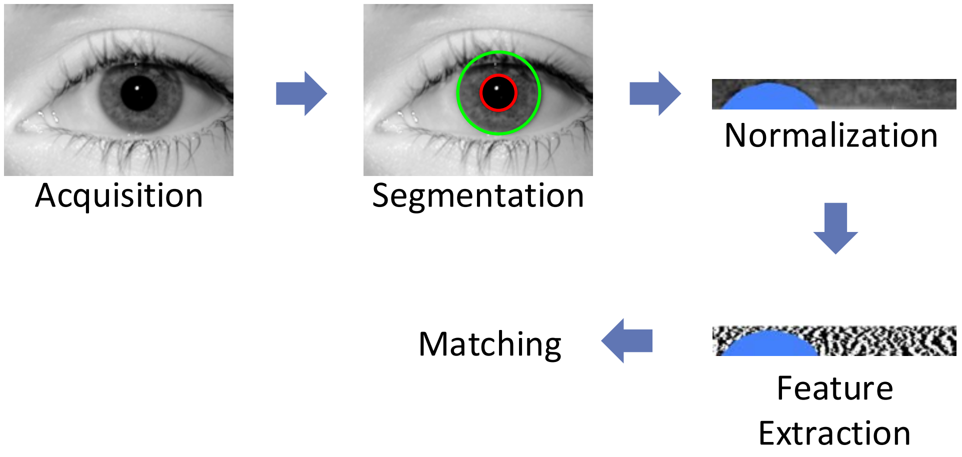


Figure 1.5: Typical workflow for contemporary iris recognition systems.

human identification. Kindly note that the term “less-constrained” is semantically a broad concept, and we mainly investigate factors of at-a-distance, lower resolution, off-angle/axis imaging and cross-database training/testing, which can be reflected by the selection of databases and experimental protocols used in this thesis.

1.4 Earlier Work

In this section we will have a detailed review on previous studies on iris recognition and periocular recognition, especially with the trend of applications from highly constrained recognition scenarios to at-a-distance and less constrained environments.

1.4.1 Iris Segmentation

Contemporary iris recognition approaches usually follow a workflow similar to that in Figure 1.5. Among the sequential procedures, iris segmentation refers to the step that identifies iris location and iris region pixels from the acquired eye images. Iris segmentation is a critical step at the beginning of the workflow and plays an important role for the final recognition accuracy [51]. Inaccurately segmented iris images are highly likely to degrade the matching performance severely. Therefore, it is necessary to ensure the robustness of iris segmentation, especially under less constrained environments.

Most of the earlier work on iris segmentation uses NIR images which are acquired from close distances. Duagman's integro-differential operator [7] is the classical algorithm for iris segmentation under NIR illumination and is adopted in most of the commercial systems nowadays. It searches for a maximum response of an integro-differential expression and then locates the circle of iris. However, as explored and addressed in [13]-[17], *etc*., under VW illumination or less-constrained environment, quality of images drops and such traditional approach performs poorly.

The iris segmentation approach developed by Tan *et al.* [12] first adopts an iterative technique to cluster the iris and non-iris region coarsely, and then uses an improved integro-differential operator to locate the iris and pupil circle coarsely. One key limitation of this algorithm is that it relies highly on the coarse clustering result so that the final accuracy will be heavily affected if the first step is not accurate. Another promising approach by Proença [11] proposes to exploit local color features and classify iris pixels using a neural network. However, the color features are not very stable, which often leads to lower reliability. A recent work detailed in [10] also offers highly competitive alternative for the iris segmentation under less constrained imaging environment. This approach first adopts a Random Walker [24] to coarsely segment the iris region to locate the iris circle, then applies a set of gray level statistics based operations to refine the boundary. This method reports a better accuracy than previous ones. However, this approach also relies on the coarse segmentation result too much, and in its post-processing operations, one common threshold value is used for the whole iris, which may not fit local features and is possible to cause global error.

Another promising work in relevant domain has been proposed by Li and Savvides [9]. In this method, a Gaussian Mixture Model (GMM) was adopted to simulate iris pixel distribution and an unsupervised training method was used to obtain the parameters for the GMM. It has shown very high segmentation accuracy and reliability. However, a critical step for iris segmentation, which is the localization of iris and pupil circles, was performed manually in the experiments presented in this work, while other methods mentioned above locate the circles automatically. In other words, the performance of [9] will highly depend on the accuracy of iris and pupil circle localization. In practice, iris and pupil circle localization is not only used in iris segmentation, but also necessary for the iris normalization, which unwraps the iris region into a polar coordinate system and is an essential step for most of the iris recognition algorithms.

1.4.2 Iris Recognition

One of the most classic and effective approaches for automated iris recognition was proposed by Daugman [7] in 2002. In his work, Gabor filter is applied on the segmented and normalized iris image, and the responses are then binarized as *IrisCode*. The Hamming distance between two *IrisCodes* is used as the dissimilarity score for verification. Based on [7], 1D log-Gabor filter was proposed in [8] to replace 2D Gabor filter for more efficient iris feature extraction. A different approach, developed in [14] in 2007, has exploited discrete cosine transforms (DCT) for analyzing frequency information of image blocks and generating binary iris features. Another frequency information based approach was proposed in [15] in 2008, in which 2D discrete Fourier transforms (DFT) was employed. In 2009, the multi-lobe differential filter (MLDF), which is a specific kind of ordinal filters, was proposed in [16] as an alternative to the Gabor/log-Gabor filters for generating iris templates.

Unlike the popularity of deep learning for various computer vision tasks, especially for face recognition, the literature so far has not yet fully exploited its potential for iris recognition. There has been very little attention on exploring iris recognition using deep learning. A deep representation for iris was proposed in [37] in 2015, but the purpose was for spoofing detection instead of iris recognition. A recent approach named DeepIrisNet in [38] has investigated deep learning based frameworks for general iris recognition. This work is essentially a direct application of typical convolutional neural networks (CNN) without much optimization for iris pattern. Our reproducible experimental comparison in Chapter 3.5 further indicates that under fair comparison, this approach [38] cannot deliver superior performance even over other popular methods. Another recent work [77] has attempted to employ deep belief net (DBN) for iris recognition. Its core component, however, is the optimal Gabor filter selection, while the DBN is again a simple application on the *IrisCode* without iris-specific optimization. Above studies have made preliminary exploration but failed to establish substantial connections between iris recognition and deep learning.

1.4.3 Periocular Recognition

Continuous research efforts have been devoted into investigating periocular recognition algorithms under different environments [39] [40]. The early feasibility study on using periocular region for human identification was performed by Park *et al.* [41] in 2009, and promising results have been reported, which provides support to subsequent research. Miller *et al.* [127] investigated personal identification using periocular skin features, followed by studies on utility of the periocular region appearance cues [128] and for soft biometrics [129] from the same group of researchers. Bharadwaj *et al.* [42] further ascertained the usefulness of periocular recognition, especially when iris recognition fails. Some of the later research focuses on cross-spectrum periocular matching [46] using techniques of neural network. Above explorative works have motivated further research efforts to continuously improve the accuracy of periocular recognition. One of the state-of-the-art approaches is proposed by [10] in 2013, which exploited DSIFT features of periocular images, followed by K-means clustering for dictionary learning and representation. This work also explored score level fusion of iris and periocular recognition and reported encouraging results. However, this approach did not investigate periocular-specific feature representation, and the employed DSIFT feature is computationally expensive. Smereka *et al.* [47] has proposed the Periocular Probabilistic Deformation Model (PPDM) in 2015, which provided a sound modelling for potential deformation existing between periocular images. Inference of the captured deformation using correlation filter is utilized for matching periocular pairs. Later in 2016, the same group of researchers improved their basic model by selecting discriminative patch regions for more accurate matching [49]. These two methods achieved promising performance on multiple datasets. Nevertheless, both of them rely on patch-based matching scheme, and therefore are less resistant to scale variation or misalignment that often violate the patch correspondence but is more likely to happen during the real deployments. More recently, Proença and Neves [123] claimed that iris and scalar regions may be less reliable for periocular recognition and proposed Deep-PRWIS, which weakened the energy of learning within these areas for CNN, and reported good results on two datasets.

1.5 Organization of Thesis

As introduced earlier, in this thesis we will focus on at-a-distance iris and periocular recognition for more accurate and robust person identification under less constrained environments. The rest of the thesis will detail my research work on developing novel approaches for improving state-of-the-art performance for at-a-distance iris and periocular recognition.

Chapter 2 will firstly introduce my work on accurate iris segmentation framework under less constrained environments, including the formulation of newly proposed RTV-L^1, improved circle detection and robust post-processing operations. Chapter 3 will mainly disclose my proposed deep learning based iris feature descriptor, which is based on a fully convolutional network (FCN) and a problem-specific extended triplet loss (ETL) function. Chapter 4 will present a novel periocular recognition approach based on semantics-assisted convolutional neural network (SCNN), which utilizes explicit semantic attributes of the training data for more comprehensive periocular feature learning. This is followed by Chapter 5, where another new approach incorporating visual attention mechanism into deep neural network for more effective and robust periocular feature extraction. Finally, Chapter 6 will draw the conclusions for my research work presented in this thesis, as well as discussions on the current limitations and future work.

CHAPTER 2

Iris Segmentation under Less Constrained Environment

2.1 Background

Iris recognition is one of the most accurate and widely employed approaches for automated personal identification. The performance of iris recognition algorithms is highly dependent on the effectiveness of segmenting iris region pixels [17]. However, the traditional iris segmentation and feature matching approaches adopt only to near-infrared illumination and require the subjects to be sampled under strictly constrained conditions [13], which is the major difficulty for deploying iris recognition system in civilian and surveillance applications on a larger scale. Automated iris segmentation has been a topic of considerable research in recent past [21]-[33] and many methods [9]-[13] have been proposed to address the problem. However the accuracy of currently available iris segmentation algorithms is still below the expectations and requires further improvement for the deployments.

This work proposes a new framework to automatically and accurately segment iris images from the distantly acquired face images. The developed approach can robustly operate using face or eye images acquired under less-constrained environments, *i.e.*, using images acquired from a distance (typically 3-8m) and under near-infrared (NIR) or visible-wavelength (VW) illumination. The key contributions from this work can be summarized as follows:

With the help of earlier studies on gradient dependent regularizer, such as relative total variation regularizer [20], we develop a new total variation formulation for iris segmentation in which the eye structure and surrounding texture are differently penalized. This formulation incorporates with an l^1 *norm* which is more effective and also computationally efficient. Our experimental results on three publicly available databases achieve significantly superior results over previous approaches presented in the most recent literature [9]-[10]. Moreover, the method developed in this work does not require any training and therefore is more attractive for the deployment in surveillance applications.

We develop a series robust post-processing operations to accurately localize limbic boundaries in noisy iris images. The adaptive and self-correcting methodology introduced in these operations can independently exploit the local features as much as possible, and helps to significantly reduce global errors. The post-processing operations can effectively use the intermediate results and adopt dynamic threshold mechanisms. Such robust strategies help to improve the overall accuracy in the segmentation of noisy iris images and can also be applied in other challenging problems in surveillances and remote sensing.

The performance of the proposed approach[1] has been evaluated on three publicly available databases, *i.e.*, UBIRIS.v2 [18], FRGC [28] using visible-light imaging and CASIA.v4-distance [27] under near infrared. The experimental results suggest average improvements of 28.82%, 30.98% and 16.05% on iris segmentation accuracy over state-of-the-art method on respective databases. Besides, we also illustrate from the experiments that using iris masks generated from our approach helps improve iris recognition performance.

The approach described in this chapter has been published as [51].

[1] The implementation codes for our algorithm are available via [34].

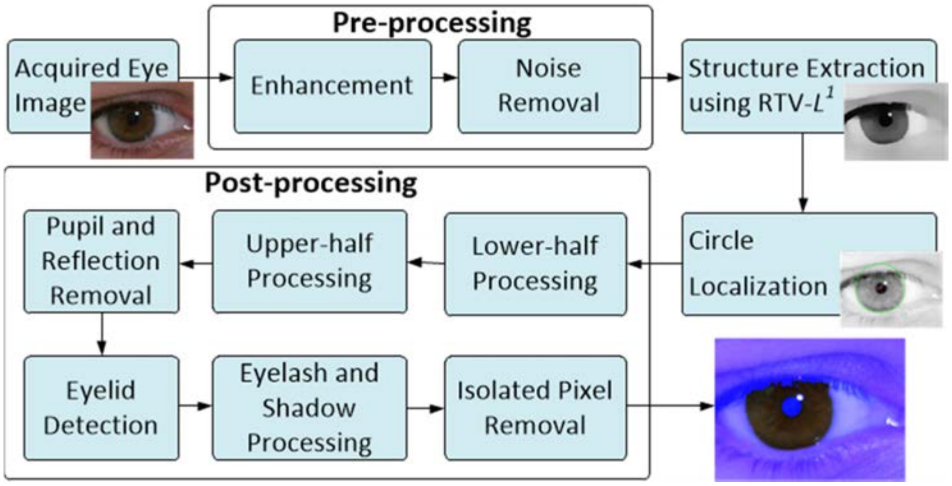


Figure 2.1: The block diagram for the proposed iris segmentation.

2.2 Proposed Methodology

This section details the methodologies used in the proposed iris segmentation approach. The overall framework of the developed approach is illustrated in Figure 2.1. The proposed approach adopts a coarse-to-fine strategy to segment iris region pixels from the background (region pixels surrounding the iris) and foreground (noisy pixels in the iris region) pixels in the acquired eye images. Our approach assumes that each of the eye images may be acquired under a relaxed imaging environment, *i.e.*, at-a-distance and under variable spectrum bands.

2.2.1 Preprocessing

Under less-constrained imaging, several factors such as varying illumination intensity and the angle of the illumination source can have adverse impact on the accuracy and quality of iris segmentation. Such unexpected changes yield severe challenges in not only the iris biometrics but also many other image understanding tasks. We use the Single Scale Retinex (SSR) approach [23] for normalizing eye image illumination. The SSR enhancement method is able to improve color consistency under severe

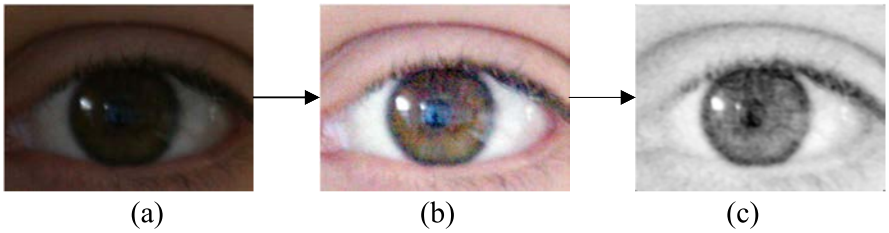


Figure 2.2: Sample image from the pre-processing stage: (a) original image, (b) enhanced image, (c) smoothed red channel.

illumination variance. A sample image after applying SSR enhancement is shown in Figure 2.2(b). After enhancement, we apply a median filter on the image to suppress isolated noisy pixels. Moreover, we only use the red channel in the following process because the imaging spectrum of red channel is closest to NIR, which retains better image quality. In Figure 2.2 (c) we can see a sample result from the pre-processing stage.

2.2.2 Total Variation-Based Iris Structure Extraction

One common characteristic for the eye images acquired under less-constrained environments is the sensitivity to noisy and complex details such as reflection and eyelashes, which are not needed in the initial structure analysis. The above factors are the major reason why the traditionally effective integro-differential operator or circular Hough transform perform poorly on images acquired under less-constrained environments, because both methods require clear contrast of structure components and least interference from noise. We exploit the total variation (TV) model to address such a problem. There have also been studies on using the total variation model for other biometric segmentation problems such as fingerprint segmentation [29].

A. Theoretical Foundation of Total Variation Model

There are several total variation (TV) regularizers for image structure separation in the literature, of which most are extended from TV-L^2 [25]. A recent reference in [20] proposed relative total variation (RTV) to measure and regularize local pixel variation. Such local gradient descriptors offer the strong capability to distinguish key image structure from the background image details. Motivated by such prior studies, we propose to use an improved RTV model to first localize the key eye structure, *i.e.*, eyelid, pupil and sclera boundaries, in the noisy eye images. Such localization of eye structure can be used to accurately locate pupillary and limbic boundaries for accurate iris segmentation. In the following, we provide a brief review on the theoretical principles of RTV which are later used to develop an improved RTV model incorporated with l^1 norm regularization to more effectively locate eye structure of key interest.

The windowed total variation of an image S within a local rectangle region R is expressed as follows:

$$\begin{aligned} D_{S,x} &= G_\sigma * | \partial_x S | \\ D_{S,y} &= G_\sigma * | \partial_y S | \end{aligned} \tag{2.1}$$

where G_σ is a Gaussian kernel with standard deviation σ, ∂_x and ∂_y are the partial derivatives on image S in two directions and $*$ represents the convolution operation. By the convolution, which gives a weighting sum of nearby absolute gradients, we can observe that $D_{S,x}$ and $D_{S,y}$ represent absolute spatial difference within a rectangular window. In earlier studies in [20], both the detail and structure patches in an image with salient textures yield large D, which indicates that the windowed total variation is responsive to visual saliency.

Another effective measure to help distinguishing prominent structures from the texture elements is to use windowed inherent variation, expressed as:

$$\begin{aligned} L_{S,x} &=| G_\sigma * \partial_x S | \\ L_{S,y} &=| G_\sigma * \partial_y S | \end{aligned} \tag{2.2}$$

Different from D, L measures overall spatial variation because $\partial_x S$ and $\partial_y S$ may be positive or negative, and therefore such values may eliminate or offset others by the convolution in frequently varying gradient region. As a result, structure patches are typically expected to yield larger *L* than those from texture patches.

The contrast between texture and structure can be further enhanced by combining *D* and *L* as *RTV*, expressed as follows:

$$RTV_{S,p} = \frac{D_{S,x}(p)}{L_{S,x}(p)+\varepsilon} + \frac{D_{S,y}(p)}{L_{S,y}(p)+\varepsilon} \tag{2.3}$$

where p is the pixel index, ε is a small positive number to avoid division by zero. From expression (2.3) we can observe that texture region is typically expected to yield larger RTV than structure since the denominator of the formulation, L , responses smaller value for texture. Making use of such a property of *RTV*, reference [20] proposed to minimize following energy to remove the texture (*e.g*., details and noise) from the input image:

$$\arg\min_{S} \sum_{p} \lambda \cdot RTV_{S,p} + \left(S_p - I_p\right)^2 \tag{2.4}$$

where *I* is the input image and *S* is the output image. Notice that equation (2.4) incorporates the square of an l^2 norm to enforce the similarity between the input and output image, which is similar to many other variants of TV regularization. We will refer to such a method as RTV-L^2 for short.

B. Extracting Eye Structure Using RTV-L^1

Each of the iris images acquired for conventional iris recognition includes surrounding eye structure. This structure essentially includes curved regions representing eyelid, pupil and sclera boundaries. Our objective is to locate the iris by automatically extracting such elements representing eye structure and other non-structural elements such as eyelash, and iris texture can be treated as noise because they could interference

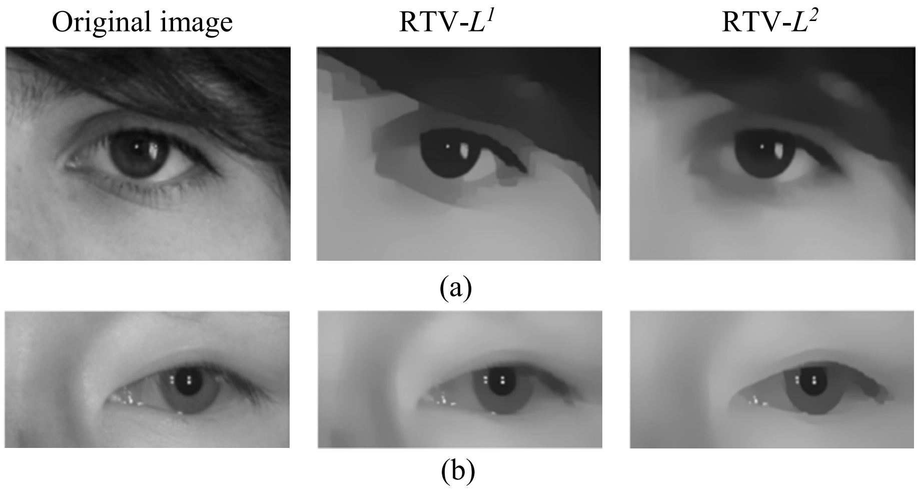


Figure 2.3: Sample results of RTV-L1 and RTV-L2 for eye images under (a) visible illumination and (b) NIR illumination.

with on our iris localization. Therefore, the RTV-L^2 approach which can remove details and texture while maintaining the main structure of the input image is a good choice for our purpose. However, it has been studied in several references [31], [32] that using l^1 norm instead of l^2 in such energy regularizers has better performance in some applications and presents more important geometric properties. We have studied the difference between l^1 and l^2 norm in RTV regularization, and propose to adopt l^1 norm instead of the original l^2 norm, i.e., we solve the following problem which we refer to as RTV-L^1:

$$\arg\min_{S} \sum_{p} \lambda \cdot RTV_{S,p} + \left| S_p - I_p \right| \tag{2.5}$$

The difference between the output images by solving problems (2.4) and (2.5) is illustrated in Figure 2.3. We can observe from Figure 2.3 that while both RTV-L^1 and RTV-L^2 can suppress texture and noise, the results from RTV-L^1 are sharper at critical edges than those from RTV-L^2. This confirms the arguments that using l^1 norm in the energy regularizer can present more important geometric properties, which is considered helpful for the subsequent iris localization process. The detailed numerical solution for problem (2.5) will be introduced in following sections.

C. Numerical Solution for RTV-L^1

The objective function in problem (2.5) is non-convex. A trivial solution for this problem is not available. In addition, by replacing the l^2 norm with l^1 norm, the structure of the objective function has changed so that the approximating solution proposed in [20] becomes unusable. Here we propose an effective dual formulation based solution similar to [30] for the RTV-L^1 problem. First, we approximate the minimization for problem (2.5) as minimizing the following new problem:

$$\arg\min_{S,V}\sum_{p}\lambda\cdot RTV_{S,p}+\frac{1}{2\theta}\left(S_p+V_p-I_p\right)^2+\left|V_p\right| \tag{2.6}$$

where V is a new variable in matrix form and the positive parameter θ is small, thus we have $V\approx I-S$. As a result, S presents the structural information and V captures the texture information from the input image. The minimization for problem (2.6) is performed with respect to S and V separately and iteratively. Thus, it boils down to the following two sub-problems:

(i) S being fixed, search for V for the problem:

$$\arg\min_{V}\sum_{p}\frac{1}{2\theta}\left(S_p+V_p-I_p\right)^2+\left|V_p\right| \tag{2.7}$$

(ii) V being fixed, search for S for the problem:

$$\arg\min_{S}\sum_{p}\lambda\cdot RTV_{S,p}+\frac{1}{2\theta}\left(S_p+V_p-I_p\right)^2 \tag{2.8}$$

Problem (2.7) and (2.8) are solved alternately and iteratively, and then the energy function in problem (2.6) keeps reducing until it converges to a satisfying level. Following we will give solutions for (2.7) and (2.8):

(a) Solution for (2.7):

Since the objective function at each pixel is independent from others, this problem is a 1-D minimization problem and can be easily solved by calculus. The solution is given by:

$$V_p = \begin{cases} I_p - S_p - \theta & \text{if } I_p - S_p > \theta \\ I_p - S_p + \theta & \text{if } I_p - S_p < -\theta \\ 0 & \text{if } | I_p - S_p | \leq \theta \end{cases} \tag{2.9}$$

Such solution is also given in [30].

(b) Solution for (2.8):

The objective function in problem (2.8) has a quadratic term, which is very similar to the original RTV-L^2 problem in [20]. Therefore, we can use a similar iterative solution to that proposed in [20] to solve problem (2.8) approximately. As shown in [20], the objective function in (8) can be approximated with a matrix form:

$$\begin{aligned} &\sum_p \lambda \cdot RTV_{S,p} + \frac{1}{2\theta}\left(S_p + V_p - I_p\right)^2 \approx \\ &\quad \left(v_S - v_{I-V}\right)^T \left(v_S - v_{I-V}\right) + 2\theta\lambda \cdot v_S{}^T \left(C_x{}^T U_{S,x} W_{S,x} C_x + C_y{}^T U_{S,y} W_{S,y} C_y\right) v_S \end{aligned} \tag{2.10}$$

where v_Q is the vector representation of matrix Q, $C_{x(y)}$ is a Toeplitz matrix from gradient operator in x or y direction. $U_{S,x}$ and $W_{S,x}$ are diagonal matrices, whose values on the diagonals are respectively

$$\begin{aligned} U_{S,x}[p,p] &= (G_\sigma * \frac{1}{\left|G_\sigma * \partial_x S\right| + \varepsilon})_p \\ W_{S,x}[p,p] &= \frac{1}{\left|(\partial_x S)_p\right| + \varepsilon'} \end{aligned} \tag{2.11}$$

where p is the pixel index in the vector representation of the image, ε and ε' are newly introduced small positive constants for preventing division by zero. After the approximation, let:

$$L = C_x{}^T U_{S,x} W_{S,x} C_x + C_y{}^T U_{S,y} W_{S,y} C_y \tag{2.12}$$

Considering L as a constant and computing the value of L using the results from last iteration, then the minimization problem (2.8) boils down to the following:

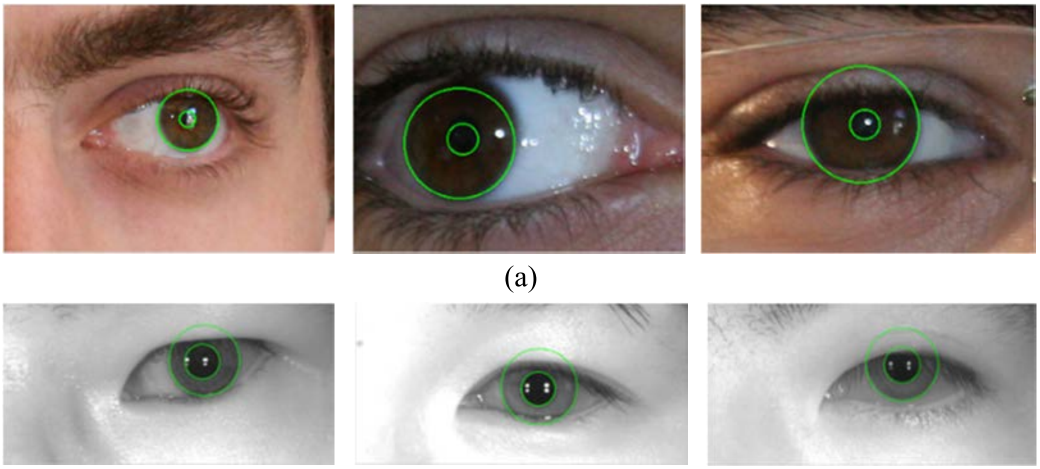
$$(\mathbf{1} + 2\theta\lambda L) \cdot v_S = v_{I-V} \tag{2.13}$$

The problem in (2.12) is easy to solve using knowledge of linear algebra. As the number of iteration increases, the output approaches to the optimal solution and the value of the energy function in (2.6) keeps reducing until it converges to a stable level. Currently we iterate five times for each eye image based on the observation on the output and receive a satisfying noise removal effect, as shown in Figure 2.3.

2.2.3 Coarse Iris Localization Using a Circle

As discussed in section 2.2, a simple circular model cannot be employed to accurately segment iris images acquired under less constrained environments. However, it is widely observed that the human iris can be coarsely approximated as a circle [7], [26]. A circular boundary that coarsely but closely fits the limbic boundary can be used to further refine the boundaries for accurate iris segmentation using a series of efficient post-processing algorithms. In this work, we refer to such a coarse localization circle as an *iris circle*. Similarly, the *pupil circle* describes the circular boundaries that coarsely fit the pupillary boundary of iris images acquired for the segmentation.

After structure extraction, the noise of the eye images is highly suppressed and it is possible to use the circular Hough transform (CHT) based approach to detect the iris and pupil circles coarsely, which highly relies on the clarity of the image structure. We implemented an improved version of CHT based on the two-phase CHT introduced in [19]. Firstly, we only detect the lower half circles to prevent possible interference from the eyelashes or eyebrow. Secondly, after the first phase in [9] which estimates the circle center, we enabled re-searching for the circle center within a rectangular region around the estimated center, to more accurately detect the center position and the radius. The robustness for coarsely localizing the iris region increases with the improved CHT. We detect the circles with empirically proper radius ranges, whose sample results are shown in Figure 2.4. The possible ranges of radius for the databases we used, *i.e.*,



(a)

(b)

Figure 2.4: Sample results from the iris and pupil circle localization for (a) VW images and (b) NIR images.

UBIRIS.v2, FRGC and CASIA.v4-distance, are [35, 120], [25, 40] and [60, 100] respectively.

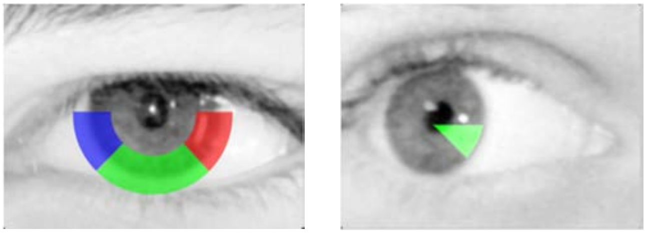
2.2.4 Iris Pixel Identification by Local Gray Level Analysis

Automated boundary refinement approach has to be developed to accurately identify the limbic boundaries after the *iris circle* is detected. We developed an adaptive histogram-based binarization approach to firstly process lower half pixels of the *iris circle* in the image.

A. Adaptive Detection of Lower Half Iris and Sclera Boundary

The reason for processing the lower part firstly is that the lower half iris is less likely to be affected by eyelash and eyelid. Therefore, accurately identifying the iris pixels in the lower half region is firstly considered in identifying noisy pixels using the thresholding. Firstly processing the lower half not only can improve segmentation accuracy but also help to detect the thresholds for accurately segmenting the upper half.

The lower half circle is firstly processed by performing *N* sector thresholdings. In one thresholding, pixels in a certain sectorial region as expressed in the following are



(a) (b)

Figure 2.5: Illustration of three sectorial regions to be processed (a) and the Otsu's thresholding result for one sectorial region (b).

identified:

$$C_{\phi_1,\phi_2} = \left\{ p \;\middle|\; t_1 r_{ir} \le \left|\overrightarrow{cp}\right| \le t_2 r_{ir} \text{ and } \phi_1 \le \theta_p \le \phi_2 \right\} \tag{2.14}$$

where c and r_{ir} are the center and radius of *iris circle* respectively, θ_p is the angle from *x axis* to the vector $\overrightarrow{cp}$, says the central angle at point p, $[\phi_1, \phi_2]$ is the range of central angles with $0 \le \phi_1 < \phi_2 \le \pi$, $[t_1, t_2]$ is the constant ratio range to the iris radius restricting the region of the sector, and is empirically set to [0.6, 1.35]. In our approach, N is set to 3, and the sequence of ranges of central angels are $[0, \frac{\pi}{4}]$, $[\frac{\pi}{4}, \frac{3\pi}{4}]$ and $[\frac{3\pi}{4}, \pi]$ respectively. These sectorial regions are also shown in Figure 2.5 (a).

If the edge is clear and the *iris circle* is accurate, we can choose a threshold value that separates the low end and high end of the pixel values inside the sectorial region. Otsu's method is a good approach for such purpose. It can automatically locate valley point between two peaks in the histogram of a set of pixel values using two-class separation metric. The significant aspect of our strategy is that we adopt different threshold values at each of the different sectors, which ensures that the overall error in the identification of iris pixels is significantly reduced. In addition, the number of sectors and range of angle sequences can be varied to accommodate iris images of degraded quality. Note that the acquired eye images suffer from serious noise and occlusions. If each segment is too small (N is large), the computed threshold may not

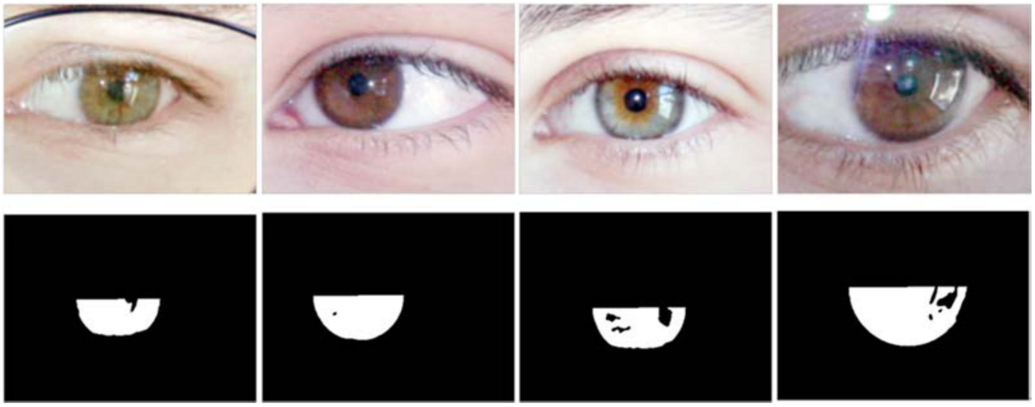


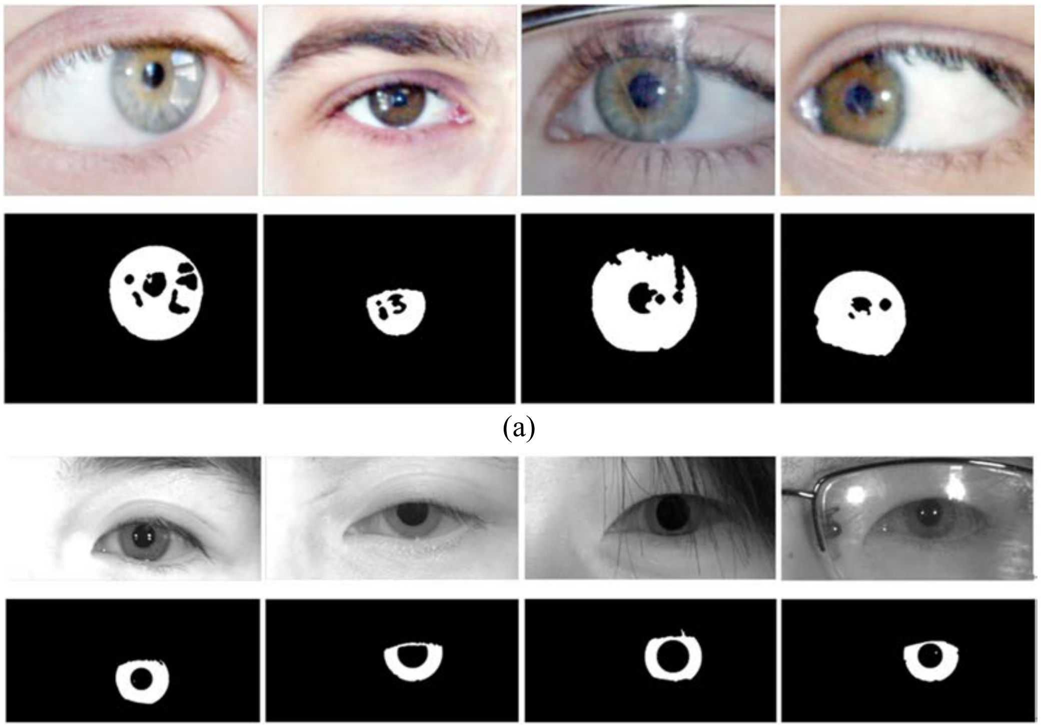
Figure 2.6: Sample iris images and corresponding results from post processing of lower half iris pixel region.

be robust and the computational time will also increase. Therefore, $N=3$ is a reasonable choice. In [10], the threshold is obtained from statistical information of pixel values within a region near the pupil. Such method applies only one fixed threshold for the whole circular boundary, which may not fit local features very well. Figure 2.6 shows the sample results from the post-processing of lower half of iris region pixels.

B. Coarse-to-Fine Localization for Upper-Half Iris, Pupil Region and Reflection

The upper half part is expected to be highly noisy, which is caused by the eyelash and shadow, and quite a significant part of the iris is occluded by eyelid and therefore the sector thresholding may not work well here. We can reuse the previous thresholds from the sectorial thresholding described in section 2.3.4.A. We just segment the upper-left 1/4 circle using threshold determined in $C_{\frac{3\pi}{4},\pi}$, and the upper right one using threshold determined in $C_{0,\frac{\pi}{4}}$. This approach is not expected to cause big error because two pairs of these regions are continuously connected, and further refinement regarding eyelid, eyelash and shadow will be performed.

Since we have already detected the *pupil circle* earlier, the pupil removal step is



(a)

(b)

Figure 2.7: Sample results after upper half masking, pupil removal and source reflection removal for (a) VW images and (b) NIR images.

to eliminate the pupil region pixels from the iris circle in previous step. Another effect of the sector thresholding is that the detected threshold can be used to identify source reflections that usually exist in the images acquired under less-constrained imaging environment and occlude the iris region. We eliminate pixels whose gray levels are higher than the highest threshold among all the three sectors in the lower half iris processing section (Figure 2.5). In summary, the pixels which are brighter than the brightest pixels in lower half of iris region are considered as source reflection. Figure 2.7 illustrates some sample results after masking upper half iris, eliminating pixels belonging to pupil and source reflection.

C. Identifying Eyelid, Eyelash and Shadow (ES)

As discussed earlier, the ES region brings much noise and ambiguity in the segmentation process. It is important to carefully identify this restricted region to

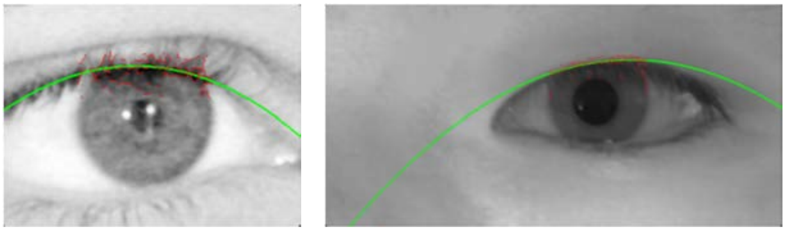


Figure 2.8: Sample results of the proposed eyelid fitting approach. Green curve is the fitted parabola representing upper eyelid, and the red points are the edge points detected by the canny edge detector in region *R*.

perform any refinement. Therefore, the position of upper eyelid should be accurately located.

(i) Eyelid Fitting

Using a parabola to approximate the eyelids is a popular approach in many iris segmentation algorithms and is found to have higher performance than other approaches [36]. Therefore we also propose to fit the eyelid with a parabola, which is in the following form:

$$y - c = a\left(x - b\right)^2 \qquad (2.15)$$

Considering the shape of the upper eyelid and in order to fasten the parameter searching, we limit the ranges of *a*, *b* and *c* as follows:

$$\begin{cases} 0 < a < 1/r_{ir} \\ x_c - 2 \cdot r_{ir} < b < x_c + 2 \cdot r_{ir} \\ y_c - 1.5 \cdot r_{ir} < c < y_c - 0.3 \cdot r_{ir} \end{cases} \qquad (2.16)$$

where (x_c, y_c) and r_{ir} are the center and radius of *iris circle* respectively. The range of *a* ensures that the parabola is orienting downwards and will not be too sharp, and the ranges of *b* and *c* make the vertex of the parabola not too far away from the iris.

The approach we propose to search the parabola is simple and yet very effective in terms of speed and accuracy. First, we define a rectangular region as the candidate eyelid area as follows:

$$R = \{(x, y) \mid x_c - r_{ir} \le x \le x + r_{ir}, y_c - r_{ir} \le y \le y_c - 0.3 \cdot r_{ir}\} \qquad (2.17)$$

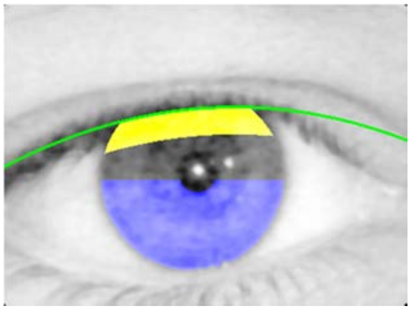


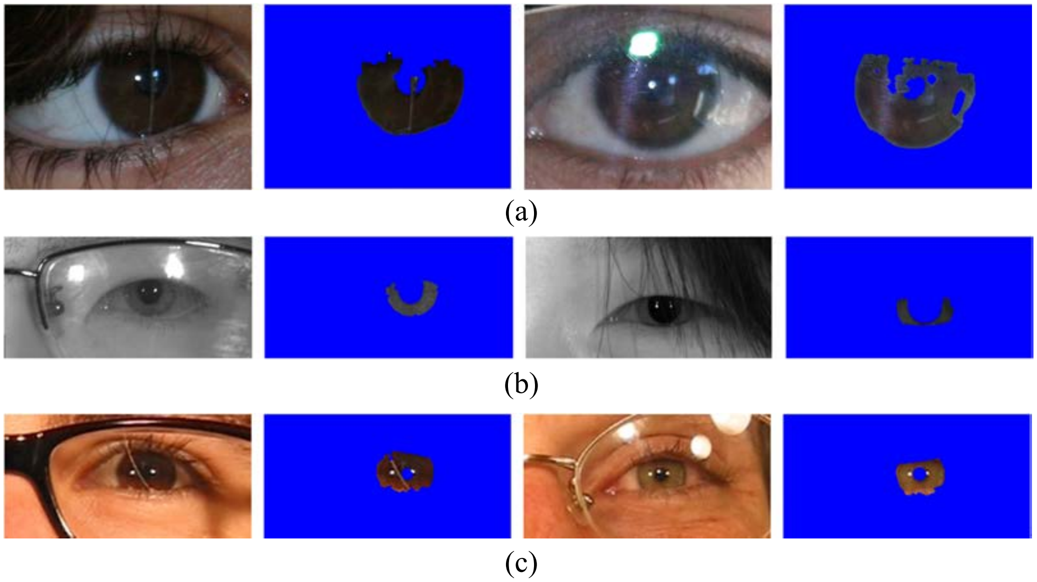
Figure 2.9: Illustration of ES processing. Pixels in the blue region are collected to calculate thresholds to process the pixels in the yellow region.

A Canny edge detector is applied in *R* and let us donate the set of detected edge points as *E*. We assume that among the edge points in *E*, some are close to the position of the eyelid, which we refer to as eyelid points, and some points belong to noise such as eyelashes and shadow, which we refer to as non-eyelid points. The spatial distribution of the non-eyelid points is highly random and less regular, while the positions of the eyelid points are very close to the parabola that can accurately fit the real eyelid. Therefore, we search for a parabola with the parameters $\{a, b, c\}$ that has maximum number of points in *E* lying on it. Moreover, we actually search for $\{a, b, c\}$ at discrete intervals so the speed can be greatly fastened. Figure 2.8 illustrates two sample results of the proposed eyelid fitting approach, which is highly accurate.

(ii) Eyelash and Shadow Processing

Having located the upper eyelid, the next step is to mask out those pixels which are belonging to the eyelashes and shadow at a certain distance below the eyelid. This step is the same as described in [10] and we choose the distance as $0.3 \times r_{ir}$. The pixel values within lower half of the currently processed iris mask are used to detect thresholds to identify those belonging to ES region. Figure 2.9 illustrates the idea.

We choose the limiting thresholds that exclude 1% of the darkest pixels and 20% of the brightest pixels as the low and high thresholds respectively. Only the pixels between these two thresholds are retained. In order to eliminate isolated noisy pixels, the iris mask is subjected to an opening operation. Figure 2.10 illustrates some sample



(a)

(b)

(c)

Figure 2.10: Sample source images and corresponding final segmentation results (non-iris region is masked with blue color) for (a) VW images from UBIRIS.v2, (b) NIR images from CASIA.v4-distance and (c) VW images from FRGC.

iris segmentation results from the databases used in this work.

2.3 Experiments and Results

2.3.1 Databases

We have used three publicly available databases, UBIRIS.v2 [18], FRGC [28] and CASIA.v4 [27] to perform experiments for iris segmentation and recognition under VW and NIR imaging. The images from these databases were acquired under less-constrained environments. It is judicious to expect that good performance on these databases indicates higher probability for the proposed approach to work well in surveillance and forensics applications. The summary of the employed datasets is presented in Table 2.1. We selected these subsets subject to the availablity of the ground truth iris masks (explained in section 2.4.2). Kindly note that images in CASIA.v4 (distance subset) and FRGC databases are full/partial face images, and we used the publicly available face and eye-pair classifiers [76] to extract the eye image

Table 2.1: Summary of databases employed in the experiments.

	UBIRIS.v2	**CASIA.v4-distance**	**FRGC**
Imaging illumination	visible	near-infrared	visible
Standoff distance	4 – 8m	⩾3m	N/A
Eye image size	400×300	about 780×400	300×150
No. of subjects	171	77	163
No. of images	1,000	581	540

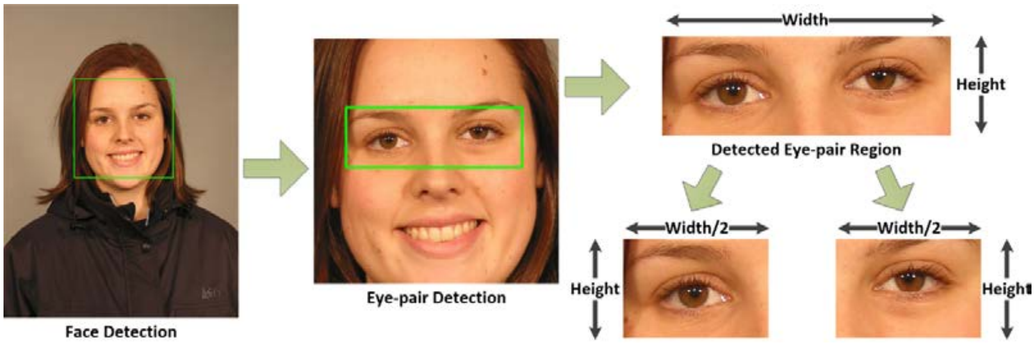


Figure 2.11: Illustration of eye image extraction from face images. Left and right eye regions are simply cropped with equivalent width from the detected eye-pair region.

from the original images in these two databases for the subsequent operations. Face detection is firstly employed for the face images from FRGC and the eye-pair is detected from the detected face area. However, for the images in CASIA.v4 database the face detection is skipped because the images contain only face region. A sample image from FRGC and the detected face image along with eye image are shown in Figure 2.11. The size of the eye-pair region is adaptive to the size of the face image. However, since the size of the face images in FRGC varies significantly, we scaled the cropped eye images to a consistant size of 300×150. For CASIA.v4-distance database, since the size of the face images varies little, we did not perform re-scaling.

As for the parameter tunning, there are mainly two types of parameters. The first one is those related to the proposed RTV-L^1 solution. We use the same set of parameters ($\lambda=0.2, \theta=0.05, \sigma=3, \varepsilon=\varepsilon'=0.005$) for all three databases, which illustrates that the proposed RTV-L^1 is highly generalizable for images captured in various condition.

Table 2.2: Comparison of average segmentation error rates for different approaches.

	Iris Segmentation Error, $\bar{e}$(%)		
Approaches	**UBIRIS.v2**	**CASIA.v4-diatance**	**FRGC**
Proposed RTV-L^1	1.21	0.68	1.27
RTV-L^2	1.41	0.75	1.28
Li & Savvides (T-PAMI'13) [9]	1.92	0.85	1.34
Tan & Kumar (T-IP'13) [10]	1.70	0.81	1.84
Tan & Kumar (T-IP'12) [13]	1.90	1.13	1.84
Proença (T-PAMI'09) [11]	3.75	1.61	2.42
Tan *et al.* (ImVis'10) [12]	3.49	1.71	3.30

Other parameters are mainly database-specific, such as the range of radius of iris circle. Such parameters should be adjusted according to the image resolution.

2.3.2 Performance Evaluation

A. Segmentation Accuracy

The accuracy of iris segmentation is evaluated using the same protocol as in the NICE. I competition [21], in which the average segmentation error rate is computed as follows:

$$\bar{e} = \frac{1}{N \times w \times h} \sum_{x \in w} \sum_{y \in h} T(x, y) \oplus M(x, y) \tag{2.18}$$

where N is the total number of images, w and h are width and height of one image, T and M are the ground truth mask and generated iris mask respectively. The symbol $\oplus$ represents an exclusive OR operation to identify the segmentation error. While ground truth of UBIRIS.v2 are manually labeled and publicly provided by NICE.I, ground truth for the other two datasets is also manually generated by authors of [10] and made publicly available. Therefore, we can use the NICE.I protocol for the consistent segmentation accuracy evaluation. Kindly note that this metric does not separately panelize false positive and false negative in the iris segmentation task. It would be difficult to determine whether it is better to have more oversegmented or undersegmented pixels for the subsequent iris recognition problem, as it should depend on the properties of the feature extractor. For instance, if the iris feature representation is robust to existance of noise to some extent, it may be desired to include more pixels, even with some falsely segmented ones, for richer information; otherwise, if the feature is sensitive to noisy pixels, undersegmentation may cause less degradation to the recognition task.

Table 2.2 summarizes the performance from state-of-the-art approaches in the recent literature while using the above protocol[2]. The approaches [10] and [13] are the work from my colleague, who also re-implemented methods [11] and [12]. The results of [9] are from my re-implementation with the assistance from the original authors. All the baselines methods have been extensively tuned to achieve their best possible performance. It can be observed from Table 2.2 that the proposed approach has achieved average segmentation error rates of 1.21%, 0.70% and 1.29% for UBIRIS.v2, CASIA.v4-distance and FRGC respectively.

The comparative statistics suggest that the proposed approach consistently outperforms other iris segmentation methods developed in the literature. As compared with the recent approach published in [10], the proposed method can achieve average improvement of 28.82%, 16.05% and 30.98% for UBIRIS.v2, CASIA.v4-distance and FRGC databases respectively, in iris segmentation accuracy. It may be noted that the method described in [12] was *ranked first* in NICE. I competition [21] and therefore provides a good benchmark for the comparison. We have also evaluated the performance when using the original RTV-L^2 approach for structure extraction and keeping other steps exactly the same. The results in Table 2.2 show that the proposed RTV-L^1 has noticable superiority over RTV-L^2 due to its ability to preserve sharpness

[2] The average error rate of algorithm in [9] is also produced from our implementation and is made available via [34].

Table 2.3: Results of t-test between the proposed approach and other state-of-the-art methods.

		UBIRIS.v2	**CASIA.v4-diatance**	**FRGC**
p-value	over [9]	3.2e-21	2.3e-10	0.10
	over [10]	4.0e-11	4.9e-4	2.6e-28

H0: Proposed method does not outperform the comparative method significantly.
H1: Proposed method outperforms the comparative method significantly.

of important edges.

We would like to give special explanation on the performance of approach [9] for comparison, which are 1.92%, 0.85% and 1.34% in terms of segmentation error rate for UBIRIS.v2, CASIA.v4-distance and FRGC databases respectively. Different from other comparative methods listed in Table 2.2, reference [9] only focuses on the steps after the iris images are normalized, and this reference method manually localizes iris and pupil circles for iris normalization, while our proposed method automatically localizes iris and pupil circles. In order to provide a fair comparison, we used our proposed approach in all the prior steps (SSR enhancement, RTV-L^1 structure extraction and circle localization) and adopted the method in [9] for post-processing.

In order to ascertain that the performance improvements achieved using the proposed method are statistically significant, we further conducted the significance test on the results from the proposed approach and those in [9] and [10]. Since the segmentation error rates from a specific approach can be treated as a sequence of independent and identically distributed data, *t*-test can be used for evaluating the statistical significance of the difference between the results from two methods. The results of the *t*-test are summarized in Table 2.3. These results suggest that under the confidence level of 95% (p-value < 0.05), the performance improvements from our method are statistically significant for most of the comparisons, despite on FRGC there is marginal improvement over method [9].

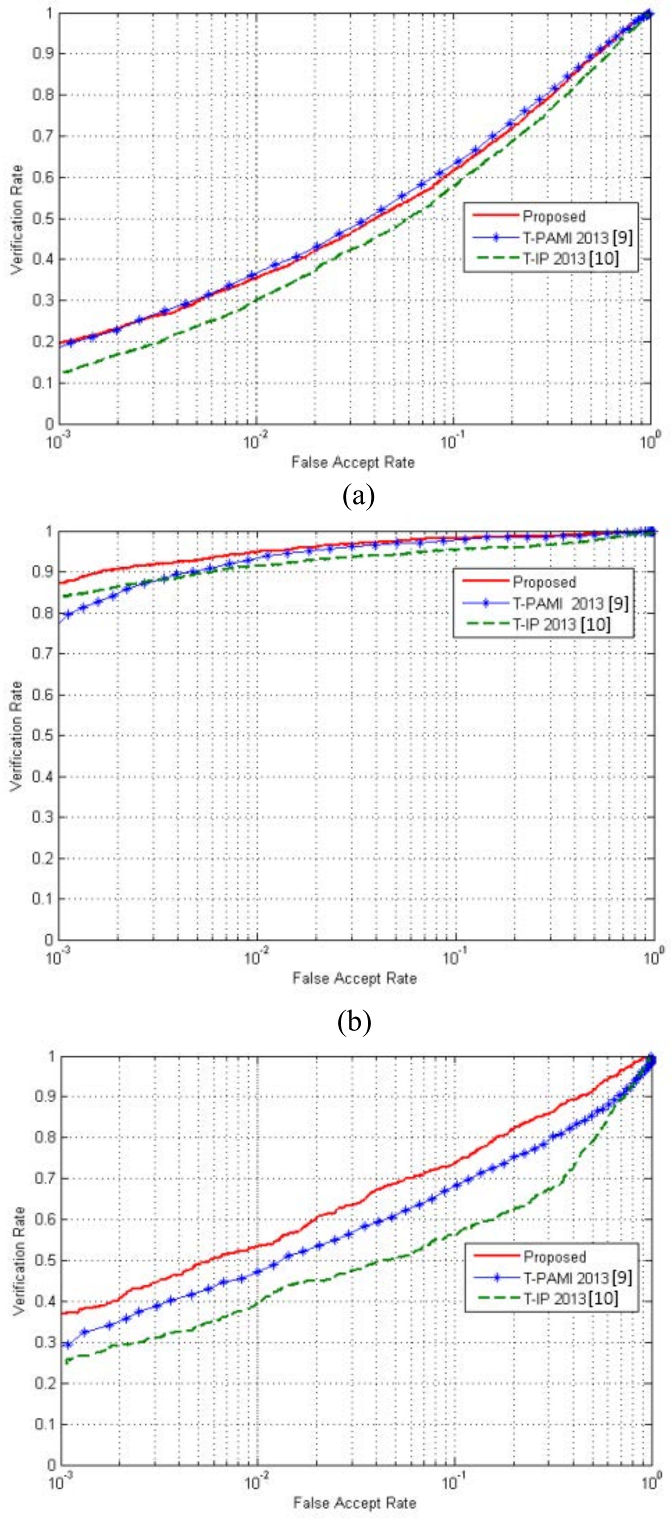
Table 2.4: Summary of training and testing protocols used in the experiments for the recognition.

	UBIRIS.v2	**CASIA.v4-diatance**	**FRGC**
#Training subjects/images*	19/96	10/79	13/40
#Test subjects/images	152/904	67/502	150/500
Optimized parameters	$\lambda = 22$ $\sigma / f = 0.30$	$\lambda = 23$ $\sigma / f = 0.35$	$\lambda = 18$ $\sigma / f = 0.45$

B. Recognition Performance

It should be noted that the objective of this work is to develop a robust approach for iris segmentation and achieve significantly improved segmentation accuracy rather than iris recognition performance. It is reasonable to argue that improved iris segmentation should lead to improved iris recognition, but not always because in many cases non-iris pixels appearing due to the poor iris segmentation can be consistent and aid to the improved matching of such iris images. However, the recognition performance is always the first concern in iris recognition systems. To answer a possible query from the readers on the performance improvement, we have also performed some experiments on the recognition on each of these public databases.

For the experiments on recognition, we adopt the 1D log-Gabor filter as the feature encoding method, which is widely used in the deployed iris recognition systems, and use iris masks generated from different segmentation approaches for comparison. The log-Gabor filter involves two parameters, central wave length (λ) and Gaussian standard deviation over central frequency (*sigmaOnf*, σ / f), which are critical for the final performance. To select the best parameters, we divided each database into non-overlapping training set and testing set, and the division method is the same as that in [10]. The parameter sets within closed regions were adopted for the training sets, and the one giving best GAR@FAR=1% was selected as the



(a)

(b)

(c)

Figure 2.12: ROC curves of iris recognition experiments using iris masks generated from different segmentation approaches for (a) UBIRIS.v2, (b) CASIA.v4-distance and (c) FRGC.

optimized set of parameters. Besides, we used the ground truth masks in the training process so that the selected parameters do not have bias on any of the segmentation approaches we are comparing. Table 2.4 gives the detailed separation of training/testing sets and the optimized parameters for three databases. It can be inferred that using the optimized parameters consistently for the log-Gabor filter for each database, the only important factor that impacts the recognition performance would be the iris segmentation approach.

The ROC curves for the employed datasets using iris masks from comparative approaches are shown in Figure 2.12. From the comparison we can see that the experiments using the proposed iris segmentation approach produce better ROC than those using other segmentation approaches, clearly for FRGC and CASIA.v4-distance. For UBIRIS.v2, the proposed approach also improves the verification rate at lower false accept rate (FAR). Above experiments illustrate that the proposed iris segmentation approach not only provides the best segmentation accuracy but also offers noticeable improvements in the final iris recognition performance. Nevertheless, it should be clarify that higher segmentation accuracy does not always constitute to better recognition results, as non-iris pixels such as eyelash edge points may also form some discriminative patterns, especially when the overall performance is not satisfactory, i.e., when it is difficult to sufficiently exploit iris textures.

C. Execution Speed

The proposed approach is computationally simpler and highly attractive for online applications. Our proposed iris segmentation framework was implemented in Matlab 2012b and run on a computer with 2GB RAM and a 2.0 GHz Intel Core2 Dual Core CPU. The average execution time for three databases is shown in Table 2.5. It is reasonable to expect that the execution time can be significantly reduced by implementing the code in C/C++ using multiple threading and GPU, *etc.* Currently no

Table 2.5: Average computational time of the proposed approach for automated iris segmentation.

Databases	**Execution Time (seconds / image)**
UBIRIS.v2	1.37
CASIA.v4-distance	1.26
FRGC	0.88

significant efforts were employed to optimization of the code.

2.4 Summary

This work has developed a more accurate iris segmentation framework to automatically segment iris images acquired under less-constrained imaging environment. The proposed approach introduces a new total-variation based energy regularizer incorporated with an l^1 norm, in which the slowly varying components of image structure such as eyelid, limbic boundaries, etc., and the surrounding texture and noise are differently penalized. In addition, an efficient solution for the proposed energy regularizing formulation is given. Such an approach allows us to reliably extract the eye structure for more accurately localizing iris and pupil circles for further segmentation. Our work also introduced a series of novel post-processing operations that exploit local (but often varying) distribution characteristics to adaptively refine pupillary and limbic boundaries. The overall framework has shown to be highly robust to achieve significant improvement in segmentation accuracy as well as iris recognition performance from publicly available iris databases that are under both VW and NIR spectrum.

The RTV-L^1 texture removal approach introduced in this work is not only significant for the noisy iris segmentation but can also be potentially employed to solve other texture or object segmentation tasks which require removal of accompanying noise. The adaptive local intensity analysis developed in our work has been greatly successful in increasing the robustness of the proposed approach under less-constrained imaging. Such adaptive decision-making strategies can also be effectively used in other challenging problems in surveillances and remote sensing that often suffer from less stable illumination conditions and unwanted occlusions. The framework developed in this work provides robust and effective prerequisite for researchers and applications which attempt to perform accurate iris recognition on noisy images acquired under less-constrained environment and at-a-distance.

CHAPTER 3

Accurate Iris Recognition Using Deeply Learned Spatially Corresponding Features

3.1 Background

Automated iris recognition systems have been widely deployed for various applications from border control [65], citizen authentication [66], forensic [67] to commercial products [68]. The usefulness of iris recognition has motivated increasing research effort in the past decades for exploring more accurate and robust iris matching algorithms under different circumstances [1, 2, 8-11]. In recent years, deep learning has gained tremendous success especially in the area of computer vision, and accomplished state-of-the-art performance for a number of tasks such as general image classification [58], object detection [59] and face recognition [56] [60]. However, unlike face, in the field of iris recognition, in the best of to our knowledge, there is almost no attention to incorporate the remarkable capabilities of deep learning and achieve superior performance over popular or state-of-the-art iris recognition methods.

3.1.1 Limitations of Existing Works

Despite the popularity of iris recognition in biometrics, conventional iris feature descriptors have several limitations. The summaries of earlier work in [5] and [50] reveal that existing methods can achieve satisfactory performance, but the performance needs to be further improved to meet the expectations for wider range of deployments. Besides, traditional iris features, such as *IrisCode*, are mostly based on empirical models which apply hand-crafted filters or feature generators. As a result, these models rely heavily on parameter selection when applied for different databases or imaging environments. Although there are some standards on iris image format [70], the selection of parameter for feature extraction remains empirical, or based on training methods such as boosting [71]. This situation can be observed from [16], where eight different combinations of parameters for ordinal filters delivered varying performance on three databases, or from [26] which employed two sets of parameters for log-Gabor filter on two databases by extensive tuning. Another limitation is that due to the simplicity of conventional iris descriptors, they are less promising to fully exploit the underlying distribution from various types of iris data available today. Learning data distribution from large amount of samples to further advance performance is one of the key trends nowadays.

Deep learning has the potential to address the above limitations, since the parameters in deep neural networks are learned from data instead of being empirically set, and deep architectures are known to have good generalization capability. However, new challenges emerge while incorporating typical deep learning architectures (*e.g.*, CNN) for the iris recognition, which can be primarily attributed to the nature of iris patterns. Different from face, iris pattern is observed to reveal little structural information or meaningful hierarchies. Iris texture is believed to be random [72]. Earlier promising works on iris recognition [1, 2, 8-10] mainly employed small-size filters or block-based operations to obtain iris features. Therefore, we can infer that the most discriminative information in the iris pattern comes from the local intensity distribution of an iris image rather than the global features, if any. CNN is known as effective for extracting features from low level to high level, and from local to global, due to the combination of convolutional layers and fully connected layers [61]. However, as discussed above, high level and global features may not be the optimal for iris representation.

3.1.2 Our Work

We aim to develop a more accurate and robust deep learning based iris feature representation framework, making solid contributions towards fully discovering the potential of deep learning for the iris recognition. Such objectives have not been sufficiently pursued in the literature. In this chapter we propose a new deep learning based iris recognition framework which not only achieves satisfactory matching accuracy but also exhibits outstanding generalization capability to different databases. With the design of an effective fully convolutional network, our model is able to significantly reduce parameter space and learn comprehensive iris features which generalize well on different datasets. A newly developed *Extended Triplet Loss* (*ETL*) function provides meaningful and extensive supervision to the iris feature learning process with limited size of training data.

While most of the contents presented in this chapter have been published in [52], in this thesis we extend this work by developing a new end-to-end binary iris feature learning mechanism to improve feature robustness as well as theoretical soundness. The previous approach adopts an ad-hoc feature binarization step to empirically exploit robustness of binary feature for iris recognition. However, such a process is not part of the deep network and largely handcrafted, which may lead to reduced adaptiveness of the matching process on generalized data. In this paper we combined the binary feature representation into training of our deep model so that it can be end-to-end, which enhances the train/test consistency and improves the recognition results.

The main contributions of this work can be summarized as follows: (i) We develop a new deep learning based iris recognition framework which is highly generalizable for operating on different databases that represent diverse deployment environments. A new *Extended Triplet Loss* function has been developed to successfully address the nature of iris pattern for learning comprehensive iris features (more details in Section

3.2.2 and 3.3). Significant advancement therefore has been made to bridge the gap between deep learning and iris recognition. (ii) Under fair comparison, our approach consistently outperforms several state-of-the-art methods on multiple datasets. Even under challenging scenario that without having any parameter tuning on the target dataset, our model can still achieve superior performance over state-of-the-art methods that have been extensively tuned. (iii) We propose a new mechanism to directly learn binary iris features from our networks, which ensures the end-to-end property of our deep model and achieve further improved results. Such mechanism also provides an effective alternative to the solutions to an open research problem in the literature, *i.e.*, learning binary hash codes with deep neural network.

The rest of this chapter is organized as follows: Chapter 3.2-3.4 detail the proposed approach in terms of network architecture, improved triplet loss function and feature encoding respectively; Chapter 3.5 presents the experimental configurations, results and analysis; finally, the brief summary from this work is presented in Chapter 3.6.

3.2 Network Architecture

We have developed a highly optimized and unified deep learning architecture, referred to as *UniNet*, for both iris region masking and feature extraction, which is based on fully convolutional networks (FCN) [56]. A new customized loss function, named *Extended Triplet Loss (ETL)*, has been developed to accommodate the nature of iris texture in supervised learning. The motivations and technical details for the proposed approach are explained in the following sections.

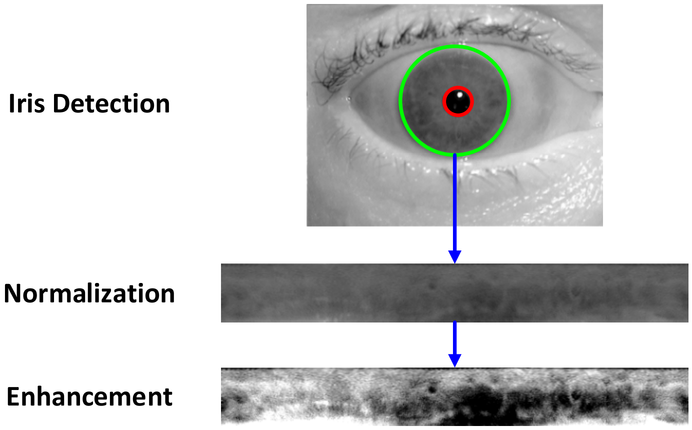


Figure 3.1: Illustration of key steps for iris image preprocessing.

3.2.1 Image Preprocessing

For all the experiments presented in this chapter, we use a recent iris segmentation approach [51] for iris detection, and normalize the iris region pixels from polar coordinate system to Cartesian coordinate system using the classic rubber-sheet model [7]. The resolution after normalization is uniformly set to 64×512 , which is considered as the upper-bound of adequate sizes for the iris images employed in this work to avoid information loss, even though some of the images may not support up to this scale. We then apply a simple contrast enhancement process, which adjusts the image intensity so that 5% darkest pixels and 5% brightest pixels are saturated at low and high intensities respectively. The enhanced images are used as input to the deep network for training and testing. Figure 3.1 illustrates the key steps of image preprocessing.

3.2.2 Fully Convolutional Network

The proposed unified network (termed as *UniNet*) is composed of two sub-networks, *FeatNet* and *MaskNet*, whose detailed structures are presented in Figure 3.2 and Table 3.1. Both of the two sub-networks are based on fully convolutional networks (FCN)

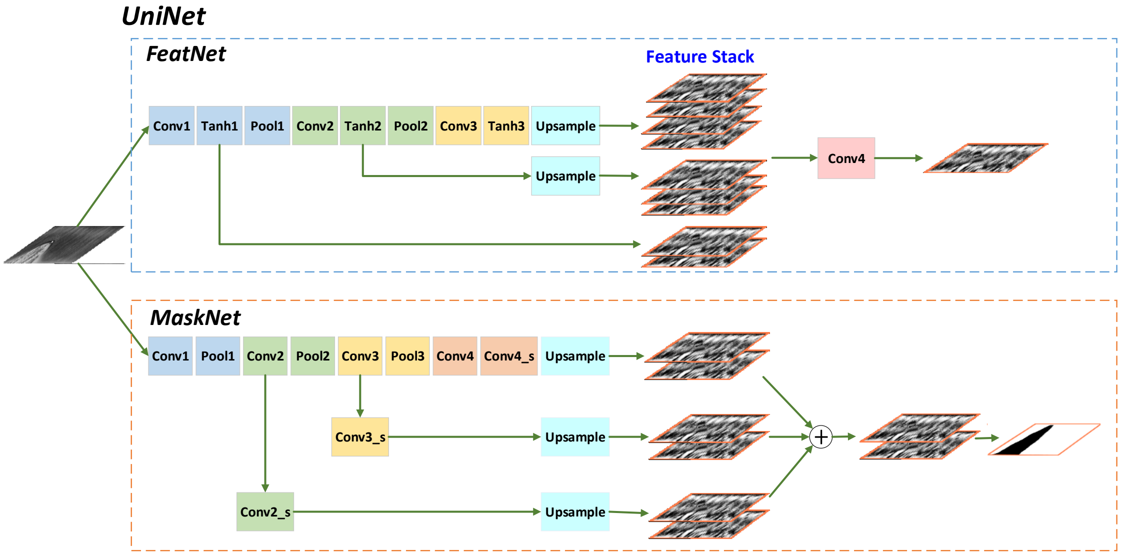


Figure 3.2: Detailed structures for *FeatNet* (top) and *MaskNet* (bottom) respectively. The *FeatNet* generates a single-channel feature map for each sample for matching. The *MaskNet* outputs a two-channel map, on which the values for each pixel along two channels represent the probabilities of belonging to iris and non-iris regions, respectively.

Table 3.1: Layer configurations for *MaskNet* and *FeatNet*.

FeatNet				
Layer	**Type**	**Kernel size**	**Stride**	**# Output channels**
Conv1	Convolution	3×7	1	16
Conv2	Convolution	3×5	1	24
Conv3	Convolution	3×3	1	32
Conv4	Convolution	3×3	1	1
Tanh1, 2, 3	TanH activation	/	/	/
Pool1, 2, 3	Average pooling	2×2	2	/
MaskNet				
Layer	**Type**	**Kernel size**	**Stride**	**# Output channels**
Conv1	Convolution	3×3	1	16
Conv2	Convolution	3×3	1	32
Conv2_s	Convolution	1×1	1	2
Conv3	Convolution	3×3	1	64
Conv3_s	Convolution	1×1	1	2
Conv4	Convolution	3×3	1	128
Conv4_s	Convolution	1×1	1	2
Pool1, 2	Max pooling	2×2	2	/
Pool3	Max pooling	4×4	4	/

which were originally developed for semantic segmentation [56]. Different from common convolutional neural networks (CNN), the FCN does not have fully connected layers. The major components of FCN are convolutional layers, pooling layers, activation layers, etc. Since all these layers operate on local regions of their bottom map, the output map can preserve *spatial correspondence* with the original input image. By incorporating up-sampling layers, FCN is able to perform pixel-to-pixel prediction. In the following we detail the two components of *UniNet*.

- ***FeatNet***

FeatNet is designed for extracting discriminative iris features which can be used in matching. As shown in Figure 3.2, the input iris image is forwarded by several convolutional layers, activation layers and pooling layers. The network activations at different scales, i.e., TanH1-3, are then up -sampled if necessary to the size of original input. These features form a multi-channel feature stack which contains rich information from different scales, and are finally convolved again to generate an integrated single-channel feature map.

The reason for selecting FCN instead of CNN for iris feature extraction primarily lies in the previous analysis on iris patterns, *i.e.*, the most discriminative information of an iris probably comes from small and local patterns. FCN is able to maintain local pixel-to-pixel correspondence between input and output, and therefore is a better candidate for the iris feature extraction.

- ***MaskNet***

MaskNet is set to perform non-iris region masking for normalized iris images, which can be regarded as a specific problem for the semantic segmentation. It is basically a simplified version of the FCNs proposed in [56]. Similar to those in [56], *MaskNet* is supervised by a pixel-wise softmax loss, where each pixel is classified into one of two classes, *i.e.*, iris or non-iris. In our practice, *MaskNet* is trained with 500 randomly

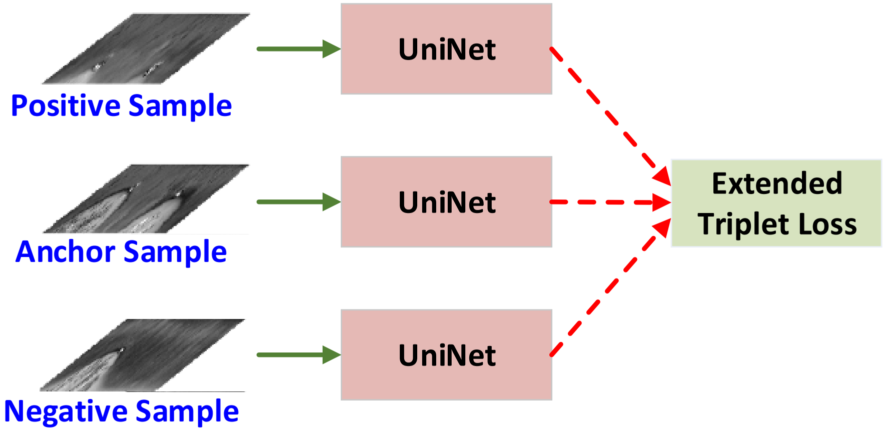


Figure 3.3: Triplet-based network organization for training.

selected samples from the training set of ND-IRIS-0405 database, and the ground truth masks are manually generated by us. We would like to declare that the main focus of this work is on learning effective iris feature representation. *MaskNet* is developed to provide adequate and immediate information for masking non-iris regions, which is necessary for the newly designed loss function (will be detailed in Section 3.3) and also for the matching process. The placement of *MaskNet* in the unified network also preserves the possibilities that iris masks may be jointly optimized/fine-tuned with the feature representations, which is one of our future research goals. At this stage, however, *MaskNet* is pre-trained and fixed during learning the iris features. A sample evaluation for its performance is provided in Chapter 3.5.

3.2.3 Triplet-based Network Architecture

A triplet network [57] was implemented for learning the convolutional kernels in *FeatNet*. The overall structure for the triplet network in the training stage is illustrated in Figure 3.3. As shown in the figure, three identical *UniNets*, whose weights are kept identical during training, are placed in parallel to forward and back-propagate the data and gradients for anchor, positive and negative samples respectively. The anchor-positive (AP) pair should come from the same person while the anchor-negative (AN) pair comes from different persons. The triplet loss function in such architecture attempts to reduce the anchor-positive distance and meanwhile increase the anchor-negative distance. However, in order to ensure more appropriate and effective supervision in the generation of iris features by the FCN, we improve the original triplet loss by incorporating a bit-shifting operation. The improved loss function is referred to as *Extended Triplet Loss* (*ETL*), whose motivation and mechanism are detailed in the next chapter.

3.3 Extended Triplet Loss Function

In this work we develop a problem-specific loss function for more effective iris feature learning. The new function has two versions, one operating on real-valued features and the other is for binary feature codes. The reason for developing the binary version is that, as indicated by a vast of studies on iris recognition (e.g., [7] [8] [14]-[17]), binary features are believed to be more suitable for iris pattern representation and can be robust to noise. Hence, an end-to-end deep learning framework which can directly learn binary iris features would be worth investigating. In the following we will introduce the two versions of the newly developed loss function.

The original loss function for a triplet network is defined as follows:

$$L = \frac{1}{N_B} \sum_{i=1}^{N} \left[\left\| \boldsymbol{f}^{A}{}_{i} - \boldsymbol{f}^{P}{}_{i} \right\|^2 - \left\| \boldsymbol{f}^{A}{}_{i} - \boldsymbol{f}^{N}{}_{i} \right\|^2 + \alpha \right]_+ \tag{3.1}$$

where N_B is the number of triplet samples in a mini-batch, $\boldsymbol{f}^{A}{}_{i}$, $\boldsymbol{f}^{P}{}_{i}$ and $\boldsymbol{f}^{N}{}_{i}$ are the feature maps of anchor, positive and negative images in the *i*-th triplet respectively. The symbol $[\bullet]_+$ is the as same as used in [57] and is equivalent to $\max(\bullet, 0)$. α is a preset parameter to control the desired margin between anchor-positive distance and anchor-negative distance. Optimizing above loss will lead to the

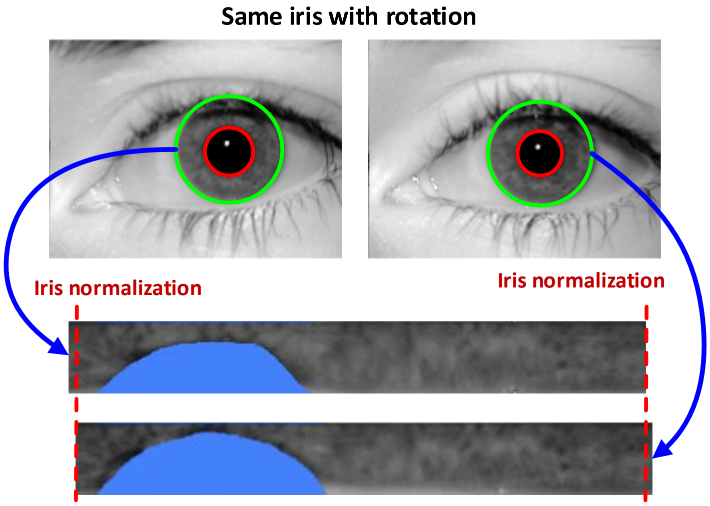


Figure 3.4: Illustration of occlusions (labeled in blue) and horizontal translation which usually exist between two normalized iris images even from a same iris.

anchor-positive distance being reduced and anchor-negative distance being enlarged until their margin is larger than a certain value.

In our case, however, using Euclidean distance as the dissimilarity metric is far from sufficient. As discussed earlier, we propose using spatial features which have the same resolution with the input, the matching process has to deal with non-iris region masking and horizontal shifting, which are frequently observed in iris samples as illustrated in Figure 3.4. Therefore in the following, we extend the original triplet loss function, which we refer to as the *Extended Triplet Loss* (*ETL*):

$$ETL = \frac{1}{N_B}\sum_{i=1}^{N}\left[D(\boldsymbol{f}^{A}{}_{i}, \boldsymbol{f}^{P}{}_{i}) - D(\boldsymbol{f}^{A}{}_{i}, \boldsymbol{f}^{N}{}_{i}) + \alpha\right]_{+} \tag{3.2}$$

where $D(\boldsymbol{f}^1, \boldsymbol{f}^2)$ represents the *Minimum Shifted and Masked Distance (MMSD)* function, defined as follows:

$$D(\boldsymbol{f}^1, \boldsymbol{f}^2) = \min_{-B \le b \le B}\left\{FD(\boldsymbol{f}_b^1, \boldsymbol{f}^2)\right\} \tag{3.3}$$

$\boldsymbol{f}_b$ represents a shifted version of f obtained by horizontally shifting it by b pixels, and *FD* denotes the Fractional Distance. The shifted and the original feature maps have the following spatial correspondence:

$$\begin{aligned} \boldsymbol{f}_b[x_b, y] &= \boldsymbol{f}[x, y] \\ x_b &= (x - b + w) \bmod w \end{aligned} \tag{3.4}$$

where x, y are the spatial coordinates and x_b is obtained by shifting the pixel to the left by a step of *b*, assuming *w* is the width of the 2-D feature map. Note that when *x* is less than *b*, the pixel position will be directed to the right end of the map, as the iris map is normalized by unwrapping the original iris circularly and the left end is therefore physically connected with the right end. When *b* is negative, the bit-shifting operation would shift the map to the right by *–b* pixels. The Fractional Distance *FD* in Eq.3 measures the relative difference between two feature maps within non-masked regions only and normalize it by the number of involved pixels:

$$FD(\boldsymbol{f}^1, \boldsymbol{f}^2) = \frac{1}{|M|} \sum_{(x,y) \in M} d(f^1{}_{(x,y)}, f^2{}_{(x,y)}) \tag{3.5}$$

where *M* is the set common non-masked pixel positions for the two feature maps.

The choice of the element-wise difference function $d(\bullet)$ in (3.5) will depend on the version of *ETL* we use, i.e., real-valued or binary version as mentioned above. In the real-valued version, the difference function is set to square of difference:

$$d_{real}(f_1, f_2) = (f_1 - f_2)^2 \tag{3.6}$$

while in the binary version, to measure the fractional Hamming distance, the difference will be result of the exclusive-or operation:

$$\begin{aligned} d_{binary}(f_1, f_2) &= f_1' \oplus f_2' \\ f' = H(f) &= \begin{cases} 1, & \text{if} f \geq 0 \\ 0, & \text{otherwise} \end{cases} \end{aligned} \tag{3.7}$$

(3.3) and (3.5) indicate that the new loss function will only evaluate the difference between features within non-masked areas and a shifting operation will be performed to address the horizontal translation, so that matching of the proposed spatially corresponding iris features is meaningful. In the following we will derive the gradients of the proposed *ETL* in order to perform back-propagation for the learning process. The cases of real-valued and binary versions *ETL* are quite different and therefore we will separately proceed.

3.3.1 Back-propagation for Real-valued *ETL*

The components of the real-valued *ETL* are all differentiable and therefore the computation of gradients is quite straightforward. Firstly, in order to maintain simplicity of the notations for the upcoming derivation, we denote the offsets that fulfill the *MMSD* of AP-pair and AN-pair as follows:

$$\begin{aligned} b_{AP} &= \underset{-B\le b\le B}{\operatorname{argmin}}\left\{FD(\boldsymbol{f}^{A}{}_{b},\boldsymbol{f}^{P})\right\} \\ b_{AN} &= \underset{-B\le b\le B}{\operatorname{argmin}}\left\{FD(\boldsymbol{f}^{A}{}_{b},\boldsymbol{f}^{N})\right\} \end{aligned} \tag{3.8}$$

During the back-propagation (BP) of the training process, the gradients (or partial derivatives) of the new loss on the anchor, positive and negative feature maps need to be computed. For simplicity, let us firstly derive the partial derivative *w.r.t* the positive feature map $\boldsymbol{f}^{P}$. From (3.2) it can be derived that for one sample in the batch:

$$\frac{\partial ETL}{\partial \boldsymbol{f}^{P}} = \begin{cases} 0, & \text{if } ETL = 0 \\ \dfrac{1}{N_B}\dfrac{\partial ETL}{\partial D(\boldsymbol{f}^{A},\boldsymbol{f}^{P})}\dfrac{\partial D(\boldsymbol{f}^{A},\boldsymbol{f}^{P})}{\partial \boldsymbol{f}^{P}}, & \text{otherwise} \end{cases} \tag{3.9}$$

Again from (3.2) we can see that $ETL=0$ is equivalent to $D(\boldsymbol{f}^{A},\boldsymbol{f}^{P}) - D(\boldsymbol{f}^{A},\boldsymbol{f}^{N}) + \alpha \le 0$. We only need to show the derivation when *ETL* is not 0. Let us denote the set of common valid iris pixel positions for AP pair as M_{AP}, from (3.3) and (3.4) we have the following pixel-wise derivatives:

$$\frac{\partial D(\boldsymbol{f}^{A},\boldsymbol{f}^{P})}{\partial \boldsymbol{f}^{P}[x,y]} = \frac{\partial FD(\boldsymbol{f}^{A}{}_{b_{AP}},\boldsymbol{f}^{P})}{\partial \boldsymbol{f}^{P}[x,y]} = \begin{cases} 0, \text{ if } (x,y)\notin M_{AP} \text{ or } ETL = 0 \\ \dfrac{-2}{|M_{AP}|}(\boldsymbol{f}^{A}[x_{b_{AP}},y] - \boldsymbol{f}^{P}[x,y]), \text{ otherwise} \end{cases} \tag{3.10}$$

And apparently $\dfrac{\partial ETL}{\partial D(\boldsymbol{f}^{A},\boldsymbol{f}^{P})} = 1$, thus from (3.9) and (3.10).

$$\frac{\partial ETL}{\partial \boldsymbol{f}^P[x,y]} = \begin{cases} 0, \text{ if } (x,y) \notin M_{AP} \text{ or } ETL = 0 \\ \dfrac{-2(\boldsymbol{f}^A[x_{b_{AP}}, y] - \boldsymbol{f}^P[x,y])}{N\,|\,M_{AP}\,|}, \text{ otherwise} \end{cases} \tag{3.11}$$

Similarly, for the partial derivatives on the negative feature map, we have:

$$\frac{\partial ETL}{\partial \boldsymbol{f}^N[x,y]} = \begin{cases} 0, \text{ if } (x,y) \notin M_{AN} \text{ or } ETL = 0 \\ \dfrac{2(\boldsymbol{f}^A[x_{b_{AN}}, y] - \boldsymbol{f}^N[x,y])}{N\,|\,M_{AN}\,|}, \text{ otherwise} \end{cases} \tag{3.12}$$

The final step is to calculate the derivatives *w.r.t* the anchor feature map. It can be seen from (3)-(5) that shifting the first map to the left by *b* pixels is equivalent to shifting the second map to the right by *b* pixels. Making use of this property, we have $FD(\boldsymbol{f}^A{}_{b_{AP}}, \boldsymbol{f}^P) = FD(\boldsymbol{f}^A, \boldsymbol{f}^P{}_{-b_{AP}})$ and $FD(\boldsymbol{f}^A{}_{b_{AN}}, \boldsymbol{f}^N) = FD(\boldsymbol{f}^A, \boldsymbol{f}^N{}_{-b_{AN}})$. It is therefore quite straightforward to obtain from (2)-(4):

$$\frac{\partial ETL}{\partial \boldsymbol{f}^A[x,y]} = -\frac{\partial ETL}{\partial \boldsymbol{f}^P[x_{-b_{AP}}, y]} + \frac{\partial ETL}{\partial \boldsymbol{f}^N[x_{-b_{AN}}, y]} \tag{3.13}$$

After calculating the derivative maps *w.r.t* $\boldsymbol{f}^A$, $\boldsymbol{f}^P$ and $\boldsymbol{f}^N$ respectively, the rest of the BP process is the same as for common convolutional neural networks. Above derivation shows that gradients will be computed only for pixels that are not masked. In this way, features are learned only within valid iris regions, while non-iris regions will be ignored since they are not of our interest. After the last convolutional layer, a single-channel feature map is generated which can be used to measure similarities between the iris samples.

3.3.2 Back-propagation for Binary *ETL*

In the case of binary version of *ETL*, the only difference with the real-valued version is the difference computation function (3.7). The step function $H(\bullet)$ generates either zero or undefined gradients and therefore it is infeasible to directly apply back-propagation with gradient descent. This can be connected to an open research problem in the literature, *i.e.*, learning binary feature or Hash codes through deep neural networks. Some existing approaches, such as [125] and [126], typically employed "smooth" versions of the step function to simulate the binarization process in order to achieve compatibility with gradient descent. In this paper, we propose a completely different strategy to address this issue, i.e., instead of simulating the step function with its "soft" versions, we originally interpret the binarization process as a problem of binary classification. Our task is to properly classify each pixel or element in the feature maps $\boldsymbol{f}^A$, $\boldsymbol{f}^P$ and $\boldsymbol{f}^N$ such that the forward loss *ETL* computed from (3.7) is minimized.

Firstly, the forward loss is computed based on (3.2) - (3.5) and (3.7), and then the learning will be casted on triplet samples which generate non-zero *ETL* (thresholded by α). As the forward loss is not able to be back-propagated, we consider the feature binarization as classification for binary case, and then a dependent backward loss is constructed with the widely used logistic (or sometimes called cross-entropy) loss function on each pixel:

$$L_{cls}(f) = -y\log(p) - (1-y)\log(1-p) \qquad (3.14)$$

where $y \in \{0,1\}$ is the latent target label for the current pixel, *p* is the probability of that pixel being class $y = 1$ and is estimated with sigmoid function:

$$p = \sigma(f) = \frac{1}{1+e^{-f}} \qquad (3.15)$$

The key issue is that the correct class label *y* for each pixel is unknown in our case. Fortunately, we can make use of the logical relationship among the anchor, positive and negative samples in the triplet architecture and infer the desired labels which can reduce the forward loss. Assume the feature maps are aligned with (3.4) and a specific pixel position is not masked, the ideal case for that aligned pixel position will be $y^A = y^P \neq y^N$, so that the anchor-positive distance shrinks and anchor-negative distance grows. We therefore assign the following *pseudo-labels* to each pixel in the anchor feature map:

$$\hat{y}^A = \begin{cases} H(f^P), & \text{if } f^A f^P > 0 \\ 1 - H(f^N), & \text{otherwise} \end{cases} \tag{3.16}$$

The motivation of the above assignment is simple, *i.e.*, to strengthen the trend that anchor has the same binary code with the positive, otherwise make it opposite to the negative. Kindly note that the feature values from other branches, i.e., f^P and f^N, are regarded as constants w.r.t the anchor feature f^A. With the pseudo-labels, it is well known that the derivative of the logistic loss function is:

$$\frac{\partial L_{cls}}{\partial f^A} = p^A - \hat{y}^A = \frac{1}{1 + e^{-f^A}} - \hat{y}^A \tag{3.17}$$

With the pseudo-labels and the resulting derivative, the backpropagation can be carried on. Currently in our implementation gradients are only computed for the anchor branch, but the weight updates will be merged to three branch networks at each iteration.

The above optimization strategy does not simulate the step function like other approaches, but incorporates the real binarization step into forward and backward processes by modelling it as a binary classification problem. A backward classification loss which is closely related to the forward target loss *ETL* is then constructed for feature optimization, which can be more effective than numeric simulation that often needs to additionally consider a quantization loss and data distribution priors.

3.4 Feature Encoding and Matching

For the real-valued features output from *UniNet*, We perform a simple encoding scheme. We perform a simple encoding process for the feature map output from *UniNet*. The feature maps originally contain real values, and it is straightforward to measure the fractional Euclidean distance between the masked maps for matching, as

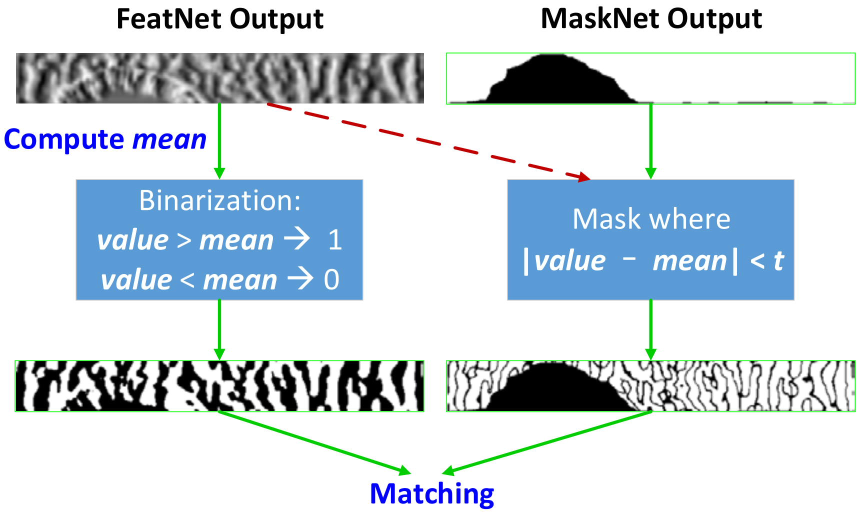


Figure 3.5: Illustration of feature binarization process.

the network is trained in this manner. However, binary features are more popular in most of the research works on iris recognition (*e.g.*, [1, 2, 8-11, 19]), since it is widely accepted by the community that binary features are more resistant to illumination change, blurring and other underlying noise. Besides, binary features consume smaller storage and enable faster matching. Therefore, we also investigated the feasibility of binarizing our features with a reasonable scheme as described in the following:

For each of the output feature map, the mean value of the elements within the non-masked iris regions is firstly computed as m. This mean value is then used as the threshold to binarize the original feature map. In order to avoid marginal errors, elements with feature values v close to m (*i.e.*, $|v-m|<t$) are regarded as less reliable and will be masked together with the original mask output by *MaskNet*. Such a further masking step is conceptually similar to "Fragile Bits" [53], which discovered that some bits in *IrisCode*, with filtered responses near the axes of the complex space, are less consistent or unreliable. The range threshold t for masking unreliable bits is uniformly set to 0.6 for all the experiments. The feature encoding process can be demonstrated in Figure 3.5. For matching, we use the *fractional Hamming distance* [8] from the binarized feature maps and extended masks. It is observed that using the binary

features does not degrade the performance compared with using the real-valued features, and even yield slight improvements in some cross-dataset scenarios, probably due to the factors discussed above.

3.5 Experiments and Results

Thorough experiments were conducted to evaluate the performance of the proposed approach from various aspects. The following sections detail the experimental settings along with the reproducible [78] results.

3.5.1 Databases and Protocols

We employed the following four publicly available databases our experiments:

- **ND-IRIS-0405 Iris Image Dataset (ICE 2006)**

  This database [73] contains 64,980 iris samples from 356 subjects and is one of the most popular iris databases in the literature. The training set for this database is composed of the first 25 left eye images from all the subjects, and the test set consists of first 10 right eye images from all the subjects. The test set, after removing some falsely segmented samples, contains 14,791 genuine pairs and 5,743,130 imposter pairs.

- **CASIA Iris Image Database V4 – distance**

  This database (subset) [27] includes 2,446 samples from 142 subjects. Each sample captures the upper part of face and therefore contain both left and right irises. The images were acquired from 3 meters away. An OpenCV-implemented eye detector [76] was applied to crop the eye regions from the original images. The training set consists of all the right eye images from all the subjects, and the

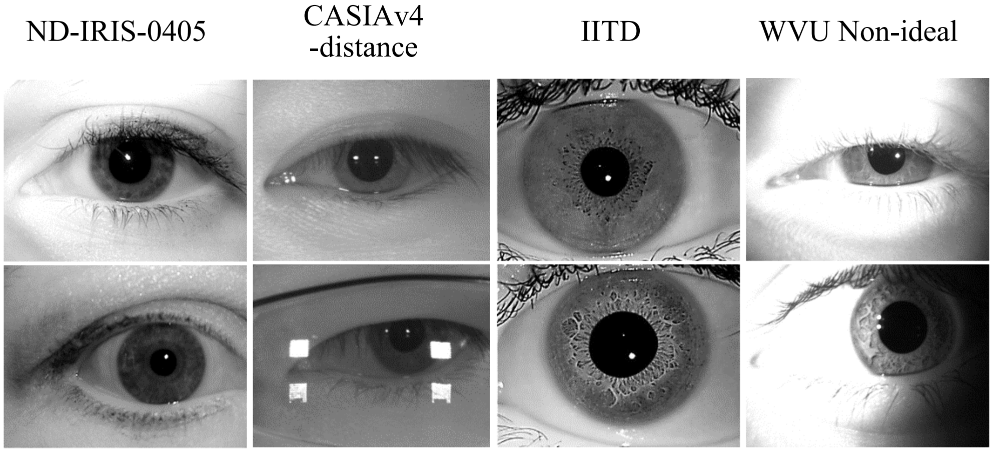


Figure 3.6: Sample raw images from four employed databases.

test set comprises all the left eye images. The test set generates 20,702 genuine pairs and 2,969,533 imposter pairs.

- **IITD Iris Database**

  The IITD database [74] contains 2,240 image samples from 224 subjects. All of the right eye iris images were used as training set while the first five left eye images were used as test set. The test set contains 2,240 genuine pairs and 624,400 imposter pairs.

- **WVU Non-ideal Iris Database – Release 1**

  The WVU Non-ideal database [75] (Rel1 subset) comprises 3,043 iris samples from 231 subjects which were acquired under different extends of off-angle, illumination change, occlusions, etc. The training set consists of all of the right eye images, and the test set was formed by the first five left eye images from all the subjects. The test set has 2,251 genuine pairs and 643,565 imposter pairs.

From the above introduction we can observe that the imaging conditions for these databases are quite different. Sample images from the four employed datasets are provided in Figure 3.6, where noticeable variation in image quality can be observed. It is therefore judicious to assume that these databases can represent diverse deployment environments. The details on the division of the training set and the test

Table 3.2: Summary of the division for training set and test set on the employed databases.

Database	Training Set				Test Set			
	#subjects	samples	side	#images	#subjects	samples	side	#images
ND-IRIS-0405	all 356	first 25	left	9,301	all 356	first 10	right	3,394
CASIA.v4-distance	all 411	first 25	left	6,840	all 411	first 10	right	3,939
IITD	all 224	all	right	1,052	all 224	first 5	left	1,120
WVU Non-ideal	all 231	all	right	1,511	all 231	first 5	left	1,137

set on the four employed databases are provided in Table 3.2. Both the training set and the test set are formed with the first X ($X = 25$, 10 or 5, shown in Table 3.2) or all of the left/right eye images from each of the subjects. If a subject has less than X images in the respective database, then all images from this subject will be included.

During the training phase of our model, the triplet-based architecture introduced in Chapter 3.2 requires the input data to be triplet sets (anchor-positive-negative entries) instead of single images. Therefore the training images in each of the databases need to be presented as triplet entries which are generated from the combinations of images. However, enumerating all the possible triplet combinations in the training set will lead to high storage and computational complexity, we therefore selectively generate part of the possible triplet entries for training, as described in the following: For each training set, we firstly enumerate all the possible anchor-positive (genuine) pairs, since the numbers of available genuine pairs are relatively small; for each anchor-positive pair, we randomly select five negative samples that are from different subjects than the anchor subject, and form the anchor-positive-negative triplet. In other words, each genuine pair in the training set will generate five triplet entries for training.

3.5.2 Test Configurations

We incorporated following two configurations during the test phase for extensive evaluation of the proposed model.

- CrossDB

  In the *CrossDB* configuration, we use the ND-IRIS-0405 as the training set. During testing, the trained model was directly applied on CASIA.v4-distance and IITD *without any further tuning*. The purpose of the *CrossDB* setting is to examine the generalization capability of the proposed framework under challenging scenario that few training samples are available.

- WithinDB

  In this configuration we use the network trained on ND-IRIS-0405 as the initial model, then fine-tune it using the independent training set from the target database. The fine-tuned network is then evaluated on the respective test set. Being capable of learning from data is the key advantage of deep learning, therefore it is judicious to examine the best possible performance from the proposed model by fine-tuning it with some samples from the target database. The fine-tuned models from the *WithinDB* configuration are expected to perform better than the one with *CrossDB*, due to higher consistency of image quality between the training set and test set.

It should be noted that left and right irises are regarded as different subjects in this work, and in both of the above configurations, training set and test set are subject-disjoint, *i.e.*, none of the irises are overlapping between the training set and test set. All the experimental results were generated under all-to-all matching protocol, *i.e.*, the scores of every image pair in the test set have been counted.

3.5.3 Comparison with Earlier Works

We present comparative experimental results using several highly competitive benchmarks. Gabor filter based *IrisCode* [7] has been the most widely deployed iris feature descriptor, largely due to the fact that few alternative iris features in the literature are universally accepted as better than *IrisCodes*. Instead, the majority of recent works on iris biometrics are more on improving segmentation and/or normalization models [51] [11], applying multi-score fusion [26] or feature bits selection [53]. In other words, in the context of iris feature representations, *IrisCode* is still the most popular and highly competitive approach, and therefore is definitely a fair benchmark for the performance evaluation. *IrisCode* has a number of advanced versions. From the publicly available ones, we selected OSIRIS [54], which is an open source tool for iris recognition and its latest version V4.1 was used. It implements a band of multiple tunable 2D Gabor filters that can encode iris patterns at different scales, therefore is a highly credible competitor. Another classic implementation of *IrisCode* is based on 1D log-Gabor filter(s) [8]. Despite the fact that this implementation is considered less competitive nowadays, and is also widely chosen as benchmark in a variety of research works (*e.g.*, [17], [51]). Therefore, this approach is also investigated here. Apart from the Gabor series filters, ordinal filters proposed in [16] can serve as a different type of iris feature extractors to complement the comparisons.

The aforementioned benchmarks have been extensively tuned on target databases during testing to ensure as good performance as possible. We iteratively adopted possible combinations of the parameters for these approaches on each of the training sets within the empirically selected ranges, similar to as in many references (*e.g.*, [8], [16] and [26]). The best performing parameters on the training sets were then employed on the respective test sets for the performance evaluation.

- Parameters for *IrisCode* (OSIRIS 2D Gabor filters):

  A Gabor filter band containing six filters is provided in the original OSIRIS implementation [54]. In addition to the default one, we generated five Gabor filter bands for tuning this tool to obtain the best performance. Based on [7], a 2D Gabor filter for generating *IrisCode* can be formulated as:

$$g(x,y) = e^{-(\frac{x^2}{\alpha^2}+\frac{y^2}{\beta^2})} e^{-i\omega x} \tag{3.18}$$

Each set of parameters (α,β,ω) can be used to produce two filters which are the real and imaginary parts of the complex filter kernel. We apply three sets of parameters to form a band of six filters. The five additional Gabor filter bands are then generated using the following parameters:

$(\alpha,\beta,\omega) \in \{(3, 1.5, 0.4\pi), (5, 1.5, 0.2\pi), (7, 1.5, 0.1\pi)\}$

$(\alpha,\beta,\omega) \in \{(3, 1.5, 0.4\pi), (5, 1.5, 0.3\pi), (7, 1.5, 0.2\pi)\}$

$(\alpha,\beta,\omega) \in \{(5, 2, 0.3\pi), (7, 2, 0.2\pi), (9, 2, 0.1\pi)\}$

$(\alpha,\beta,\omega) \in \{(3, 2, 0.3\pi), (6, 2, 0.2\pi), (9, 2, 0.1\pi)\}$

$(\alpha,\beta,\omega) \in \{(5, 1.5, 0.3\pi), (7, 1.5, 0.2\pi), (9, 1.5, 0.1\pi)\}$

- Parameters for *IrisCode* (1D log-Gabor filter):

  Based on the model presented in [8], two parameters were tuned as follows:

  σ/f (bandwidth over frequency): ranges from 0.3 to 0.6, with a step of 0.05.

  λ (wavelength): ranges from 15 to 40, with a step of 1.

  182 combinations in total.

- Parameters for ordinal filter based method:

  Based on the model presented in [4], four parameters were tuned as follows:

  n (number of lobes): ranges between {2, 3}.

  s (size of each lobe): ranges among {5, 7, 9}.

  d (distance between lobes): ranges among {5, 9, 13, 17}.

  σ (standard deviation of each lobe): ranges among {1.5, 1.7, 1.9}.

  72 combinations in total.

The best parameters automatically selected using the above detailed steps are provided in Table 3.3. It can be observed that such optimal parameters vary for one dataset to another, which underlines the *need* for selecting parameters for conventional

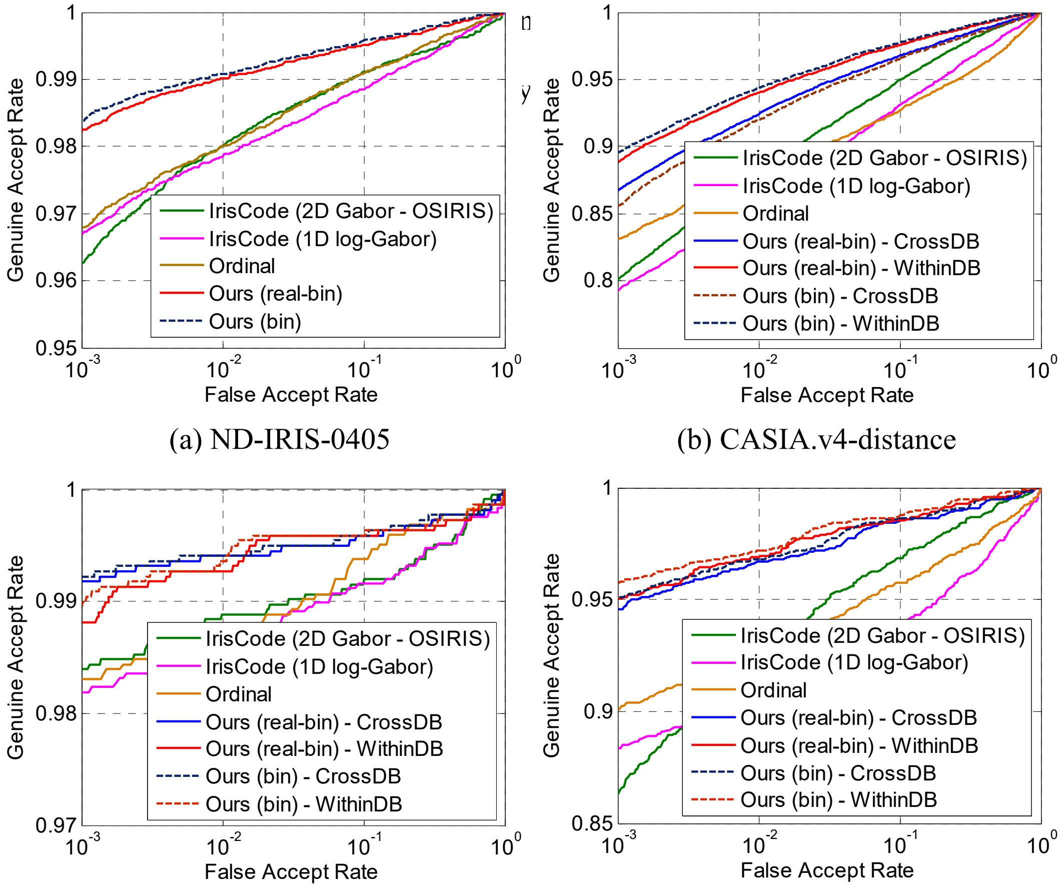
Table 3.3: Best performing parameters for *IrisCode* and Ordinal filters on four employed databases.

Method	Parameter	ND-IRIS-0405	CASIA.v4-distance	IITD	WVU Non-ideal
IrisCode (2D Gabor - OSIRIS)	config.	default	(iii)	(ii)	(i)
IrisCode (1D log-Gabor)	σ/f λ	18 0.45	24 0.35	18 0.4	15 0.55
Ordinal filter	n s d σ	3 5 9 1.9	3 9 13 1.7	2 7 5 1.9	3 9 5 1.9

methods according to the imaging environments and the quality of images for different databases. In contrast, our *CrossDB* model is able to deliver stable and satisfactory performance on the four public databases without any tuning, as would be shown in this chapter later.

It is worth mentioning that we did *not* use the original or built-in iris segmentation/normalization procedures from OSIRIS and Masek's 1D log-Gabor implementation. Iris segmentation has been shown to have significant impact on the recognition accuracy. Therefore to ensure the fairness in the evaluation of proposed iris feature representation, we uniformly adopt our previous proposed method in Chapter 2 [51] for iris detection and normalization (as this method has shown superior results on multiple public databases), and use the output of *MaskNet* as the iris masks for our method and other investigated methods in this part.

The comparison results for recognition are shown in Figure 3.7 and Table 3.4. There are mainly two aspects which can be observed from the results. Firstly, significant and consistent improvements from our method over others have been shown on all of the four databases, under both *WithinDB* and *CrossDB* configurations. Such results suggest that the proposed iris feature representation not only achieves superior accuracy but also exhibits outstanding generalization capability. Even without additional parameter tuning, the well-trained model from our framework is promising



(a) ND-IRIS-0405

(b) CASIA.v4-distance

(c) IITD

(d) WVU Non-ideal

Figure 3.7: ROCs for comparison with other state-of-the-art methods on for employed databases. Best viewed in color.

Table 3.4: Summary of false reject rates (FRR) at **0.1%** false accept rate (FAR) and equal error rates (EER) for the comparison.

	ND-IRIS-0405		**CASIA.v4-distance**		**IITD**		**WVU Non-ideal**	
	FRR	**EER**	**FRR**	**EER**	**FRR**	**EER**	**FRR**	**EER**
IrisCode (OSIRIS)	3.73%	1.70%	19.93%	6.39%	1.61%	1.11%	13.70%	4.43%
IrisCode (log-Gabor)	3.31%	1.88%	20.72%	7.71%	1.81%	1.38%	11.63%	6.82%
Ordinal	3.22%	1.74%	16.93%	7.89%	1.70%	1.25%	9.89%	5.19%
Ours (real-bin) - CrossDB	/	/	13.27%	4.54%	0.82%	0.64%	5.46%	2.83%
Ours (real-bin) - WithinDB	1.78%	0.99%	11.15%	3.85%	1.19%	0.73%	5.00%	2.28%
Ours (bin) - CrossDB	/	/	14.35%	5.06%	**0.77%**	**0.61%**	5.02%	2.69%
Ours (bin) - WithinDB	**1.62%**	**0.93%**	**10.27%**	**3.34%**	1.01%	0.73%	**4.35%**	**2.23%**

life applications. An interesting finding is that on IITD database, the *CrossDB* model performs better even than the fine-tuned one. This is possibly because most of the images in IITD are with high qualities and less challenging, and its training set is not large enough, which causes slight over-fitting problem.

Secondly, in most scenarios the binary iris features which are learned end-to-end yield slight improvements over the real-valued version that has been binarized in ad-hoc manner during test phase. Such results ascertain the effectiveness of the proposed end-to-end binary feature learning scheme, which is promising for addressing the problem of learning to hash with deep neural networks.

3.5.4 Comparison with Other Deep Learning Configurations

In order to ascertain the effectiveness of the proposed network architecture for spatial feature extraction and the extended triplet loss, we also compared our method against typical deep learning architectures that are widely employed in various recognition tasks. The tested configurations are introduced in the following:

(i) *CNN+softmax/triplet loss*

CNN+softmax is the most widely employed deep learning configurations in the community, such as in [58] and [61]. Besides, CNN+triplet loss is gaining increasing popularity after it was proposed in [57], and therefore may also be interesting and worth evaluating. For the CNN model, we have chosen the popular VGG-16 which has achieved superior performance in face recognition.

(ii) *FCN+triplet loss*

Comparative evaluation has also been performed on using the proposed FCN (*FeatNet* only) and the original triplet loss function without incorporating bit-

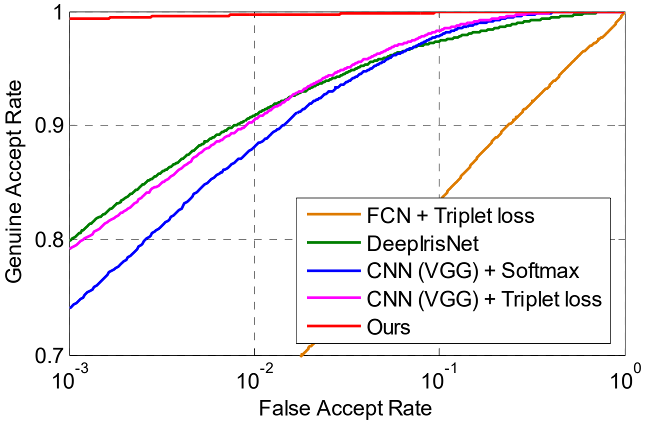


Figure 3.8: ROC curves for typical deep learning architectures available in the literature and our method on ND-IRIS-0405.

shifting and masking. Such comparison may assert the necessity of extending the original triplet loss.

(iii) *DeepIrisNet* [38]

We also compared our method against the recent deep learning based iris recognition framework, DeepIrisNet, which reports promising results. This architecture actually belongs to the CNN+softmax category, but we separately inspected it as it is directly proposed for iris recognition. Since the original model in their paper is not publicly available, we carefully implemented and trained the CNN according to all the details in [38].

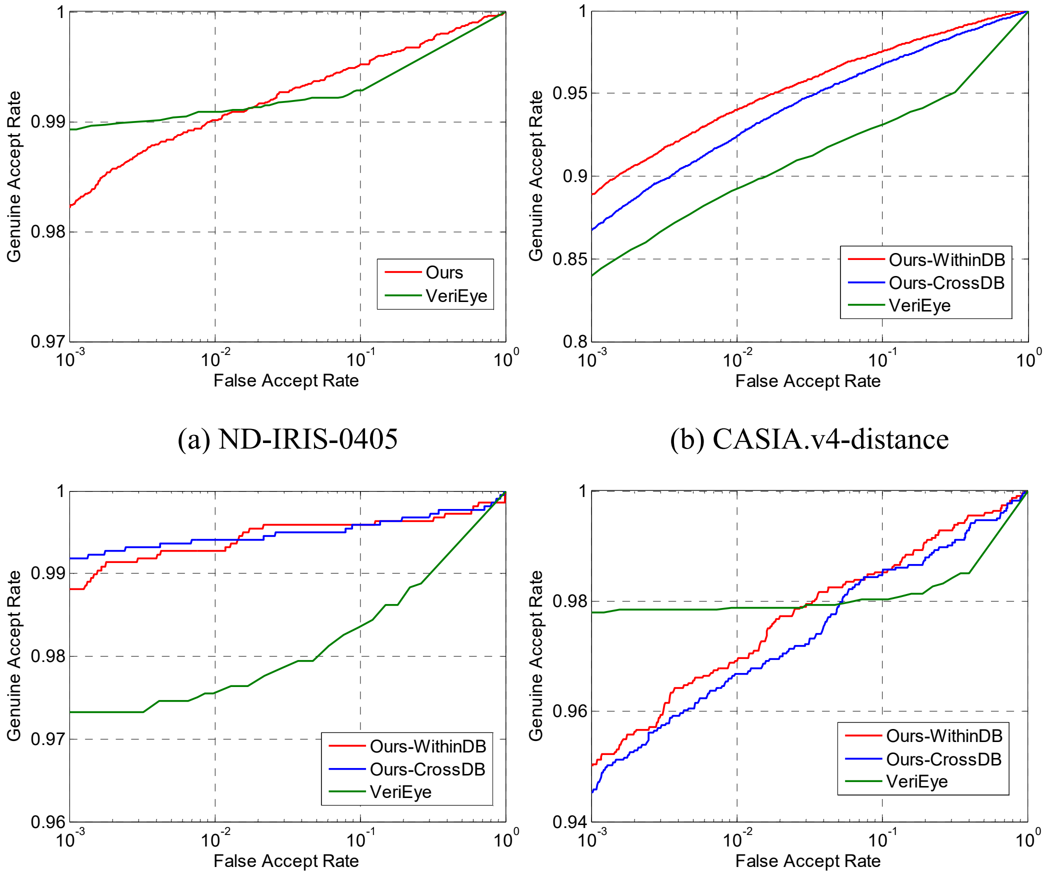
The comparison with aforementioned configurations was performed on ND-IRIS-0405 dataset, which has the largest number of training images among employed ones. The test set is kept consistent during the comparison. Hyper-parameters of the training processes for above architectures have been carefully investigated to achieve best possible performance. The results on the same test set are presented in Figure 3.8.

It can be observed from Figure 3.8 that our newly developed architecture significantly outperforms other deep learning configurations. CNN based configurations have failed to deliver satisfactory results especially at lower FAR. Such results support our previous analysis that global and high-level features extracted by CNN may not be suitable for iris recognition. The poor performance from FCN+triplet loss strongly suggests that it is necessary to account for bit-shifting and non-iris region masking when learning spatially corresponding features through FCN.

3.5.5 Sample Comparison with Commercial System

In the earlier chapter we have provided reproducible performance comparison with the *IrisCode* and ordinal filter based method. Although these methods are widely cited and have shown to offer competitive performance in the literature, it can be interesting to provide comparison with some commercial solutions for iris recognition, as they are considered to be more suitable and optimized for real-life deployment. We therefore performed comparative evaluation using a popular commercial product, VeriEye iris recognition SDK from Neurotechnology [55], which released the latest version 9.0 in October 2016 and is available with us. The VeriEye SDK accepts original eye images (without normalization) as input and has its built-in iris segmentation components. Since this software is not open-source for its core functions, we are not able to describe its iris segmentation process. Therefore, the comparison results presented in this section may not be fully representing the effectiveness of iris feature representation, which is the key focus of this work. Instead, it can be a sample reference for overall performance evaluation. The results for the comparison are shown in Figure 3.9.

As shown in the figure, on ND-IRIS-0405 and WVU Non-ideal databases, VeriEye has better genuine accept rates (GAR) at lower false accept rates (FAR), while our approach consistently outperforms VeriEye on CASIA.v4-distance and IITD datasets. As discussed earlier, the difference in the segmentation process may have certain impact on the final recognition results. Besides, VeryEye has a built-in quality assessment function that it does not match images with low quality, which may



(a) ND-IRIS-0405

(b) CASIA.v4-distance

(c) IITD

(d) WVU Non-ideal

Figure 3.9: ROC curves from our approaches and the commercial product VeriEye SDK on four databases.

improve its overall performance to a certain extent, while our approach does not evaluate image quality at the current stage. Considering above factors, it is judicious to believe that our prototype model can already offer highly competitive performance compared with the well optimized commercial system.

3.5.6 Computational Complexity

The computational complexity of our model has been evaluated in order to address the potential concerns on the feasibility for the deployment. Since our FCN does not employ fully connected layers, the number of parameters is significantly reduced and therefore it is much spatially simpler than conventional CNN based architectures.

Table 3.5: Summary of number of parameters, model storage size and feature extraction time per image, run with Matlab wrapper and C++ implementation, on Intel i7-4770 CPU, 16G RAM and Nvidia GTX670 GPU.

Approach	**#Parameters**	**Model Size (Byte)**	**Feature Extraction Time**	
			GPU	**CPU**
Ours	~ 110.7 K	1.5 M	7.6 ms	236 ms
DeepIrisNet [38]	~55,420 K	289.0 M	12.7 ms	335 ms

Table 3.5 summaries the computational time for feature extraction and the storage required by our model, as compared with the CNN based approach in [38]. It can be noted that the space and time complexities for our approach are quite small.

3.5.7 Sample Evaluation for *MaskNet*

As mentioned earlier, our key focus is on learning more effective iris feature representation. *MaskNet* is an essential component of *UniNet* for providing immediate and appropriate non-iris masking information to the proposed *Extended Triplet Loss* (*ETL*) function. In order to assert the adequateness of the masking information during the feature learning process, we have performed a sample evaluation of *MaskNet*. For the evaluation benchmark, we use our previously proposed iris segmentation framework in Chapter 2 [51] as this method has been published recently and already provided comparison with other promising methods in the literature. Similar to as used in Chapter 2, the average segmentation error is measured using Equation (2.18), *i.e.*, the NICE.I protocol. The difference with [51] is that we measure the segmentation error after iris normalization.

The *MaskNet* employed in our experiments was trained with 500 randomly selected left eye images from ND-IRIS-0405 database, with manually labeled iris masks as the ground truth. The test sets for its evaluation are also generated from the same database, excluding the training samples. We used the following two sets for the testing: (a) 100 randomly selected samples and their ground truth masks manually

Table 3.6: Comparison of average segmentation errors from *MaskNet* and [51].

Approach	**Average Segmentation Error**	
	Set (a)	**Set (b)**
MaskNet	5.89%	9.00%
ICCV'15 [51]	6.73%	11.83%

created by us; (b) 792 samples and their ground truth masks which are available from a public iris segmentation ground truth database, IRISSEG-EP [22]. The average segmentation errors of *MaskNet* and [51] are shown in Table 3.6.

The results shown in Table 3.6 suggest that for both test sets, the developed *MaskNet* can achieve superior segmentation accuracy compared with state-of-the-art iris segmentation approach. It is therefore reasonable to conclude that *MaskNet* is able to provide appropriate information for identifying valid iris region during the feature learning process via *ETL*.

3.6 Summary

This chapter has developed a novel deep learning based iris feature representation which can offer superior matching accuracy and generalization capability for the iris recognition. The specially designed *Extended Triplet Loss* function can provide effective supervision for learning comprehensive and spatially corresponding iris features through the fully convolutional network. Further extension of this work should focus on learning more robust iris mask information through the deep networks, which is expected to further exploit the spatially corresponding features for more accurate iris recognition.

CHAPTER 4

Deep Learning Based Periocular Recognition Using Explicit Semantic Information

4.1 Background

Periocular recognition is an emerging biometric modality that has attracted noticeable interest in recent years and a lot of research effort has been devoted to advance accuracy from the automated algorithms. The periocular region usually refers to the region around the eye, although there is no strict definition or standard from research bodies like NIST [105]. Periocular recognition is believed to be useful when accurate iris recognition cannot be ensured, such as under visible illumination [41], unconstrained environment [44] or when the whole face is not available, as illustrated from some real-life samples in Figure 4.2. It has also been shown that the periocular region is more resistant to expression variations [47] and aging [82] as compared with the face. In addition to serving as an independent biometric modality, periocular information can also be simultaneously combined with iris [10], [84] and/or face [86] to improve the overall recognition performance. However matching periocular images, particularly under less constrained environment, is a challenging problem as this region itself contains less information than the entire face and often accompanied by high intra-class variations along with occlusions like from glasses, hair, *etc*.

In recent years, deep learning techniques, *e.g.,* Convolutional Neural Network (CNN), have gained popularity for their strong ability to extract comprehensive features from the input data, especially for visual patterns. It has demonstrated its robustness to the real-life intra-class spatial variations. The CNN has many successful

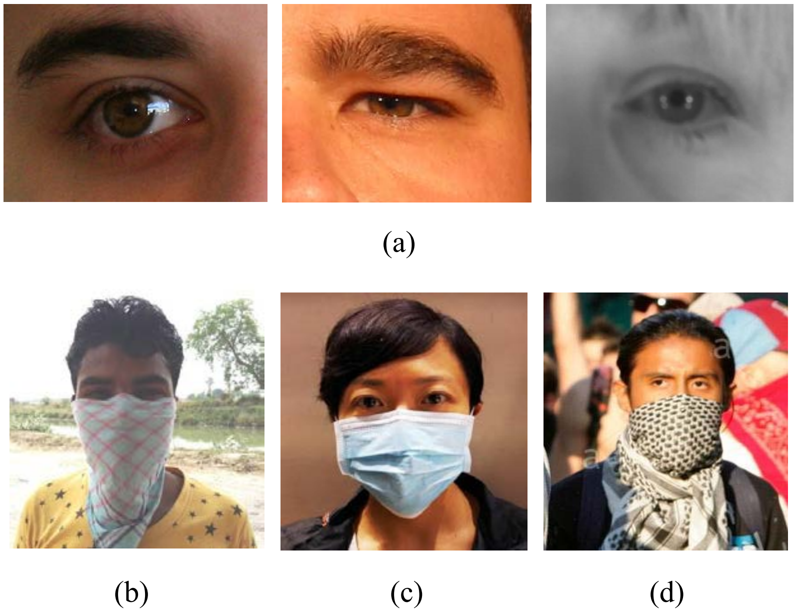


Figure 4.1: Periocular recognition is useful when (a) iris texture is degraded or when the faces are covered for (b) protection from environment, (c) during sickness or (d) during demonstrations or riots [106].

applications like hand-written character recognition [80], object detection [59], large-scale image classification [58] [81] and face recognition [57] [61] [87], where CNN has significantly outperformed traditional methods using handcrafted features or other learning based approaches. Therefore we have been motivated to use CNN to achieve better performance for the challenging periocular recognition problem.

4.2 Our Work

Automated periocular recognition under less constrained environment has shown promising performance and underlined the need for further research. Several databases, acquired under visible and near-infrared illuminations, have been introduced in the public domain [28], [99]-[100] and it can be observed that researchers require/use training samples from respective databases, primarily to select or learn best set of parameters. The performance achieved on these less-constrained databases is encouraging but requires further work. This work attempts to address these two limitations for the automated periocular recognition.

In addition to successfully investigating the strengths of CNN for the less-constrained periocular recognition, this work introduces the Semantics-Assisted CNN (SCNN) architecture to fully exploit the discriminative information within limited number of training samples. The key contributions of our work can be summarized as in the following.

Our approach for periocular recognition using SCNN does not require training samples from target datasets, while achieving outperforming results, which is a key advantage over state-of-the-art approaches [10] and [47]. In our experiments, the SCNN is trained with one database and tested on totally independent/separate databases. The testing and training sets have mutually exclusive subjects and highly different image quality as well as imaging conditions and/or equipment's. The SCNN architecture can also enable recovery of more comprehensive periocular features from the limited training samples. Another key advantage of the proposed method in this work is its computational simplicity, *i.e.*, our trained model requires much less computational time for feature extraction and matching compared with other methods. Unlike earlier works, the trained models and executable files of our work are made publicly available [104] so that other researchers can easily reproduce our results or evaluate on new databases. Finally, the use of SCNN is not only limited to the periocular recognition but can also be useful for general image classification task. By attaching branch CNN(s) that are trained with semantic supervision from the training data, the SCNN architecture can be easily used to extend and improve existing CNN based approaches while limiting the general requirement of increase in training data for such performance improvement. The SCNN enables the deep neural network to fully learn the training data in conjunction with the semantical correlation and therefore can benefit the final classification task, especially when the size of training data is limited to build a very deep network. The structure of SCNN is easy to implement, and semantic annotation of the training samples is often included with the release of many public databases.

The work introduced in this chapter has been published as [48].

4.3 Proposed Methodology

As discussed earlier, we were motivated to incorporate CNN for the challenging periocular recognition problem due to its known ability to extract comprehensive feature from image. In this section we will first introduce the theoretical background of CNN and the practical architecture of our SCNN model in Chapter 4.3.1, followed by detailing the application for the periocular recognition problem in Chapter 4.3.2 and 4.3.3.

4.3.1 Semantics-Assisted Convolution Neural Network (SCNN)

A. Basic Introduction to CNN

CNN is a biologically-inspired variants of multilayer perceptron (MLP) and well-known as one of typical deep learning architectures. CNN has shown strong ability to learn effective feature representation from input data especially for image/video understanding tasks, such as handwritten character recognition [80], large-scale image classification [58] [81], face recognition [57] [61] [87], etc. In the following, we will briefly introduce the basic knowledge of a typical CNN architecture that is used in our and many other work.

CNN is usually composed of convolution layers, pooling layers and fully connected (FC) layers. At the output of each layer, there is often a nonlinear activate function, such as sigmoid, ReLU [79], *etc.* In our work, we adopt the basic CNN

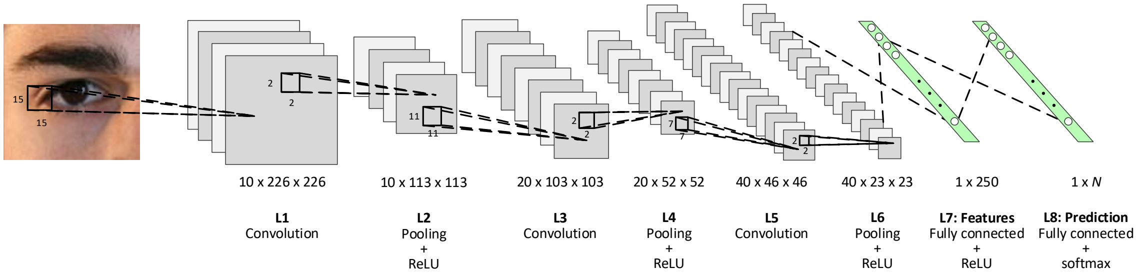


Figure 4.2: Structure of the employed deep convolutional neuron network.

structure similar to AlexNet [81] and is shown in Figure 4.2 (say the periocular recognition problem as an example). The input image is passed through several convolutional units and then a few fully connected layers. The output of the last FC layer with *N* (number of classes) nodes would represent probabilistic prediction to the class labels.

Each of the convolution units is composed of three components - a convolution layer, a max-pooling layer and a non-linear activation function, *e.g.,* ReLU (Rectified Linear Unit), as shown in Figure 4.2. For the convolutional layer, each channel of its output is computed as:

$$\boldsymbol{y}^{(i)} = \sum_{j} (\boldsymbol{b}^{(ij)} + \boldsymbol{k}^{(ij)} * \boldsymbol{x}^{j}) \tag{4.1}$$

where $\boldsymbol{y}^{(i)}$ is the *i*-th channel of the output map, $\boldsymbol{x}^{(j)}$ is the *j*-th channel of the input map, $\boldsymbol{b}^{(ij)}$ is called the bias term, $\boldsymbol{k}^{(ij)}$ is the convolution kernel between $\boldsymbol{y}^{(i)}$ and $\boldsymbol{x}^{(j)}$, and * denotes the 2D convolution operation. $\boldsymbol{b}^{(ij)}$ and $\boldsymbol{k}^{(ij)}$ will be learned by back-propagation so that the convolution kernels are trained to extract most useful features that are discriminative among different subjects.

The pooling layer extracts one maximum or average value from each patch of the input channel. In our application, we use max-pooling with non-overlapping patches. As a result, the input maps, after convolution, are down-sampled with a scale determined by the pooling kernel. The pooling operation aggregates low-level features from the input to high-level representation and thus could achieve spatial invariance among different samples.

At the output of each pooling layer and the first FC layer (e.g., L7 in Figure 4.2), we choose the ReLU (Rectified Linear Unit) [79] as the activation function:

$$y_i' = \max(y_i, 0) \tag{4.2}$$

The ReLU activation ensures the nonlinearity of the feature extraction process and is more efficient for training, compared with the traditional activation functions like sigmoid or tanh employed in other approaches [85].

The FC layers process the input as in conventional neural networks:

$$y_i = b_i + \sum_j x_j \cdot w_{ij} \tag{4.3}$$

where x_j is the *j*-th element of the vectorized input map to the current layer, y_i is the *i*-th element of the output map, which is also a vector. b_i and w_{ij} are elements of the bias and weights to be learned through training. The last FC layer, as usually configured in classification problem, is not followed by ReLU but a *softmax* function:

$$y_i'' = \frac{e^{y_i}}{\sum_j e^{y_j}} \tag{4.4}$$

The use of *softmax* function in the final output of the network results in a $1 \times N$ vector with positive elements which are summed up to one. Each element then is treated as the probabilistic prediction of the class label. The cross-entropy loss function is to be minimized, which is formulated as:

$$L(\boldsymbol{y}'') = -\log y_t'' \tag{4.5}$$

where *t* is the ground truth label of the training sample. The loss function is minimized via back-propagation so that the predictions of the ground truth class of the training samples will approach to unity.

Table 4.1: Examples of several deep learning based approaches and their required number of training images.

Approach	Task	Size of Training Data	
		No. of Classes	No. of Samples
CVPR [58]	Image classification	1,000	1,281,167
ICML [91]	Handwritten digits recognition	10	60,000
T-PAMI [92]	Object detection	200	456,567
CVPR [61]	Face recognition	10,177	202,599
CVPR [87]	Face recognition	4,030	~ 4,400,000

B. Limitation of Contemporary CNN Based Approaches

A common way to achieve superior performance using CNN based methods is to add more layers to make the network deeper and more comprehensive, and/or devote more labeled training data because CNN is usually trained in a supervised manner. For instance, the famous CNN architecture GoogLeNet [58] has 22 layers and later comes the Microsoft's deep network with 152 layers [89]. Apparently, common researchers or companies could hardly afford to train such deep networks due to the lack of enough computational power. Also, as the network goes deeper, the need for training data grows accordingly, while in many research areas, it is difficult to acquire enough labelled training samples like ImageNet [90]. Table 4.1 provides examples of several typical deep learning based approaches and their employed training data. In reference [61] in Table 4.1, for instance, where the developed CNN is not very deep (nine layers), a total of ~200,000 face images from more than 10,000 people were used for training to achieve superior performance. However for other popular biometrics modalities like iris or periocular, in the best of our knowledge, there is currently no single public database with that many images.

Therefore, we are motivated to improve the performance of existing CNN based architecture in another way - to enhance CNN with supervision from explicit semantic information. When human recognizes objects, for example while recognizing a face image, one would analyze not only the overall visual pattern but also the semantic information, such as gender, ethnicity, age, *etc*., to judge whether the face image belongs to a certain known person. Therefore, it is reasonable to believe that semantic information is helpful for the visual identification task. For a CNN that is trained with the identity label only, it is possible that the network is already capable of acquiring semantic information. For instance, for the well-known deep learning model for face recognition, *DeepID2+*, researchers discovered that although the network was trained using subject identities, certain neurons turn out to exhibit selectiveness to attributes like gender, ethnicity, age, *etc.* These semantic attributes contribute to discriminating identities [60]. However, such useful semantic information is expected to be *implicitly* learned by the CNN. It is not easy to answer the following questions:

(1) How many types of semantic information can be acquired? Since the discriminative capacity of a certain CNN is limited, we cannot guarantee that all the semantic information we prefer to have has already been included.

(2) To what extent the semantic information can be analyzed by the trained CNN? Does it really help in the final identification task, or could it be further improved?

Above problems arise due to the nature of training popularly employed for the CNN, *i.e.*, the loss function is usually only related to the class labels, therefore it is hard to reveal how the semantic information can be *implicitly* acquired. In order to address this issue, we propose to empower the CNN with the ability to analyze semantic information *explicitly*. The idea is very simple and illustrated in Figure 4.3.

C. Semantics-Assisted CNN

As illustrated in Figure 4.3, we simply add a branch, which is also a CNN, to the existing CNN. The attached CNN is not trained using the identity of the training data but the semantic groups. For example, we could train CNN2 using the gender

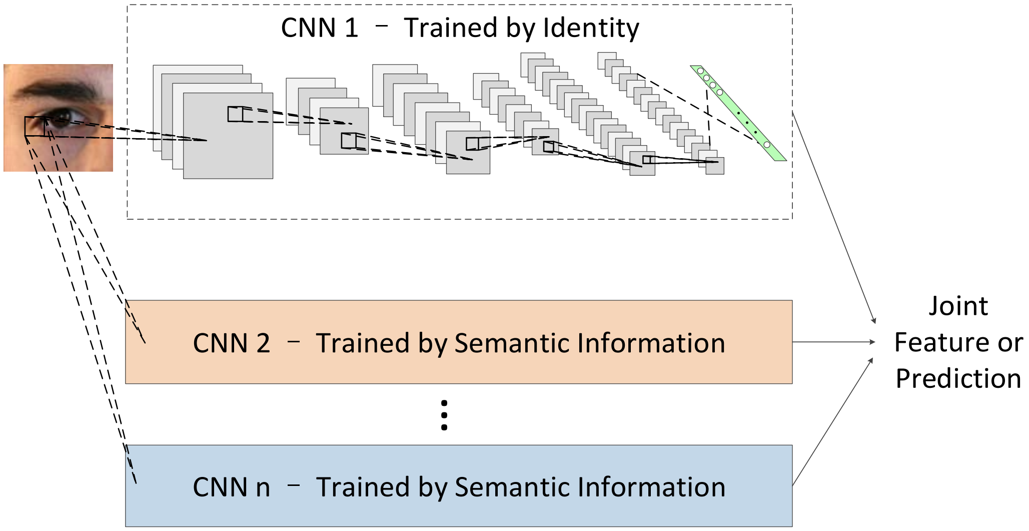

CNN 1 – Trained by Identity
CNN 2 – Trained by Semantic Information
CNN n – Trained by Semantic Information
Joint Feature or Prediction

Figure 4.3: Structure of the proposed Semantics-Assisted CNN (SCNN). While first branch is trained by the label of the intended tasks, other branch CNNs are trained using different semantic information, then the branches are joint in the end to get a comprehensive feature representation or perform score fusion.

information of the training sample, *i.e.*, let the CNN2 be able to estimate the gender instead of identity, and train CNN3 using the ethnicity information. After the CNNs are trained, we can combine the output of each CNN in the way of feature fusion. We refer to such extended structure of the CNN as *Semantics-Assisted CNN* (*SCNN* for short). Despite the simplicity of this idea, it can inherently improve the original CNN by adding more discriminative power to it, which has been shown from the experiments described in Section 3. Theoretically, the SCNN has the following benefits:

- Instead of having the CNN learn the semantic information from the identities in an unpredictable and uncontrolled way, SCNN allows us to *explicitly* recover the preferred semantic information that can be helpful for the identification task. As a result, the feature representation from the SCNN is accompanied by more reliable semantic information that is closer to mechanism in human visual system.

- The training scheme for SCNN can reuse the same set of training data but just labeled in another way than the simple identities. Since the labeling scheme is variable, the branches of SCNN learn the training data from different points of view, which is equivalent to increasing the data volume without really adding the number of training samples. This can relax the constraints on the requirements of enormous training data for deep neural networks to some extent, *i.e.*, instead of pursuing for superior performance from a single CNN, we enhance the joint performance of branches of CNNs with fewer amounts of training data.
- The SCNN architecture and training scheme is naturally compatible for most of the existing CNN based approaches. What we need is just to train some independent CNNs with semantic grouping labels and judiciously combine the features from multiple CNNs to benefit from such training, as the semantic annotations of training samples are also available for many public databases. In addition, the architecture of SCNN is highly friendly for parallel computing platforms.

4.3.2 Application for Periocular Recognition

As discussed earlier, CNN has been successfully used for the face recognition in several state-of-the-art approaches [61] [87]. Considering that the periocular region is actually a part of face and also presents some structural information (eyebrow, eyelids, eyeball, *etc*.), it is reasonable to expect that CNN can be effective for the periocular recognition problem. However, as compared with such related work, we are constrained by lack of large-scale periocular databases that are usually required to sufficiently train a deep neural network. Therefore we developed and investigated SCNN for the periocular recognition problem.

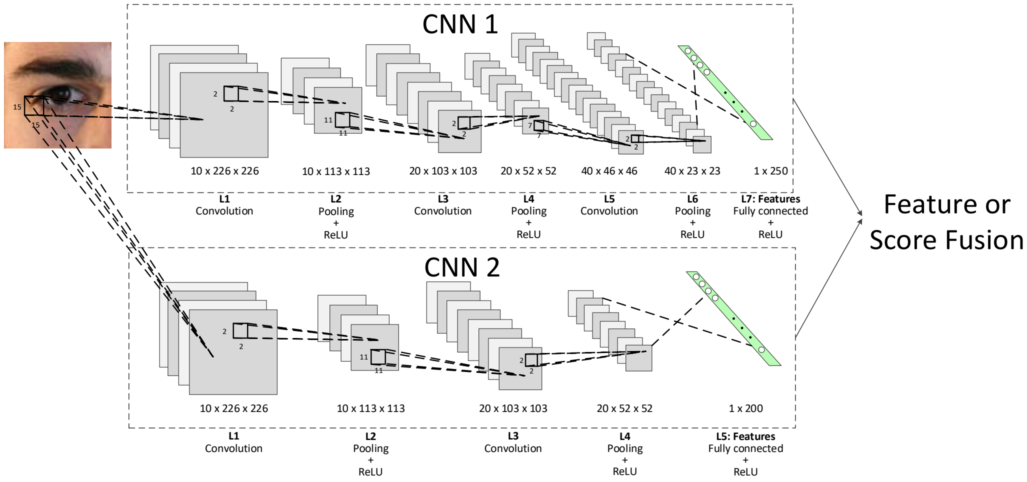


Figure 4.4: Structure of the employed SCNN for the periocular recognition.

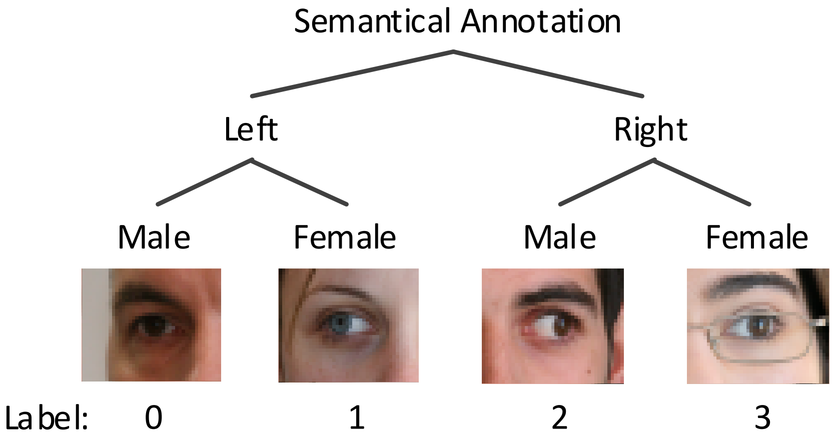


Figure 4.5: Semantical labeling used in our implantation to train CNN2.

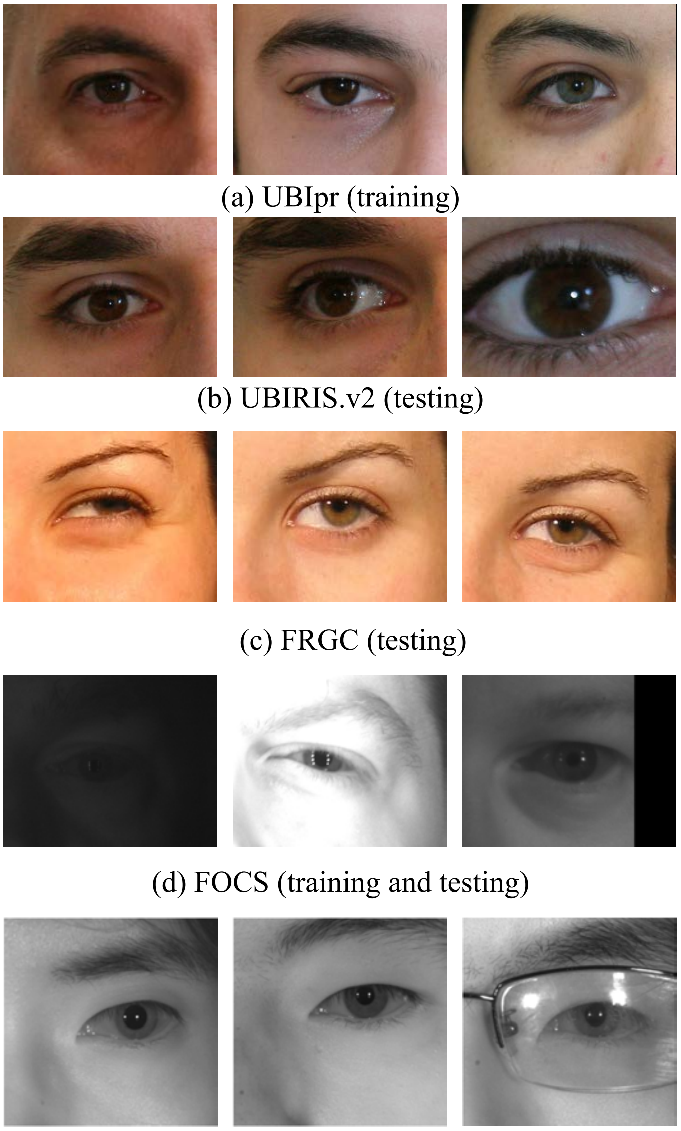
A. Network Structure and Supervision Information

The detailed SCNN structure used for the periocular recognition is shown in Figure 4.4. In order to examine the impact of adding a branch to an existing CNN, we simply designed one branch that is trained with semantic information, denoted as CNN2 in Figure 4.4. While CNN1 is like the ones commonly trained with the subject identities from the training samples, CNN2 is designated to be trained with the side (left or right) and the gender information. More specifically, we labelled the training data as follows, also shown in Figure 4.5:

$$\begin{cases} \textit{0 - Left and Male} \\ \textit{1 - Right and Male} \\ \textit{2 - Left and Female} \\ \textit{3 - Right and Female} \end{cases}$$

The reason for using left/right and gender information is that humans also tend to incorporate such judgment by visually inspecting the presented periocular images, although such accuracy may not reach a hundred percent. Therefore there is some scientific basis to believe that CNN can learn to distinguish above semantic information from the periocular patterns and assist in the identification task. Another reason for using gender information is that the genders of subjects are often included in the metadata of many publicly available datasets, such as UBIpr [100]. Therefore we can directly use those labels to train CNN2. Other possible and useful semantic information include iris color (light/dark), ethnicity, shape of eyebrow, *etc*.

Using such additional semantic information to supervise the network makes the overall architecture and learning process of SCNN similar to multi-label learning [107] to some extent. However, the principal difference is that, the introduction of semantic labeling in our model aims to assist/supplement the prediction of subject identity labels, *i.e.*, they are inequally important, while in traditional multi-label learning, the multiple labels are usually in equal positions. In addition, the learning processes of identities and other semantic information are separately undertaken to maximally ensure the explicitness of semantic learning and compatibility to other CNN based model, while in general multi-label learning, features are usually jointly learned for predicting different lables. Nevertheless, in spite of the diffrentiation between the identity labels and other supportive labels, the semantic learning process (*e.g.*, CNN2 itself) can also be conducted in the manner of multi-label learning alternatively.



(a) UBIpr (training)

(b) UBIRIS.v2 (testing)

(c) FRGC (testing)

(d) FOCS (training and testing)

(e) CASIA.v4-distance (testing)

Figure 4.6: Sample images from the databases we used in the experiments. Scale variance and misalignment are common in the testing environment.

B. Training Protocol and Data Augmentation

Among the original training samples, the last sample of each subject is selected to form the validation set, which is tested in every certain amount of iterations to observe whether the training process is converging in a right direction or not.

Furthermore, it is observed that the periocular images from the training set are well aligned and scaled to a similar level, while the samples from independent test

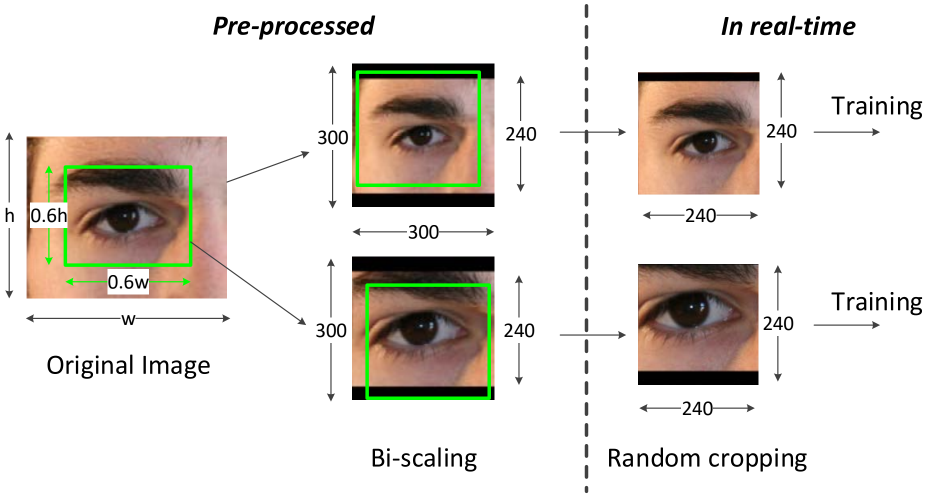


Figure 4.7: Illustration of scale augmentation and random cropping. Each original image is augmented to two samples with different scales, and each augmented sample would be cropped by a smaller window that is randomly placed before entering the network for each epoch of the training process.

datasets and real applications may have misalignments and scale variations. Such inconsistency can also be observed from the image samples in Figure 4.6.

If the deep network is trained with the well aligned and scaled images, it may not be effectively generalized to other datasets or data acquired by real applications. In order to address such problems, we firstly augmented the training data with a different scale to simulate scale inconsistency in the test environment. Then we applied random cropping during the training process to ensure that the network can accommodate spatial variations among the periocular images. The scale augmentation and random cropping process is also illustrated in Figure 4.7. As illustrated in this, each original of the image in training set is automatically cropped from its center with a size of $0.6w \times 0.6h$, where w and h are its original width and height respectively. The original images and its cropped patch are resized to 300×240, then padded with symmetric edges filled with zeroes to a size of 300×300. So far one original periocular image could generate two training samples. As a result, we have 6,270 samples for training and 448 samples for validation while training for each side of the periocular images. Furthermore, during the training process, each training sample would be cropped by a

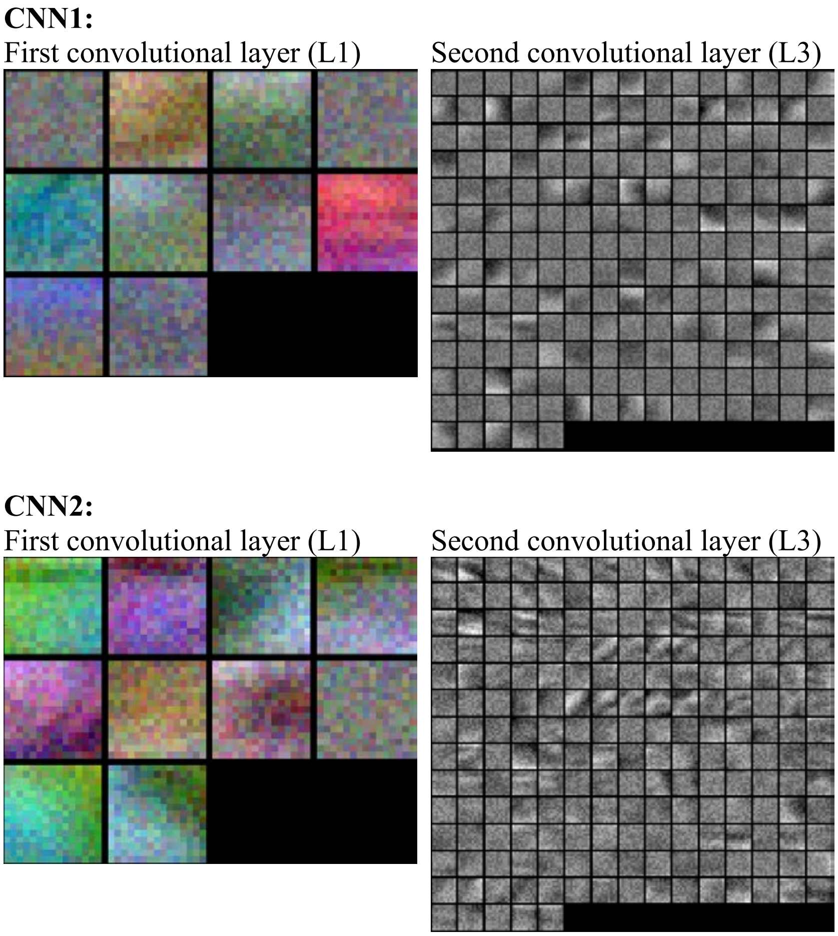

CNN1:
First convolutional layer (L1)
Second convolutional layer (L3)
CNN2:
First convolutional layer (L1)
Second convolutional layer (L3)

Figure 4.8: Visualization of the filter kernels from the first two convolutional layers of trained CNN1 and CNN2 respectively.

240×240 window randomly placed within the image region before entering the first layer of the network. Such randomized cropping process from one training sample could produce abundant samples that have randomized misalignments with others. In this way, the network can be enforced to learn to extract features that are robust to the misalignments.

C. Visualization of Trained SCNN

Once the networks have been trained, CNN1 is expected to lock-into features that are directly relevant to the subject identities, while CNN2 is expected to analyze the features that are more related to side and the gender difference. In order to observe the difference among features extracted by the two CNNs, we have visualized the filter kernels from the first two convolutional layers of trained CNN1 and CNN2 in Figure 4.8.

We can visually observe from Figure 4.8 that: 1) Overall both CNNs were not trained sufficiently. Compared with convolutional kernels trained with large amount of samples (*e.g.*, those in [81]), a number of kernels here remain flat or noisy, for which it is less likely to extract useful information. Insufficiently trained network parameters usually results in certain levels of over-fitting. 2) Despite the over-fitting concern, the convolutional filter kernels of CNN1 and CNN2 are quite different. Critical kernels in CNN2 are sharper and present more visual salience, therefore might be more sensitive to small texture, edges or corners than the filters in CNN1. This indicates CNN2 can provide complementary information that CNN1 was not able to learn due to lack of sufficient training data. Although the features extracted by CNN2 are not directly related to the subject identities, it is reasonable to expect that those visual features could assist CNN1 to form a more comprehensive visual representation of the periocular image, therefore help to distinguish different subjects finally.

4.3.3 Feature Vector and Verification Score Generation

The CNNs we use are trained in a classification protocol, *i.e.*, the category or identity of the input data is known and fixed. Therefore this network can be directly used in some classification or identification tasks. However, in biometrics, one-to-one matching for probably unseen subjects is the key problem and needs to be evaluated. Therefore, we need to generalize the trained model to separated subjects that are not included in the training set, and formulate one-to-one matching scheme.

Similar to [61], we use the output of second last layer (L7 in CNN1 and L5 in CNN2) as the feature representation of the input data. While the last layer represents the class prediction during the training process, the second-to-last layer should contain the most relevant and aggregated information that can contribute to distinguishing the classes or identities. Therefore, it is reasonable to use the output of the second last layer as the feature representation and generalize the model to unseen subjects. Once we get the layer output vectors, we first normalize them by l^2 norm, then apply PCA to reduce the dimensionality of the vector. For the SCNN architecture, we simply concatenate the two independently normalized output vectors to form a longer vector before PCA. In our experiments, the dimension of output vectors after PCA is set to 80, for both the single CNN and SCNN cases. Then the joint Bayesian scheme [98] is utilized to predict the similarity between a pair of feature vectors. The joint Bayesian is primarily designed for face verification, in which a face (equivalent to the periocular feature vector here) is represented by:

$$\boldsymbol{f} = \boldsymbol{\mu} + \boldsymbol{\varepsilon} \tag{4.6}$$

where $\boldsymbol{f}$ is the observation, in this work the feature vector after PCA, $\boldsymbol{\mu}$ is the identity of the subject, $\boldsymbol{\varepsilon}$ is the intra-class variation. $\boldsymbol{\mu}$ and $\boldsymbol{\varepsilon}$ are assumed to be two independent Gaussian variables following $\mathcal{N}(\mathbf{0}, \boldsymbol{S}_{\mu})$ and $\mathcal{N}(\mathbf{0}, \boldsymbol{S}_{\varepsilon})$ respectively, then the covariance of two observation is:

$$\mathrm{cov}(\boldsymbol{f}_1, \boldsymbol{f}_2) = \mathrm{cov}(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2) + \mathrm{cov}(\boldsymbol{\varepsilon}_1, \boldsymbol{\varepsilon}_2) \tag{4.7}$$

The joint distribution of a pair of observations $\{\boldsymbol{f}_1, \boldsymbol{f}_2\}$ is considered. Let H_I denote the intra-person hypothesis indicating that two observations are from the same person, and H_E the extra-person hypothesis. Under H_I, since $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ are the same, $\boldsymbol{\varepsilon}_1$ and $\boldsymbol{\varepsilon}_2$ are independent, the covariance matrix of the distribution $P(\boldsymbol{f}_1;\boldsymbol{f}_2 \mid H_I)$ is:

$$\boldsymbol{\Sigma}_I = \begin{bmatrix} \boldsymbol{S}_{\mu} + \boldsymbol{S}_{\varepsilon} & \boldsymbol{S}_{\mu} \\ \boldsymbol{S}_{\mu} & \boldsymbol{S}_{\mu} + \boldsymbol{S}_{\varepsilon} \end{bmatrix} \tag{4.8}$$

On the other hand, under H_E, $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ are also independent, therefore the covariance matrix has become:

$$\boldsymbol{\Sigma}_E = \begin{bmatrix} \boldsymbol{S}_\mu + \boldsymbol{S}_\varepsilon & \mathbf{0} \\ \mathbf{0} & \boldsymbol{S}_\mu + \boldsymbol{S}_\varepsilon \end{bmatrix} \tag{4.9}$$

With above conditional joint probabilities, the log likelihood ratio which tells the difference between intra- and extra-person probabilities can be obtained in a closed form:

$$r(\boldsymbol{f}_1, \boldsymbol{f}_2) = \frac{P(\boldsymbol{f}_1, \boldsymbol{f}_2 \mid H_I)}{P(\boldsymbol{f}_1, \boldsymbol{f}_2 \mid H_E)} = \boldsymbol{f}_1^{\mathrm{T}} \boldsymbol{A} \boldsymbol{f}_1 + \boldsymbol{f}_2^{\mathrm{T}} \boldsymbol{A} \boldsymbol{f}_2 - 2 \boldsymbol{f}_1^{\mathrm{T}} \boldsymbol{G} \boldsymbol{f}_2 \tag{4.10}$$

where

$$\boldsymbol{A} = (\boldsymbol{S}_\mu + \boldsymbol{S}_\varepsilon) - (\boldsymbol{F} + \boldsymbol{G}) \tag{4.11}$$

$$\begin{bmatrix} \boldsymbol{F} + \boldsymbol{G} & \boldsymbol{G} \\ \boldsymbol{G} & \boldsymbol{F} + \boldsymbol{G} \end{bmatrix} = \begin{bmatrix} \boldsymbol{S}_\mu + \boldsymbol{S}_\varepsilon & \boldsymbol{S}_\mu \\ \boldsymbol{S}_\mu & \boldsymbol{S}_\mu + \boldsymbol{S}_\varepsilon \end{bmatrix}^{-1} \tag{4.12}$$

The covariance matrix $\boldsymbol{S}_\mu$ and $\boldsymbol{S}_\varepsilon$ can be estimated using an EM based algorithm as detailed in [98], and the log likelihood ratio $r(\boldsymbol{f}_1, \boldsymbol{f}_2)$ is used as the similarity score in our one-to-one matching scenario.

4.4 Experiments and Results

In this chapter we provide the details on the experiments and analyze the results. The experimental details on the periocular identification are firstly provided and this is followed by details on supporting experiments for the image classification.

4.4.1 Periocular Recognition

A. Training and Testing Datasets and Protocol

We use following publicly available databases for the experiments. Two different databases were employed for training the deep neural networks and three separate databases were employed for the testing.

- UBIpr [100] - *for training*

  We employed UBIpr periocular database for training the SCNN for the visible spectrum. This database originally contains 5,126 images for each of left and right periocular from 344 subjects. However, we are also employing a subset of UBIRIS.v2 database [18] for separate test experiments, which has some overlapping subjects with the UBIpr database. In order to ensure that subjects of training set and testing set are mutually exclusive, we removed these overlapping subjects from UBIpr database before we perform training on the network. As a result, we only have 3359 periocular images from each of the two sides of 224 subjects. Such a scale is relatively small as compared with those in the training protocols in other typical deep learning work like ImageNet [93] or LFW [94]. Therefore, the application scenario is good for validating the ability of SCNN for learning comprehensive information from limited size of training data.

- UBIRIS.v2 [18]

  The UBIRIS.v2 database is primarily released for evaluation of at-a-distance iris segmentation and recognition algorithms under visible illumination and challenging imaging environment. Since the eye images in this database contain surrounding regions of the eye, it is possible to perform periocular recognition on the UBIRIS.v2 database. Similar to as in [10], we use a subset of 1,000 images from this database that is released in NICE.I competition [21]. This subset contains left and right eye images together from 161 subjects that are captured from 3m to 8m, bringing serious scale inconsistency. Some images only contain the eye region without eyebrow and other surrounding texture which makes the task of periocular recognition highly challenging. Some sample images are shown in Figure 4.6 (b).

- FRGC [28]

  The dataset of Face Recognition Grand Challenge (FRGC) is released by the National Institute of Standards and Technology (NIST) and has been primarily for the evaluation of new algorithms for the automated face recognition. Similar to as in [10], we automatically extracted the periocular region from the original face images of FRGC using publicly available face and eye detector [96]-[97]. A subset of 540 right eye images from 163 subjects, same as also the ones used in [10], were employed in the experiments. Some sample images are reproduced in Figure 4.6 (c).

- FOCS [99] - *for training and testing*

  The Face and Ocular Challenge Series (FOCS) dataset is also released by NIST and contains face, ocular images and videos. We employed the "OcularStillChallenge1" section, which consists of 4,792 left and 4,789 right periocular images from 136 subjects that are cropped from face video clips acquired under near-infrared (NIR) spectrum. The periocular samples from this dataset, as shown in Figure 4.6, suffer from serious illumination inconsistency and misalignments, therefore this dataset is considered as highly challenging. We used 3,262 left and 3,259 right periocular images of the first 80 subjects to train the CNNs and used the remaining images from 56 subjects for testing. Again, such a scale of training samples and subjects is small compared with other typical deep learning tasks.

- CASIA.v4-distance [27]

  CASIA.v4 is the first publicly available long-range iris and face database acquired under NIR illumination, which is released by the Center for Biometrics and Security Research (CBSR) from the Chinese Academy of Sciences (CASIA). The full database contains 2,567 images from 142 subjects in single session. The

Table 4.2: Summary of the employed databases for training and testing.

Spectrum	**Visible**			**Near Infrared (NIR)**		
Division	**Train**	**Test**		**Train**	**Test**	
Dataset	UBIpr	UBIRIS.v2	FRGC	FOCS	FOCS	CASIA.v4-distance
Standoff distance	4 – 8m	3 - 8m	N/A	N/A	N/A	⩾3m
No. of subjects*	224	171 (19/152)	163 (13/150)	80	56	141 (10/131)
No. of images*	left: 3,359 right: 3,359	1,000 (96/904)	540 (40/500)	left: 3,262 right: 3,259	1,530	1,077 (79/998)

* In the bracket (*a*/*b*) means *a* subjects or images were used for training for methods [10] and [47] (not for our method), remaining *b* subjects or images were used for testing.

standoff distance of the subjects to the camera is from 3 meters away. Similar to FRGC, we used publicly available eye detector [96]-[97] to automatically segment left periocular images which are used in our experiments. The first eight samples of each subject, excluding a few badly segmented images, were used for the periocular matching experiment.

Above datasets were selected for evaluation because of the availability of periocular images acquired under less constrained environments that are close to real world scenarios. The selected subsets from FRGC and UBIRIS.v2 contain multi-session data and exhibit obvious scale/illumunation variation. Samples in FOCS database suffer from significant illumination degradation and misalignment. Images from CASIA.v4-distance are more consistent than the other three databases, but were acquired at a distance and some contain artifacts like glasses and/or hair, therefore also represent less constrained scenarios. In addition, networks for visible and NIR spectrums were trained separately due to the significant difference between the image properties.

It is important to clarify that during our (reproducible [104]) experiments, the SCNN is tested in totally *cross-database* manner, *i.e.*, not only the subjects from the training and test set sets are totally separated, the databases themselves are independent from training for *three sets* of experiments. However, the methods we are going to compare with, [10] and [47], both require some samples of the target databases for the training. In order to compare with the best performance of [10] and [47] as well as to ensure the fairness in such comparison, we still divide the target datasets into training and testing sets, as summarized in Table 4.2. For example, 96 samples of the first 19 subjects in UBIRIS.v2 were used to train the models [10] and [47], the remaining were used for test as in [10], [47] and also for our method. Such a configuration is highly disadvantageous to our methods because the inter-database variance is always a key factor for the performance of all learning based methods. However, our method has still been able to achieve outperforming results as detailed later.

We perform periocular matching using the all-to-all protocol, *i.e.*, every image is matched to all the other images in the testing set, and all the generated matching scores are taken into calculation of the receiver operating characteristic (ROC) curve. Such a protocol is considered to be highly challenging because one bad sample may result in several poor genuine scores, which drops the overall matching performance.

B. Effectiveness of SCNN

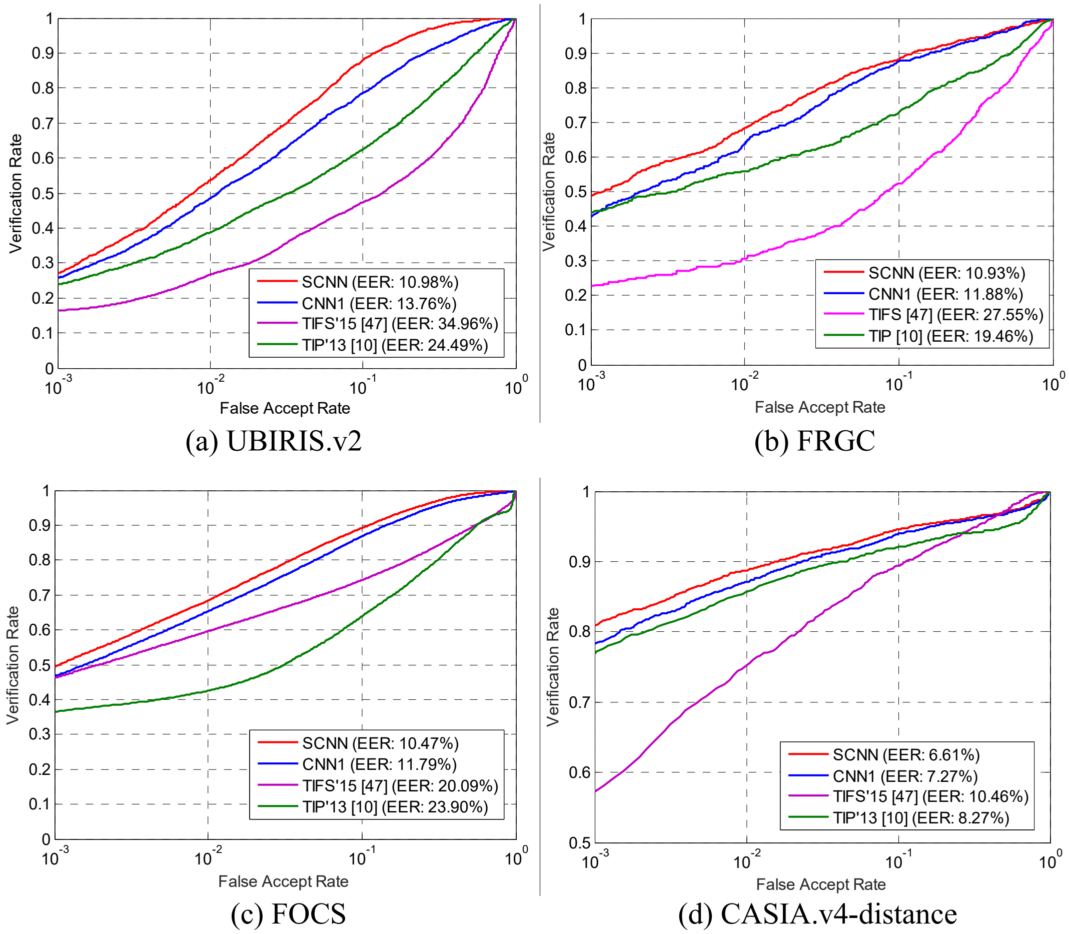
We firstly examine the impact of the added branch that has been trained with the semantic information. We have compared the performance of a single CNN, *i.e.*, only CNN1 in Figure 4.3, with the performance of the extended SCNN. The results from the verification experiments are illustrated in Figure 4.9.

We can observe from Figure 4.9 that the SCNN consistently achieves better performance than that of original or single CNN. This observation suggests that the newly added CNN2 which is trained with semantic supervision has been successful in contributing to some useful information that is not reinforced in CNN1, and therefore improving the overall discriminative power of the network. In theory, we can add more branches that are trained with different semantic information (*e.g.*, iris color) to further improve the final recognition accuracy. However, the need for computational power would also increase and the trade-off may need to be made according to the applications. In our example, since CNN2 shown in Figure 4.4 has a relatively simplified structure, the additional training cost is minor.

C. Comparison with Earlier Work on Periocular Recognition

We also compared the performance of our approach with state-of-the-art approaches [10], [47] on the periocular recognition problem. While [10] is our previous work, we have carefully implemented the methods in [47] with the help of the original authors. The test protocols were kept exactly the same for different approaches during the experimental process and therefore the comparisons of ROC/CMC curves are fair. However several factors can be firstly clarified here to ensure clarity in understanding the experimental comparisons.

1) For UBIRIS.v2, we use the 1,000 image set that was employed for the NICE.I competition. This subset is the same as was used in [10] but different from the one in [47]. In [47], test images were gathered from the full dataset, but only those acquired from 6-8 meters were used, while the 1,000 image set in [10] included samples acquired from 3-8 meters. Due to the relatively consistent imaging distance, the subset used in [47] involves much less scale variance than those in [10] and also in this work. As a result, the performance from our experiment using exact method in [47] is not reproduced as good as what appears to be in [47] and this is reasonable due to the difference in selection of images as explained above.
2) For FRGC, we also used the same subset as in [10] but different from the one used in [47]. As described before, the subset we used contains 540 periocular images which were *automatically* segmented from the original face images and therefore may suffer from some misalignment. Moreover, images in this subset were acquired from *various sessions* with certain time lapse and different imaging



(a) UBIRIS.v2

(b) FRGC

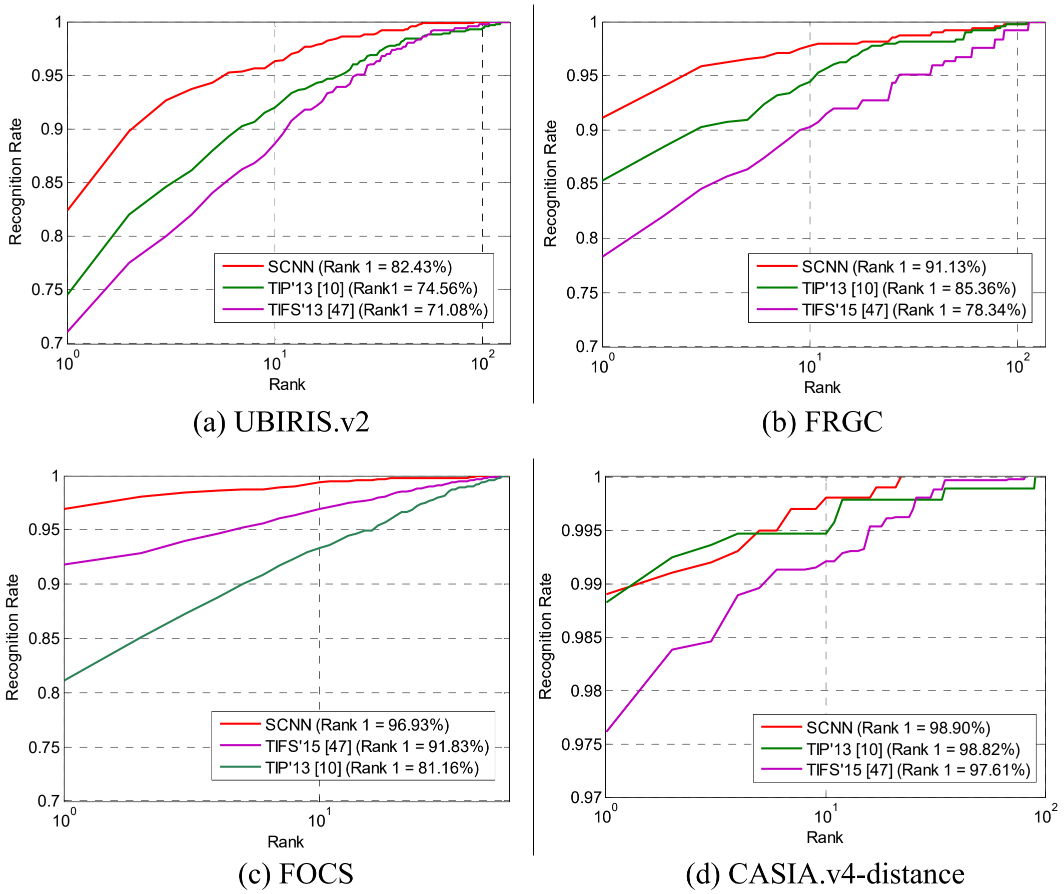
(c) FOCS

(d) CASIA.v4-distance

Figure 4.9: ROC curves of the periocular verification using SCNN and comparison with single CNN and other state-of-the-art methods for different databases.

environments, which increases the difficulty for accurate recognition. However, the subset used in [47] only consists of images captured in consistent illumination and background in *single session*, and the periocular regions were *manually* segmented. Therefore, it is also a reasonable explanation for the drop in performance in our reproduced results, over the ones shown in [47] using manual segmentation.

3) For FOCS, we used fixed division of training and testing sets as shown in Table 4.2, while the original setup in [47] used 5-fold cross validation for the entire dataset. Although the subsets used in our experiment and their original experiment



(a) UBIRIS.v2

(b) FRGC

(c) FOCS

(d) CASIA.v4-distance

Figure 4.10: CMC curves of the periocular verification using SCNN and comparison with state-of-the-art methods for different databases.

are not exactly the same, the quality of images is observed to be quite similar. Therefore our reproduced result is very close to those appearing in reference [47].

The verification results (ROC) for above comparisons are also shown in Figure 4.9, while the identification results (CMC) are shown in Figure 4.10. It can be observed from the experimental results in these two figures that the proposed approach using SCNN consistently outperforms the two state-of-the-art approaches.

In order to ascertain statistical significance of the improvements, we have conducted the significance test for the ROC curves using the method described in [83], which judges from the area under the curve (AUC). Table 4.3 shows the significance level (*p*-value) of the difference of the SCNN based method over the comparative

Table 4.3: Results of significance test for comparison of ROCs using method [83]. *p*-value indicates the probability of the null hypothesis that two methods have no difference statistically.

Comparison	***p*-value***			
	UBIRIS.v2	**FRGC**	**CASIA.v4-distance**	**FOCS**
SCNN & TIP'13 [10]	< 1e-4	< 1e-4	< 1e-4	< 1e-4
SCNN & TIFS'15 [47]	< 1e-4	< 1e-4	< 1e-4	< 1e-4

* The computed *z*-statistics are too large that the corresponding *p*-values exceed double precision, therefore expressed as < 1e-4.

Table 4.4: Comparison of time required to match two periocular images by different approaches, from Matlab implementation running on a computer with Linux OS, 16 GB RAM, 3.4 GHz Intel i7-4770 CPU (4 cores) and NVIDIA GeForce GTX 670 GPU.

Approach	**Major Time Consuming Operations**	**Matching Time (s)**	
		GPU	**CPU**
proposed	convolution, matrix multiplication	**0.013**	**0.183**
TIP'13 [10]	DSIFT feature extraction, K-means clustering	/	15.478
TIFS'15 [47]	Gabor feature extraction, correlation filter matching	/	1.441

methods [10] and [47]. The results indicate that, by the commonly used confidence level of 95%, our approach significantly outperforms these two methods (p-value < 0.05) on all the employed datasets.

It may be noted that [47] performed poorly on the UBIRIS.v2 set because it adopts the patch based matching scheme while, as explained above, the 1,000-image set of UBIRIS.v2 used in our experiment suffers from serious scale variations among the samples, which results in significant loss of patch correspondence. The approach from [10] which uses DSIFT features is more robust to scale variance, however the extraction of DSIFT feature is especially time consuming. In contrast, our approach not only performs better than both of the baseline approaches on different databases, but is also computationally simpler for the deployment using the trained network. Table 4.4 presents the summary of the average time required for the feature extraction for the considered state-of-art approaches. These tests were performed using the Matlab wrapper and C++ implementation running on a computer with Linux OS, 16 GB RAM, 3.4 GHz Intel® Core™ i7-4770 CPU (4 cores) and NVIDIA® GeForce GTX 670 GPU. It can be observed that the proposed approach is much faster due to the straightforward architecture and the use of GPU could further reduce the computational time.

4.4.2 Image Classification

In order to examine that the proposed SCNN architecture is not only effective for the periocular recognition but can also be useful for more general problems, we performed experiment for image classification on the CIFAR-10 dataset [101].

The CIFAR-10 dataset contains 60,000 32×32 color images from 10 classes. Among these images, 50,000 images are for training and 10,000 are for testing. Figure 4.11 shows some randomly selected samples from each class. As we can see from Figure 4.11, although the number of classes is not large, the intra-class variation is significant and the resolution is also smaller, which brings certain challenge for classifying those images. The CIFAR-10 has therefore emerged as a popular dataset for evaluating image classification algorithms along with others like ImageNet and CIFAR-100, *etc*.

Since the SCNN is developed to enhance existing CNN based approaches, we select a baseline CNN to ascertain the improvement. We adopt the CNN originated from Krizhevsky's cuda-convnet [102], re-implemented and introduced in the Caffe tutorial [103]. Although the selected CNN is not the state-of-the-art for CIFAR-10 in terms of performance, we chose it because this model is publicly available under Caffe, the deep learning framework employed in the work, and it is also quick to train. For

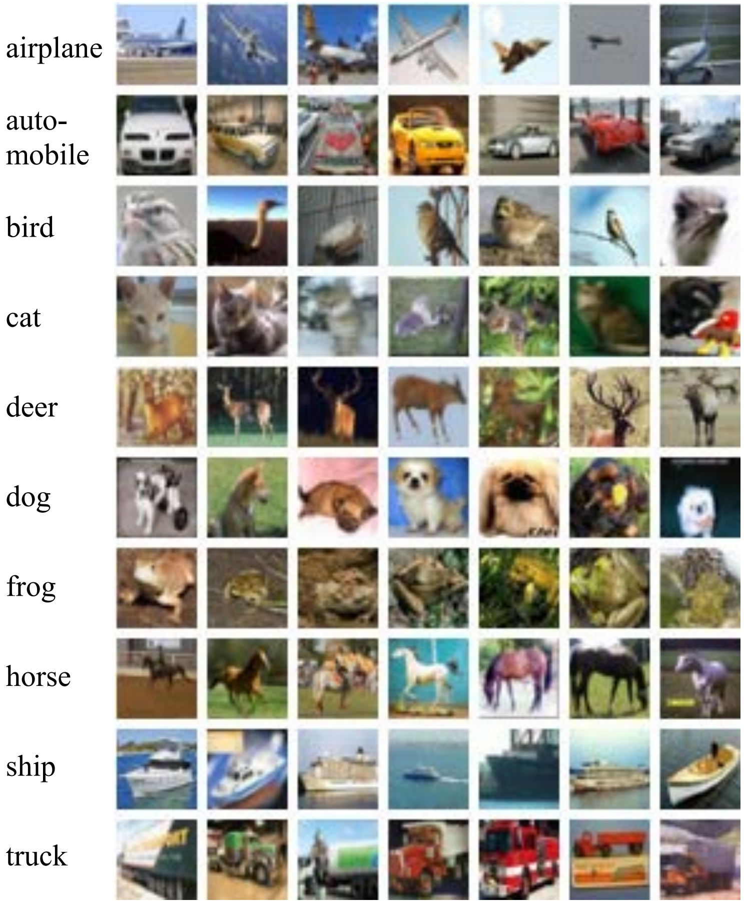


Figure 4.11: Sample images from each class of CIFAR-10 dataset.

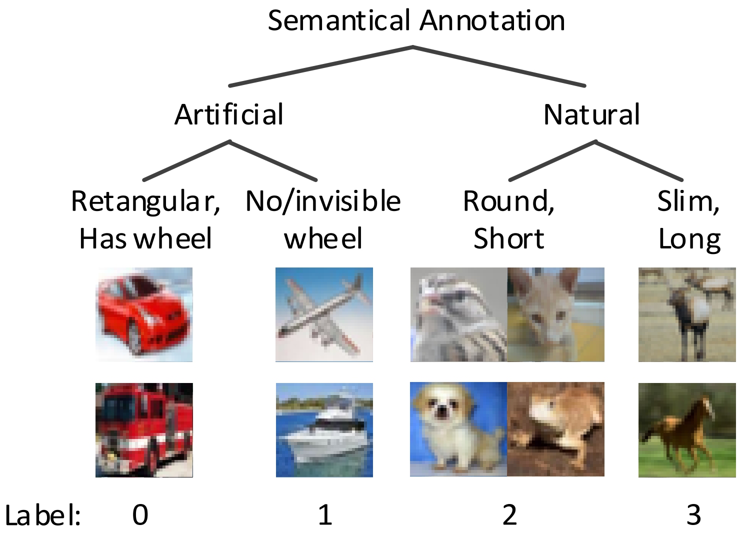


Figure 4.12: The semantical group labelling used in our experiment to train the cuda-convnet-s.

simple annotation, we refer to this network as *cuda-convnet*. By following the tutorial, we can quickly get an accuracy of about 75% on the CIFAR-10 test set. Then we trained a branch CNN to learn the semantic features of the images in CIFAR-10 in order to build the SCNN architecture. We define one possible grouping of semantic information for the classes in the CIFAR-10 dataset as follows, also shown in Figure 4.12.

- artificial
  - rectangular, has wheel: (automobile, truck)
  - no/invisible wheel: (airplane, ship)
- natural
  - round, short: (cat, dog, bird, frog)
  - slim, long: (deer, horse)

With above division, the entire dataset is grouped into four semantical classes. It may be noted that this is not the unique or the optimal division, but it is an easy-to-understand scheme to start with. In order to obtain a branch CNN that was trained to acquire above semantic features, we simply duplicate the structure of the base cuda-convnet but replace the last fully connected layer having 10 neurons with a new fully connected layer with four neurons, since the task now is to recognize the four semantic groups. We then just repeat, as described in Caffe tutorial, but train the new network with newly labeled data. We refer to this new CNN as cuda-convnet-s. Again, above configuration is made because of the ease to execute and one has many choices for actual applications. We then built an SCNN with the architecture as in Figure 4.13. As shown in this figure, we combine the branch CNN and the original one to obtain an extended structure. The components highlighted in red are retrained after the combination to aggregate the long concatenated features, and this process can be considered as a kind of finetuning. Since the number of layers to be retrained is small, the finetuning is very fast. Table 4.5 shows the classification results on the test set using the original cuda-convnet and the extended SCNN.

We can observe from the results that the proposed SCNN can achieve an

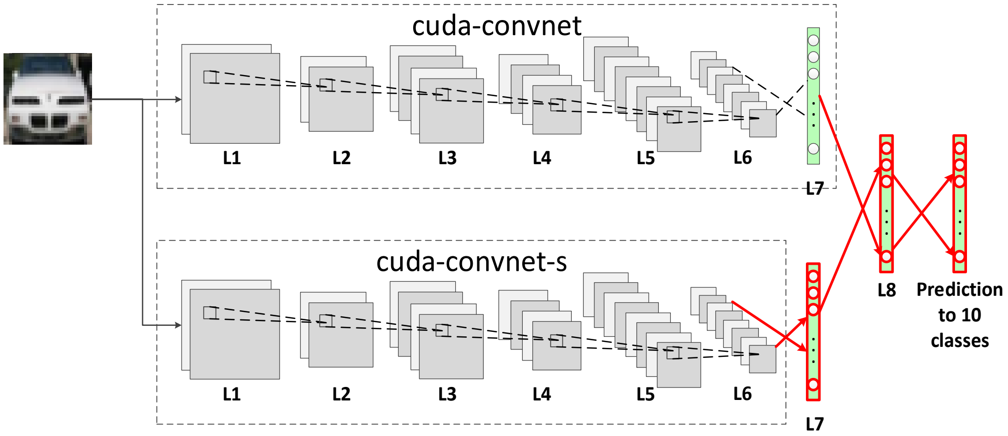


Figure 4.13: The structure of SCNN used in the experiment for CIFAR-10 dataset. The *cuda-convnet* is from the original Caffe tutorial, and the *cuda-convnet*-s is newly trained by the semantic information.

Table 4.5: Results of classification on the CIFAR-10 testing set using original existing *cuda-convnet* and the proposed SCNN enhancement on the *cuda-convnet*.

Approach	Accuracy
cuda-convnet	74.95%
cuda-convnet-SCNN	77.06%

improvement of 2.11% over the original result. Although this may not be considered as a very large improvement, the achieved results reinforce the motivation for SCNN is to make solid and consistent enhancement on existing CNN based approaches, especially for the scenario when the training data may not be enough to feed a complex network. In the CIFAR-10 dataset, the number of images per class is actually quite large and therefore the effect of SCNN is not significant, but it still offers a noticable improvement with minor addition in the complexity. Moreover, as discussd above, the experimental setup is reproducible and made to execute in a straightforward manner. Therefore it is reasonable to expect certain space for further improvement.

4.5 Summary

This chapter has presented automated periocular recognition using CNN with outperforming results and significantly smaller complexity. In particular, we proposed a robust and more accurate framework for the periocular recognition using the semantics-assisted convolutional neural network (SCNN). By training one or more branches of CNNs with semantical information corresponding to training data, the SCNN is capable of recovering more comprehensive features from the images and therefore achieve superior performance. Our experimental results on four publicly available databases suggest that the proposed approach can achieve outperforming results while requiring much smaller computational time for the matching process. The SCNN architecture can also be generalized for other image classification tasks, which can improve the performance over the single CNN based approaches. The source and executable files of our approach are made publicly available [104] to encourage other researchers to easily reproduce our results and further advance research on accurate periocular recognition.

It may be noted that at the current stage, we decouple the identity supervision and other semantic supervision, in order to ensure high level of explicitness of semantic learning and compatibility to existing CNN based approaches. However, it is believed that a well-designed network structure may explicitly incorporate semantic information itself and facilitate efficient training in an end-to-end training manner. It will be our future work to investigate improved architecture which enables joint learning of semantic information explicitly as well as preserving the network integrity.

CHAPTER 5

Periocular Recognition by Strengthening Attention to Critical Regions in Deep Neural Networks

5.1 Background

As discussed in Chapter 4, periocular recognition has been receiving increasing attention for its promising performance especially under less constrained conditions [39] [40]. Periocular region has been validated to be discriminative for different persons, and is considered highly effective as an independent biometric modality or as supplement to face and/or iris recognition.

In spite of usefulness of periocular recognition, matching periocular images accurately under less constrained environments remains a challenging problem in the community. By reviewing the recent development of periocular recognition algorithms, we can conclude that there is still considerable space for the matching accuracy improvement in order to meet the need for large scale real applications, and therefore further research efforts are necessary to advance state-of-the-art performance for periocular recognition.

5.1.1 Limitations of Existing Works

Despite the significant and encouraging research progress gained by aforementioned studies in Chapter 1.4 as well as the proposed approach in Chapter 4, the performance of periocular recognition still needs to be further improved in order to meet the expectation for real applications. Besides, existing periocular feature extraction methods seldom consider the underlying regional significance that may exist in periocular images. In summary, the following aspects require further research in order to facilitate the performance of periocular recognition:

- Hand-crafted features and shallow learning models are still in the majority of focus for periocular recognition algorithms. Advanced deep learning architectures and technologies, whose effectiveness has already been largely ascertained, have immense potential but not yet been fully exploited in this area, possibly due to the need for large amount of training data;
- Several studies already revealed the importance of eye and eyebrow regions for periocular recognition, but most of existing approaches only consider including these regions for the input/acquired images, and little effort has focused on emphasizing these regions during feature extraction process.

Based on the above facts as well as earlier studies on the human visual attention, this chapter proposes an attention based CNN architecture for more accurate and robust periocular feature learning, under the assumption that eyebrow and eye regions preserve higher importance and deserve more attention than the surrounding skin areas. As discussed earlier, employing visual attention mechanism may address the regional significance for the deep feature extraction and benefit the recognition accuracy [111]-[114]. Besides, several mechanisms including customized network structure, pair-wise training and dynamic data augmentation are adopted to relax the need for training data.

5.1.2 Our Work

In this chapter we propose the attention based deep learning architecture, referred to as *AttNet*, for more accurate and robust periocular recognition under less constrained environments. The key assumption of our approach is that, the eyebrow and eye region are critical for periocular recognition and should attract additional attention for feature learning. This is inspired by human perception as well as the recent trend in the deep learning community, which suggests that incorporating visual attention to potentially

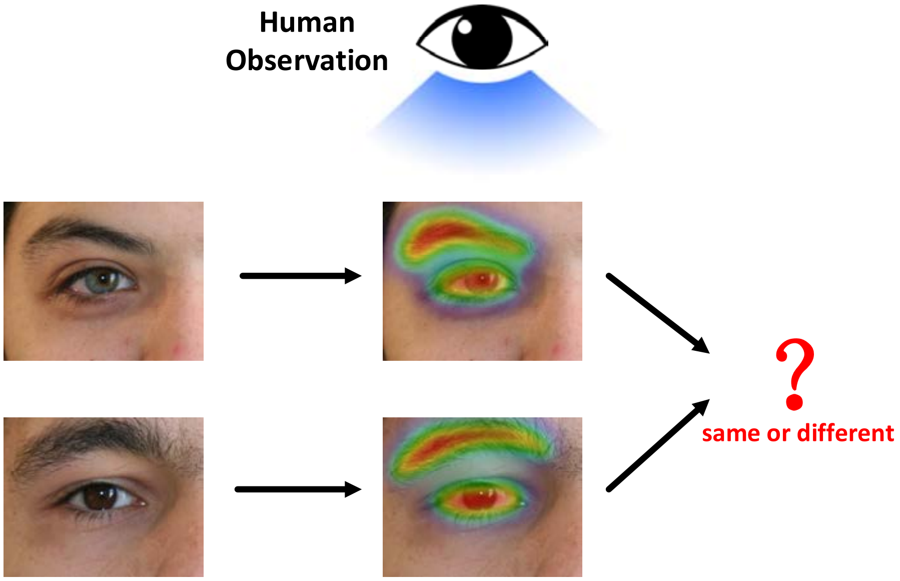


Figure 5.1: Illustration of implicit human visual attention while performing recognition tasks such as periocular verification. Critical regions that can provide more discriminative information attract more attention, especially for the find-grained recognition.

more important regions can significantly benefit the performance for a number of image understanding tasks [111]-[114]. As illustrated in Figure 5.1, when human performs recognition tasks, salient regions such as eye and eyebrow within periocular may provide more discriminative information, and naturally attract more attention than the surrounding regions.

With such assumption, we develop the explicit attention based deep neural network, which incorporates a region of interest detection network and attention implication module. The proposed framework is shown to extract more comprehensive periocular features with higher discriminative capability. The main contributions of our work can be summarized as follows: 1) the proposed approach achieves superior accuracy for periocular recognition under less constrained environments with visible and near-infrared (NIR) imaging. Extensive experimental results on four publicly available databases suggest that our attention based model outperforms several state-of-the-art methods significantly. Such results provide strong support to our assumption on the importance of critical regions, *i.e.*, eye and eyebrow, for more accurate periocular recognition. 2) We also present a customized loss function, referred to as *Distance-driven Sigmoid Cross-entropy* (*DSC*) loss. The *DSC* loss is shown to offer a marginal effect for both positive and negative training samples during the verification oriented learning, which results in more effective supervision compared with other loss functions such as contrastive loss and triplet loss.

The trained models and source codes of our approach are provided in [116] for reproducing our experimental results, so that other researchers can easily follow our work for further research progress on periocular recognition.

The rest of this chapter is organized in the following way: Chapter 5.2 explains the methodology of the proposed approach, including the visual attention based model and the customized *DSC* loss function; Chapter 5.3 and 5.4 provide analysis on the importance of attention-drawing regions and convergence status of training respectively; Chapter 5.3 details the experimental configurations and the analysis on the results; Chapter 5.4 draws conclusions of this work and introduces our future research goals.

5.2 Proposed Methodology

As discussed earlier, the key innovation of our method is the incorporation of attention model which draws the network attention to specific region of interest (RoI) during feature learning and matching for the periocular recognition. The overall framework is illustrated in Figure 5.2. The proposed network structure, referred to as *AttNet* in this work, firstly exploits a convolutional unit (i.e., conv1) for extracting low-level features from the input image. The network is then split into two branches, where the

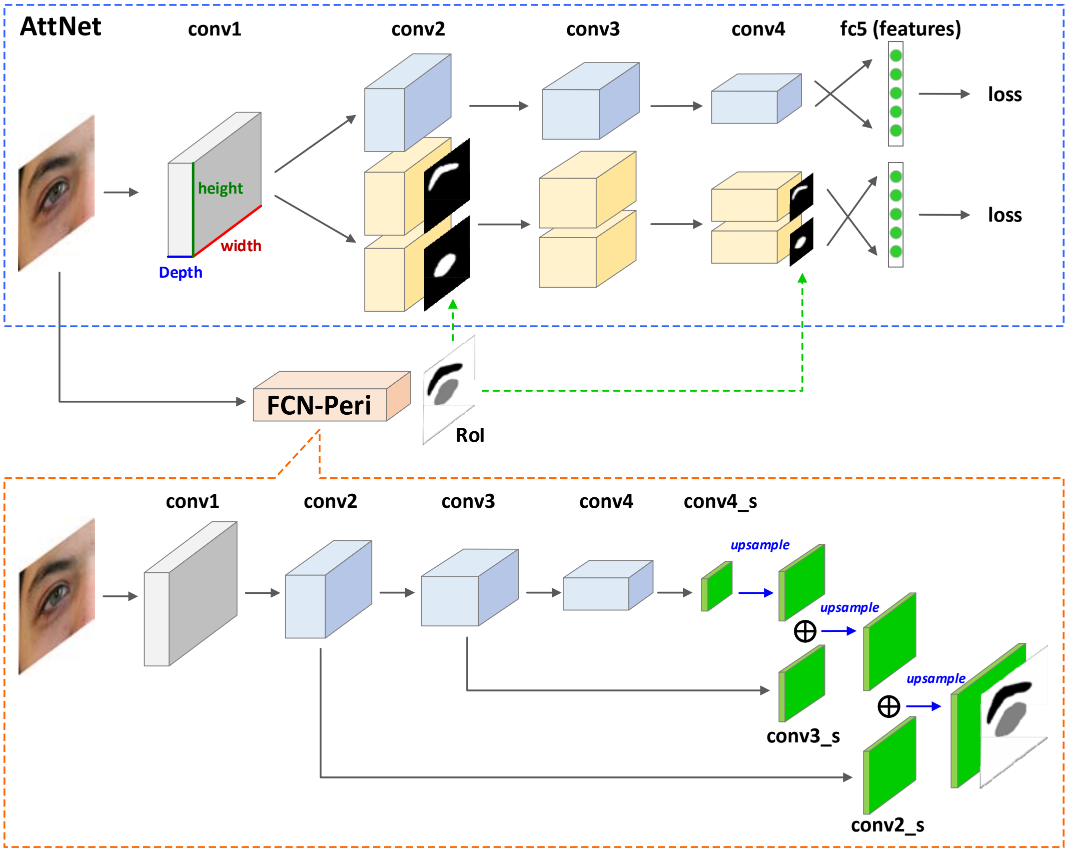


Figure 5.2: Architecture of the proposed attention based convolutional neuron network, referred to as *AttNet* (top), and the utilized fully convolutional network for specific region detection, called *FCN-Peri* (bottom).

first branch process the bottom inputs as usual CNNs, while the second branch incorporates RoI information in its intermediate layers (i.e., conv2 and conv4) so that higher attention is imparted to the specific areas of the input periocular image. The first branch without utilizing attention mechanism is designed to recover global features that a typical CNN can perform, which is able to maintain the robustness of the network when RoI information is incorrect, and improve overall performance by feature conjunction. The RoI information is provided by a fully convolutional network (FCN) [56], i.e., *FCN-Peri* in Figure 5.2. The detailed layer configuration of these two networks are provided in Table 5.1. Kindly note that both networks employed in this work are relatively simple compared with popular and very deep architectures such as VGG [62] and ResNet [89], considering the availability of training data. Besides, we

Table 5.1: Detailed layer configurations for *AttNet* and *FCN-Peri*.

Unit	Layer	Type	#Output channels	Kernel size	Stride
AttNet					
conv1	conv1_1	convolution	32	5×5	1
	relu1_1	ReLU	/	/	/
	conv1_2	convolution	32	5×5	1
	relu1_2	ReLU	/	/	/
	pool1	max pooling	/	2×2	2
conv2	conv2_1	convolution	32	3×3	1
	relu2_1	ReLU	/	/	/
	conv2_2	convolution	32	3×3	1
	relu2_2	ReLU	/	/	/
	pool2	max pooling	/	2×2	2
	att2*	attention	/	/	/
conv3	conv3_1	convolution	64	3×3	1
	relu3_1	ReLU	/	/	/
	conv3_2	convolution	64	3×3	1
	relu3_2	ReLU	/	/	/
	pool3	max pooling	/	2×2	2
conv4	conv4_1	convolution	64	3×3	1
	relu4_1	ReLU	/	/	/
	conv4_2	convolution	64	3×3	1
	relu4_2	ReLU	/	/	/
	pool4	max pooling	/	2×2	2
	att4*	attention	/	/	/
fc5	fc5	fully connected	64	/	/
FCN-Peri					
conv1	conv1	convolution	16	5×5	1
	relu1	ReLU	/	/	/
	pool1	max pooling	/	2×2	2
conv2	conv2	convolution	32	3×3	1
	relu2	ReLU	/	/	/
	conv2_s	convolution	3	1×1	1
	pool2	max pooling	/	2×2	2
conv3	conv3	convolution	64	3×3	1
	relu3	ReLU	/	/	/
	conv3_s	convolution	3	1×1	1
	pool3	max pooling	/	4×4	2
conv4	conv4	convolution	128	3×3	1
	relu4	ReLU	/	/	/
	conv4_s	convolution	3	1×1	1

* Two branches of *AttNet* as shown in Figure 5.2 have the same layer configuration, but attention layers are only placed in the second branch.

adopt the Siamese infrastructure for training the network in end-to-end verification protocol, and develop a new compositional loss function which is referred to as Distance-driven Sigmoid Cross-entropy (DSC) loss. This new DSC loss has shown to offer superior performance than traditional verification oriented loss functions like contrastive loss and triplet loss.

In this chapter, the detailed mechanisms for RoI detection and attention implication are explained in Chapter 5.2.1 and Chapter 5.2.2 respectively; Chapter 5.2.3 presents the newly developed *DSC* loss function, followed by the details on the training and test configurations in Chapter 5.2.4.

5.2.1 *FCN-Peri* – Semantical Region Detection

The key issue for incorporating visual attention model is to identify potentially important regions that deserve more attention than other regions during learning. In general image classification/understanding, the inference of important regions is often jointly learned with the specific tasks [111] [114], as the input data generally involves significantly different background information and those regions could not be predefined. Such strategies, however, require huge amount of training data with sufficient variation to regularize the learning process. For fine-grained tasks such as periocular recognition, predefined region detection is preferred [113] as prior knowledge about the input images is usually available, so that the learning process can be better regularized with limited training data. In our approach, based on human perception model, we assume that the regions containing eyebrow and eye are relatively important for periocular recognition. Under such assumption, we firstly exploit a fully convolutional network (FCN) to detect the eyebrow and eye regions.

The FCN employed in our work was firstly proposed for the semantic segmentation in [56]. Different from common CNNs, FCN does not contain fully

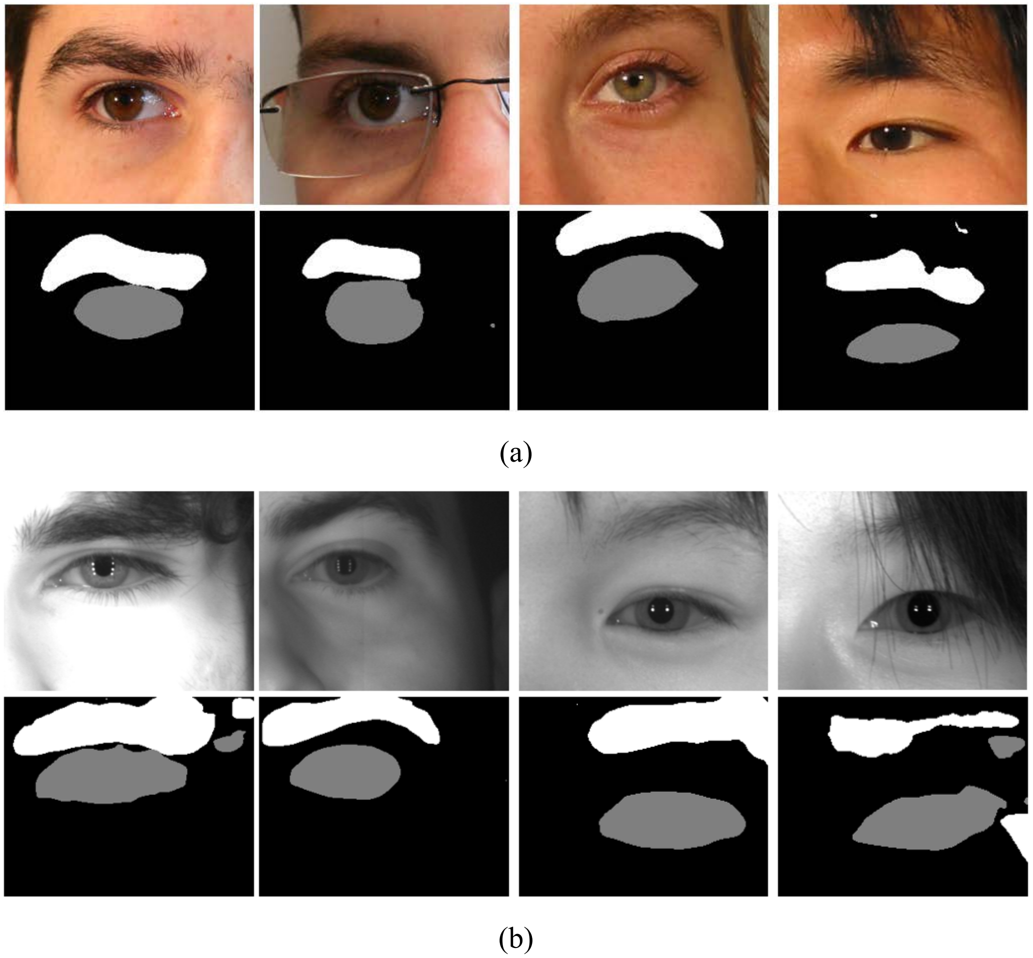


Figure 5.3: Samples outputs of *FCN-Peri* for test images with visible (a) and near infrared (b) imaging. The black pixels represent predicted background, and the white and gray pixels identify predicted eyebrow and eye respectively.

connected layers, and the upsampling layers are utilized to integrate intermediate convolutional feature maps at different scales. The spatial correspondence between the input image and the output features is therefore maintained to achieve pixel-to-pixel prediction. The FCN is supervised by a pixel-wise *softmax* loss function using groundtruth labels. In our approach, we employed a simplified version of the FCN proposed in [56] for segmenting eyebrow and eye from background in the input periocular image, which we refer to as *FCN-Peri*. The detailed architecture of *FCN-Peri* is illustrated in Figure 5.2 (bottom), which contains about 0.1M parameters.

The original FCN in [56] was developed to classify each pixel into one of 21 classes. In our work, eyebrow and eye are regarded as two different classes, and pixels in the original input image are to be segmented into three classes, *i.e.*, eye, eyebrow and background. We manually labeled the eyebrow and eye regions for about 100 images from the training sets of visible and near infrared (NIR) data (details of datasets are in Section III) respectively as the ground truths to train *FCN-Peri* from scratch. It should be noted that by "eye region", we refer to the region including the iris, sclera, eyelid and eyelash, *etc*., rather than just the iris region. Figure 5.3 shows several region segmentation results from trained *FCN-Peri* on the test sets. It can be observed that the region predictions are quite robust despite that it makes some mistakes for some challenging samples. The proposed attention based deep neural network, *i.e.*, *AttNet*, is however expected to be tolerant to such level of errors in a few samples. Also kindly note that the networks for visible and NIR spectrums are trained separately.

5.2.2 *AttNet* – Incorporating Visual Attention for Periocular Feature Learning

With the detected regions containing eyebrow and eye for an input image from *FCN-Peri*, we then incorporate the resulting RoI in *AttNet* for attention model implementation. As shown in Fig. 2, after convolutional units conv2 and conv4, the output map from *FCN-Peri* indicating eyebrow and eye positions is utilized to adjust the convolutional features. There is no standard procedure for accomplishing attention in deep neural networks. Some methods use the RoI for affine transformation and alignment [112], while others consider bluring/masking the background for the input images or intermediate features [114], or feed cropped areas into multiple deep networks [113]. In our approach, we apply a straightforward yet effective mechanism for emphasizing important areas inferred by *FCN-Peri*, *i.e.*, increasing the magnitudes of the convolutional features within the RoI and decreasing those outside the RoI.

More specifically, an *attention layer* is placed after a convolutional unit and performs the follow operation:

$$f'_{x,y} = \begin{cases} \alpha f_{x,y} & \text{, if } (x, y) \in R \\ \dfrac{1}{\alpha} f_{x,y} & \text{, otherwise} \end{cases} \tag{5.1}$$

where R is the set of x-y coordinates where the current position is considered as RoI, $\boldsymbol{f}$ is the convolutional feature map from the previous layer, $\boldsymbol{f'}$ is the processed feature map before entering the next layer, and α is a positive parameter controlling the intensity of adjustment. It was empirically fixed to 5 for all our experiments. Such operation attempts to simulate human visual attention by weighting the features within the RoI more than those in the background for the subsequent layers of the network. The feature adjustments for eyebrow and eye are separately performed, each on half of the channels of the feature maps respectively, as these two regions present quite different characteristics. We selectively incorporate such attention mechanism for conv2 and conv4 to account for both low-level and high-level convolutional features. Since conv1 is shared by the RoI-aware and common branches, conv2 is therefore more appropriate to incorporate for the low-level attention. On the other hand, conv4 is right before the fully connected layer fc5 (*i.e.*, the layer generating feature vectors) and is also judicious to be selected to impart high-level attention. Figure 5.4 visualizes the effect of the employed attention model for the features from the two convolutional units. It can be observed that the background features which do not belong to the RoI "fade" after the operation by attention layers. In this way, the foreground features make more impact on the feature extraction process by subsequent layers. Although simply increasing the feature magnitudes inside the RoI may not be an *optimal* approach to incorporate visual attention, it is a reasonble and easy-to-implement scheme to achieve key objective of our research, *i.e.*, to investigate and evaluate the importance of eye and eyebrow regions to advance periocular recognition through the deep periocular

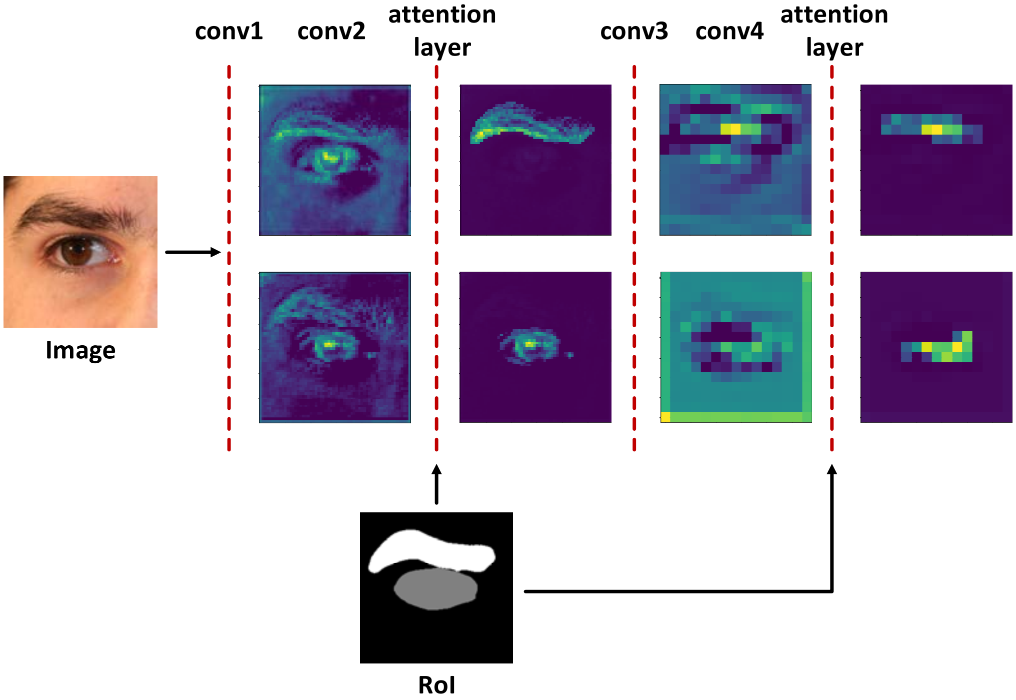


Figure 5.4: Visualization of convolutional features from intermediate layers before and after attention layers. *Attention layers* increase the feature values within the RoI, and meanwhile decrease those in background. Feature maps of different scales are upsampled to the same size for better illustration.

feature extraction.

5.2.3 Distance-driven Sigmoid Cross-entropy (*DSC*) Loss for Verification Oriented Supervision

We adopt Siamese-like pair-wise network infrastructure for training our *AttNet*, *i.e.*, instead of classifying a single image into a standalone class, a pair of images are jointly evaluated to predict whether they belong to the same class or not. Such configuration is illustrated by Figure 5.5. Contrastive loss [115] or triplet loss [57] are often used for the pair-wise training. Compared with the classification training protocol which usually uses a *softmax* loss function for supervision, the pair-wise protocol is closer to the verification problem (one-to-one matching) which is a fundamental application scenario for most biometric systems. A classification based model, in contrast, may

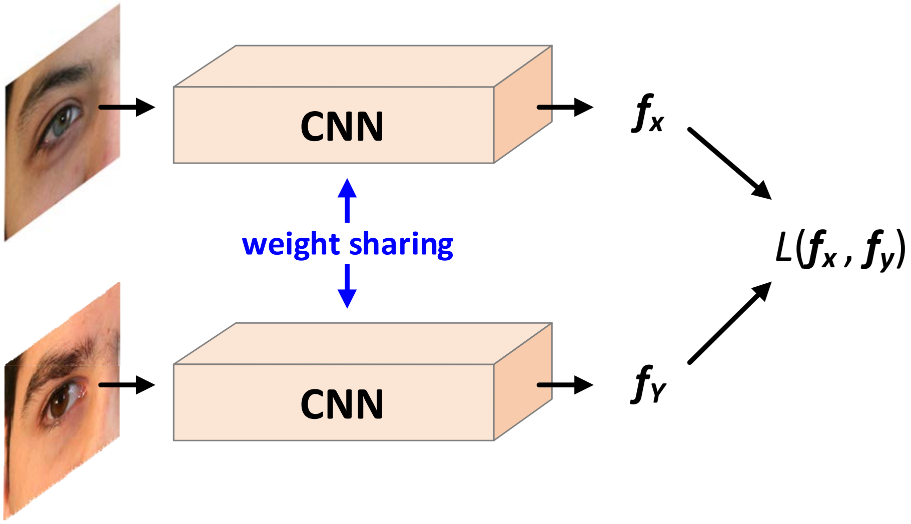


Figure 5.5: Illustration of Siamese architecture for training CNN in verification protocol. Two identical CNNs are placed in parallel to process a pair of samples. Specific pair-wise loss function (*e.g.*, contrastive loss) is employed to supervise the training, and the weights (parameters) of the two networks are kept the same (weight sharing) during the entire training process.

require additional transfer learning to make itself more effective and scalable, such as in [57]. Besides, the pair combination from training samples introduce more data variation, which is believed to reduce the overfitting of trained model. In the following, we present a brief introduction to conventionally used loss functions for the pair-wise training, followed by our newly designed *DSC* loss function.

A. Conventional Verification Oriented Loss Functions

The conventional contrastive loss function for training Siamese network is formulated as follows:

$$L_{con} = td^2 + (1-t)\max(0, m-d)^2 \tag{5.2}$$

where t is the label of the current pair, *i.e.*, t=1 if the two samples come from a same class and t=0 otherwise, and d is simply the Euclidean distance between the two input feature vectors $\boldsymbol{f}_X$ and $\boldsymbol{f}_Y$:

$$d = \left\| \boldsymbol{f}_X - \boldsymbol{f}_Y \right\|_2 \tag{5.3}$$

m is a preset margin for regularizing the distance from a negative pair (*i.e.*, a pair for samples from different classes). The contrastive loss is designed to reduce the distance between a positive pair as a quadratic energy term, while for negative pairs, the distance between a negative pair would be increased until it exceeds the hard margin m. The effect of m is to force the network to concentrate on relatively challenging negative pairs only. However, there is no regularization on the positive pair samples. As the training progresses, more and more negative pairs do not produce any losses due to the hard margin, while all the positive pairs still have continues impact on the backpropagation. This causes unbalanced training for positive and negative pair samples.

The above side effect is to some extent alleviated by triplet loss, which can be considered as a variant of contrastive loss. Instead of evaluating a simple pair, the triplet loss composes positive and negative pair into a triple structure, and measures the loss by:

$$L_{tri} = \max(\left\| \boldsymbol{f}_{X_1} - \boldsymbol{f}_{X_2} \right\|_2^2 - \left\| \boldsymbol{f}_{X_1} - \boldsymbol{f}_Y \right\|_2^2 + m', 0) \tag{5.4}$$

where $\boldsymbol{f}_{X_1}$ and $\boldsymbol{f}_{X_2}$ are features from a same class while $\boldsymbol{f}_Y$ is extracted from another class. Different from contrastive loss, which uses an absolute margin to regularize negative pairs, the triple loss relies on a relative margin m' to enlarge the difference between the positive pair distance and negative pair distance. In this way, the balance of positive and negative pair samples is always retained during the training process. Verification oriented applications, however, mostly use an absolute value as threshold instead of relative margin for decision making, and therefore slight inconsistency exists between the training process supervised by triplet loss and the actual test (matching) process.

B. Distance-driven Sigmoid Cross-entropy (DSC) Loss

In order to address the above limitations, in this chapter we introduce a customized compositional loss function called *Distance-driven Sigmoid Cross-entropy* (*DSC*) loss. Given the distance d between a pair of features to be evaluated, we firstly perform following mapping on it:

$$s = b - ad^2 \tag{5.5}$$

$$p = \frac{1}{1+e^{-s}} \tag{5.6}$$

where a and b are positive constants which are used for linear transformation on the square of the Euclidean distance, p is obtained by a sigmoid function on the transformed s and can be regarded as the probability that the two samples come from a same class. The motivation of using sigmoid function is that it maps any real value into (0, 1), and varies significantly near zero but much slower at two ends. Such property essentially enables a kind of soft margins for the low and high values of s. In this way, the learning process for both positive and negative pairs can be regularized, so that it mainly focuses on challenging samples with s values near zero. The loss for the obtained probability p is then measured by the cross-entropy function:

$$L_{DSC} = -[t \log p + (1-t)\log(1-p)] \tag{5.7}$$

The sigmoid cross-entropy loss is widely used when the task is to predict probabilities of certain events. In this case, we regard our task as predicting the probability of a binary event – same class or different classes. Different from common approaches which feed a single neuron output spanning over $(-\infty, +\infty)$ into the sigmoid function, we originally map the Euclidean distance d to a term s that spans over $(-\infty, b]$, then transfer to approximated probability p. The constant b should be selected such that its sigmoid value $1/(1+e^b)$ is very close to one. Such transfer is the key to the new *DSC* loss function which utilizes the soft margins of sigmoid function in a straightforward

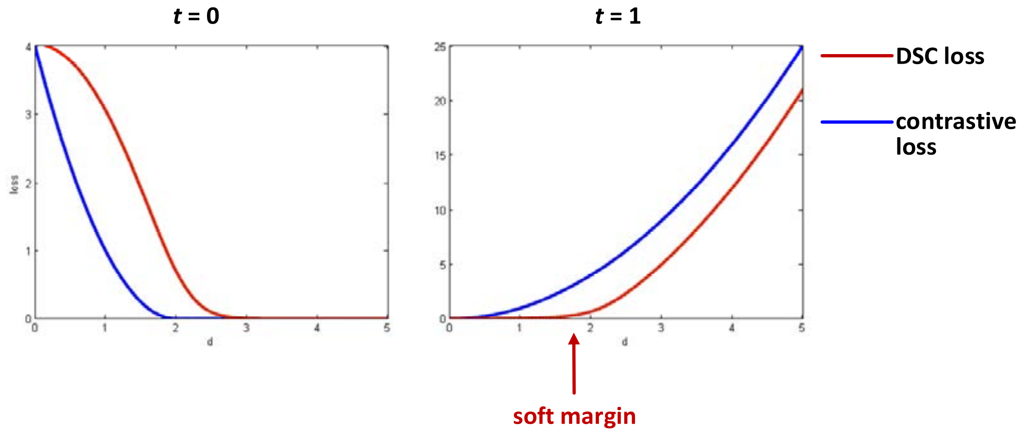


Figure 5.6: Comparison of *DSC* loss (*a* = 1, *b* = 4) and conventional contrastive loss (*m* = 2) with respect to *d*. The *DSC* loss provides a (soft) margin for positive cases (*t* = 1) which achieves better regularization for genuine pairs, such that the learning process mainly focuses on challenging samples.

way.

Figure 5.6 demonstrates the comparison of the newly developed *DSC* loss function and conventional contrastive loss function w.r.t *d*, for both positive (*t* = 1) and negative (*t* = 0) cases. It can be clearly observed that for negative cases the two losses have similar distribution that, when *d* is greater than certain values, the losses approach to zero. Such marginal effects make sure that the learning process does not waste energy on unchallenging negative pairs that already have large distance. For positive cases, however, notably different characteristics are presented by the two losses. The contrastive loss simply evaluates the distance with a quadratic term, which results in the fact that unchallenging positive samples would have continuous impact on the learning process. In contrast, a number of negative samples would be ignored due to the hard margin *m*. Such imbalance may mislead the training process to focus too much on positive samples, even for unchallenging ones. On the other hand, our *DSC* loss provides a (soft) marginal effect for positive cases as well, *i.e.*, when *d* is in certain small range, it produces a loss close to zero. Such minor loss values indicate that the current samples are typically unchallenging, and they do not generate noticeable gradients for the backpropagation of the training process. In this way, the learning keeps focusing on challenging samples, for both positive and negative cases, to maximally increase the discriminating capability of the network.

As would be shown from the experiments in Chapter 5.3.3, the proposed *DSC* loss contributes to better discriminating power than conventional contrastive loss and triplet loss, especially for lower false acceptance rates.

5.2.4 Training and Test Configuration

In order to improve the network generalizability and feature effectiveness, we have adopted several commonly used data augmentation techniques for the training process, as well as feature composition during the matching phase. These measures are explained in the following.

A. Training Data Augmentation

All the training images are resized to 300×240 in advance. Besides, we have performed several *on-the-fly* image augmentation approaches which are commonly adopted in various deep learning studies and proven to help improving the performance. These approaches are randomly applied before each image is fed into the network, and are described in the following:

- Scaling – There is 80% probability for each image to be enlarged, with a factor randomly drawn from a uniform distribution over (1, 1.3).
- Cropping – Each image is cropped with a window of 240×240 that is randomly placed across the entire image region.
- Color/intensity jittering – For an RGB image (visible imaging), a color augmentation method called Fancy PCA as described in [88] is applied. For a

grayscale image (NIR imaging), a random value drawn from $\mathcal{N}(0, 0.02)$ is added to its pixel intensities to simulate illumination variation.

Above random parameters are drawn once for each image in the mini-batch during the training process. When a same image appears again in a later iteration, the parameters will be randomly drawn again to create a different variant of that image. In this way, one source image can produce a good amount of different versions without consuming much of the storage space. Such augmentation measures can effectively reduce the risk of over-fitting when training deep neural networks, especially when the number of training samples is not very large.

B. Test Feature Composition

As mentioned earlier, our network model accepts 240×240 square image as the input. On the other hand, the source periocular images used in our experiments have rectangular aspect ratios close to 5:4. During the test phase, we adopt feature composition similar to [61] and [62], to make our model adaptive to (slightly) different resolutions / aspect ratios, and also to obtain multi-scale feature representation. The composition process is described sequentially in the following:

a) The input image is resized to *w*×240, where *w* is larger than 240 and subject to the image's original aspect ratio.

b) The resized image is cropped with three 240×240 windows that are placed on the left end, center and right end of it respectively.

c) The resized image is enlarged with a factor of 1.2, then another 240×240 window is placed in the center of it, to create the fourth cropped version.

d) Four cropped versions are fed into the network separately, each generating a 128-D feature vector. These four vectors are then concatenated into a 512-D vector for the matching.

The Euclidean distance between two vectors is regarded as the dissimilarity score. Above feature composition process can cover the entire image region and account for the multi-scale feature representation to certain extent.

5.3 Analysis on Region Selection

In this section we will investigate the reasonableness of the pre-defined regions for visual attention enhancement. As mentioned before, we select eyebrow and eye as the RoI mainly due the following two reasons:

a) Inspired by human perception, eyebrow and eye regions will attract most of attention when humans observe periocular images. Kindly note that many machine learning / deep learning algorithms are inspired by human perception / behaviors, including neural networks, reinforcement learning, long-short term memory (LSTM) and also the referenced attention models in this paper.

b) The importance of eyebrow and eye characteristics for periocular recognition has been ascertained by a number of earlier research works [41][42][47][84][124], where excluding or masking eyebrow or eye regions will lead to performance degradation in most cases.

In order to statistically ascertain the effect of selecting these areas for attention enhancement, we have attempted training different versions of AttNet by adjusting the feature weights α in Equation (1), detailed as follows:

- Eye + Eyebrow: $\alpha_{\text{eye}} = \alpha_{\text{eyebrow}} = 5$
- Eye only: $\alpha_{\text{eye}} = 5,\ \alpha_{\text{eyebrow}} = 1$
- Eyebrow only: $\alpha_{\text{eye}} = 1,\ \alpha_{\text{eyebrow}} = 5$
- No attention: $\alpha_{\text{eye}} = \alpha_{\text{eyebrow}} = 1$

The above settings enable preliminary investigation into the effect of selected

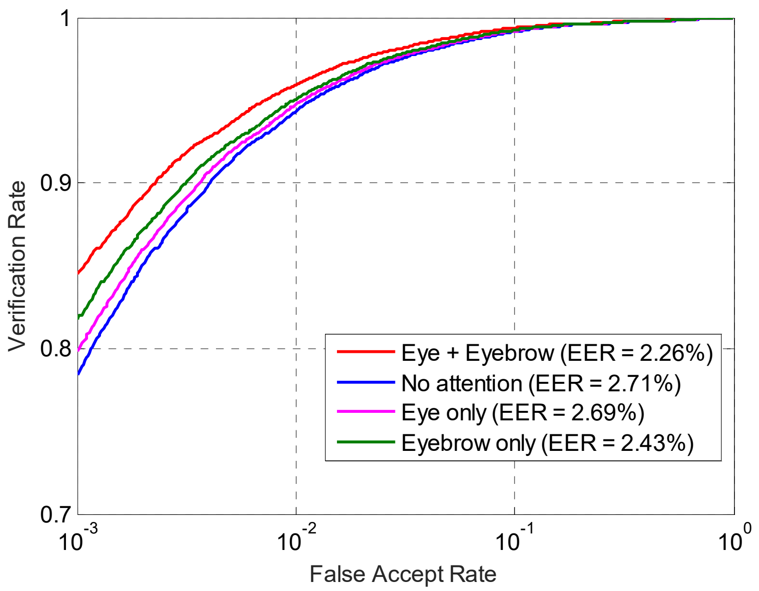


Figure 5.7: Comparison of different weights on the selected regions of interest for attention incorporation.

regions for attention enhancement on the recognition results. Comparative study was performed on the UBIPr database and the results are shown in Figure 5.7. It can be observed that with explicitly enhanced attention on eye and eyebrow regions simultaneously can mostly benefit the recognition accuracy. Emphasizing eyebrow region separately yields higher improvement than focusing on the eye region only. This is probably because the eyebrow characteristics are more stable and resistant to illumination variation, eyeball movement, *etc*. The above observations have validated the positive effect of incorporating visual attention within the detected eyebrow and eye regions during deep feature extraction for more accurate periocular recognition.

5.4 Analysis on Training

The effectiveness of training is a key aspect to consider for deep learning based approaches, which is related to a number of factors such as the classification task, network complexity, volume of training data and learning algorithm. Compared with

Table 5.2: Comparison of network configurations for our work and other typical architectures.

Architecture	Problem	#Classes	#Param.	# Train Images
AlexNet [88]	Image class.	1,000	60M	~1M
VGG-16 [62]	Image class.	1,000	138M	~1M
ResNet-152 [89]	Image class.	1,000	60M	~1M
DeepIrisNet [38]	Iris recog.	356	138M	~30K
PRWIS [123]	Periocular recog.	518	248M	~8K
AttNet	Periocular recog.	224	**7.7M**	~3K
FCN [56]	Semantic segm.	21	134M	~8K
FCN-Peri	Semantic segm.	3	**0.1M**	100

typical deep learning solutions on ImageNet classification [58][89][110], semantic segmentation [56], *etc.*, one of the most critical challenges when researchers explore deep learning's potential on biometrics may lie in the availability of labeled training data. Insufficient training data may cause severe over-fitting, *i.e.*, the model fits too well on the small scale of training data but is not able to properly classify test data which was unseen during training phase. In this section, we perform analysis on the training processes of *AttNet* and *FCN-Peri* to validate that our models are adequately trained and the level of over-fitting is within acceptable range.

5.4.1 Training of AttNet

There is no definite conclusions so far about the minimum required number to properly train a CNN for classification purpose. Generally, it is accepted that when there are more parameters to learn and the problem is more complicated, the required amount of training data will be larger in order to avoid over-fitting. A practical way is to refer to some typical architectures and the training configuration which have been widely adopted by researchers/developers in the literature. Table 5.2 presents the summary of scale of our networks as well as some existing architectures for different classification tasks.

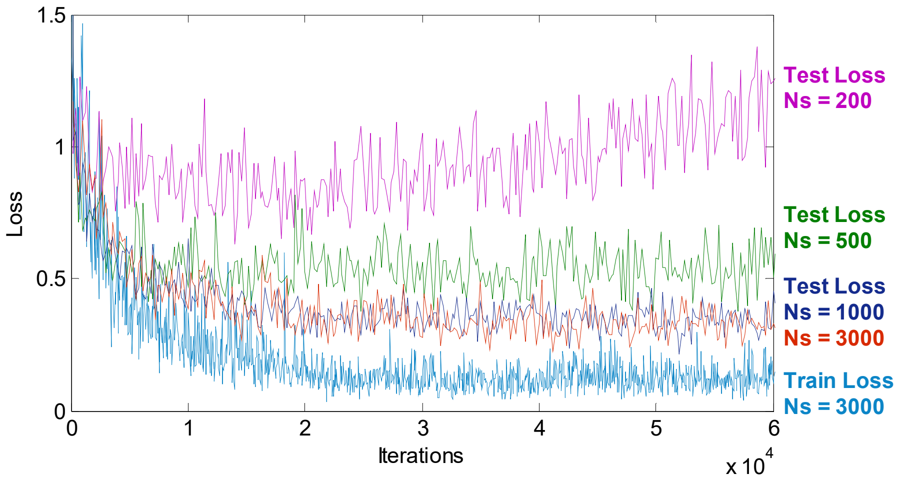


Figure 5.8: Learning status of *AttNet* with different number of training samples (*N*s). With *N*s no less than 1,000, test loss converges to a stable level. Train losses with different *N*s are similar and therefore only one is plotted for clarity. *Best viewed in color*.

It can be implied from the table that: (1) Our network is much smaller than other typical network architectures in terms of parameter scale, and it is therefore reasonable to assume that the required number of training samples should be less than other examples in the Table 5.2. (2) For general image classification such as [88] and [89], dramatic intra- and inter-class variation exists and large volume of training data should be applied for sufficient learning; On the other hand, for typical biometric problems such as iris and periocular recognition, relatively small amount of training samples was employed but promising results can still be obtained. This is probably because smaller inter-image variation for biometric recognition may not require that many training samples to supply over-complicated information. The periocular recognition problem discussed in this paper belongs to the latter. Considering the above two factors, our configuration for training the small *AttNet* with about 3,000 (on UBIPr dataset which will be detailed in the next section) images should be reasonable.

In order to statistically examine the convergence condition of our configuration, we vary the number of training samples to train *AttNet* on UBIPr database for several times and observe the convergence status. The results are shown in Figure 5.8. It can be observed that employing several hundreds of training images may easily cause over-fitting as there is a large gap between the train loss and the test loss. However, when the number increases to 1,000 or above, test loss converge to similar level and the gap becomes smaller. Note that it is difficult to totally eliminate the gap for most deep learning approaches. The above results indicate that the actual configuration we applied in this paper, in which approximately 3,000 images were used for training *AttNet*, is practically appropriate for sufficient training.

5.4.2 Training of FCN-Peri

The case of training an FCN for semantic segmentation is quite different from training a CNN for image classification. Semantic segmentation (*e.g.*, detecting eyebrow and eye regions in this paper) is a task of pixel-wise classification, rather than entire image classification. In other words, with semantic segmentation, each pixel in the input image is classified into one of several pre-defined classes. Therefore, analysis on the number of training samples, or data points, should be casted at pixel level instead of image level. However, not all the pixels should be considered as independent data points, as adjacent pixels will have highly redundant information. The concept of *receptive field* can help to more scientifically estimate meaningful data points in an image when training FCN.

In single or multiple regular convolution/pooling operations, one output element or pixel is computed from a certain region in the input image/map, and this region is referred to as the *receptive field*. For example, with one convolutional layer in CNN/FCN having a 3×3 kernel, the receptive field is 3×3. With two such convolutional layers, the receptive field from input to output is 5×5. can illustrate the concept. Since FCN mainly comprises convolutional layer and pooling layer, each output

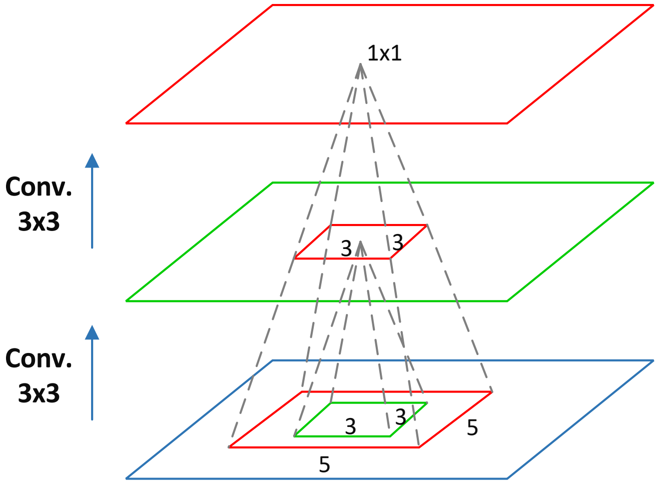


Figure 5.9: Illustration of receptive fields. Through one or more convolutional or pooling layers, each output neuron in the top layer is determined by a patch in the bottom/input layer.

element/pixel is determined by a patch from the input rather than the entire image. We can therefore compute the receptive field of *FCN-Peri* first to estimate the approximate number of non-redundant data points available in the training process.

The receptive field can be computed in a top-down manner to identify the region at bottom layer determining one pixel at the topmost layer. Following the longest path from input to output in *FCN-Peri*, the process is illustrated in the following:

Layer	**Kernel, Stride**	**Receptive Field**
output	-	1×1
upsample×3	-	2×2
conv4	3×3, 1	4×4
pool3	4×4, 4	16×16
conv3	3×3, 1	18×18
pool2	2×2, 2	36×36
conv2	3×3, 1	38×38
pool1	2×2, 2	76×76
conv1	5×5, 1	80×80

The result indicates that each output pixel of FCN-Peri is determined by a patch of 80×80 from the input image. We can roughly assume that two patches can be considered as independent data points when the overlap between them is no less than 25% (otherwise the information will be highly redundant). As a result, a 300×240 image we used as input can provide approximately 108 (9×12) non-redundant data points. As discussed earlier, we have labelled about 100 images for training FCN-Peri, generating approximately 10,000 data points for learning classification of three classes (i.e., eyebrow, eye and background). On average, about 3,000 training samples per class are available for training. Note that network is more than 1,000 times smaller than the original FCN as revealed from Tab. 2, which suggest that the number of available training samples should be sufficient. In fact, the segmentation results on test data shown in Fig. 3, which were visually appropriate, can also validate that our FCN-Peri has been properly trained.

5.5 Experiments and Results

Thorough experiments have been performed to evaluate the proposed approach from various perspectives, and comparisons are made with several state-of-the-art methods. All of our experimental results are reproducible via [116]. We have conducted two sessions of experiments, which focuses on *Open-World* problem and *Closed-World* problem respectively. In this chapter we detail the problem definition, experimental configurations as well as observation and analysis on the results.

5.5.1 Open-World vs. Closed-World Verification

The open-world problem refers to the configuration that the subjects to be enrolled into the gallery in the deployment process may be *unseen* during the training phase. On the other hand, the closed-world problem has a constraint that all the subjects to be recognized in the deployment process are already *known* during the training phase.

The open-world problem is apparently more challenging but closer to the real deployment environments for most applications, such as citizen authentication, general access control and searching for missing people, as it is impractical for these systems to collect data from all possible subjects in advance during training/development phase. The closed-world setting may result in higher recognition accuracy as more precise data adaptation can be achieved during training, but the system may be less scalable for the deployment, which is also clarified by [123].

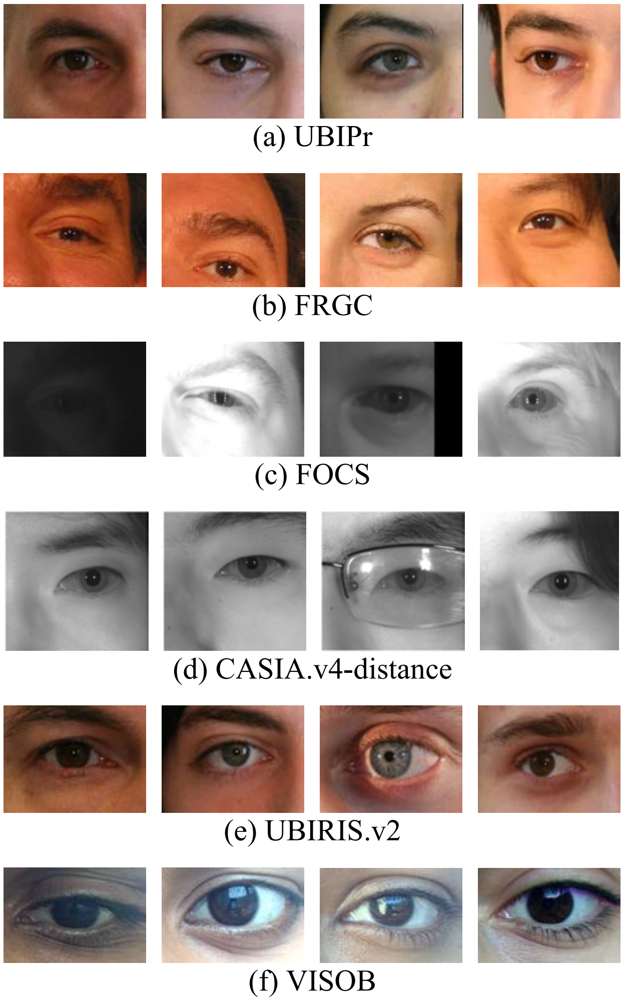
It should be clarified that the approached presented in this paper, especially the newly developed DSC loss function, are proposed for the open-world problem. However, we noticed that some recent method and contest [118] [123] in the literature focus on closed-world problem only, and therefore we investigate the performance under both settings.

5.5.2 Baseline Methods

Several state-of-the-art methods, *i.e.*, SCNN proposed in Chapter 4 [48], [10], [47] and [123], are selected as baselines to evaluate the performance of proposed approach. These methods are selected as baselines because they focus on the same problem with us, *i.e.*, less constrained periocular recognition under either visible or NIR imaging, and report state-of-the-art performance on multiple datasets in the recent years and with judicious theoretical significance. Kindly note that methods [10], [47], [48] and also ours are adaptive to the open-world setting, while [123] is only developed for closed-world setting as clarified in their paper.

5.5.3 Datasets and Protocols

We employ six publicly available databases for the experiments. Four of them are acquired under visible spectrum while the other two are with NIR imaging. While four of these databases have been introduced in Chapter 4.4.1, two of them are newly


(a) UBIPr

(b) FRGC

(c) FOCS

(d) CASIA.v4-distance

(e) UBIRIS.v2

(f) VISOB

Figure 5.10: Sample images from the employed databases, which present noticeable pose, illumination variation and occlusions due to the less constrained imaging environments.

included or with new training/test protocol applied, which are detailed in the following:

- UBIRIS.v2 [18]

As introduced earlier, this dataset is released for noisy iris recognition under visible spectrum. The full set contains 11,101 eye images from 518 subjects, which are acquired from 3-8 meters away. In this chapter, experiments on this dataset is mainly set for *closed-world* verification and comparison with method [123], but will also attach *open-world* results for comparative study. In the closed-world setting, 80% of images from all 518 subjects are used for training and the remaining 20% are selected for testing. In the open-world setting, images from the first 400 subjects are used for

Table 5.3: Summary of the employed databases for training and testing. The training sets of FRGC and CASIA.v4-distance are used for training [10] and [47]. Our method and [48] only adopt UBIPr and FOCS for training.

Database	**UBIPr**		**FRGC**		**FOCS**		**CASIA.v4-dist.**		**UBIRIS.v2**		**VISOB**	
Spectrum	visible		visible		NIR		NIR		visible		visible	
Imaging distance	4 – 8m		N/A		N/A		≥3m		3 - 8m		8 - 12 in.	
World scenario	Open		Open		Open		Open		Open/closed		Closed	
Division	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
#Subjects	224	120	13	150	80	56	10	131	518	518	484	475
#Images	3,359	1,767	40	500	3,262	1,530	79	998	8,886	2,215	5,270	5,103
#Genuine scores (Test)	12,351		826		39,614		3,371		2,215		4,914	
#Imposter scores (Test)	1,547,910		123,425		1,130,071		494,132		1,145,155		2,464,938	

training while the remaining are used for testing.

- VISOB [118]

This competition dataset comprises ocular images captured with three different smartphones under three illumination conditions. The Visit-1 involves 550 subjects and was released for algorithm development. The Visit-2 has images from 290 subjects and was used for performance evaluation in the competition. Kindly note that the data we acquired from the competition organizer only contains Visit-1, and therefore our experimental results were obtained on Visit-1 only and should not be directly compared with the published ranked methods in [118]. *Closed-world* setting was applied on the experiments on this dataset.

The six employed datasets cover both visible and NIR spectrums, and were collected under varying and less constrained imaging environments that are close to real world application scenarios. A few sample images from them are provided in Figure 5.10. More detailed information about the employed databases and training/test set division is provided in Table 5.3.

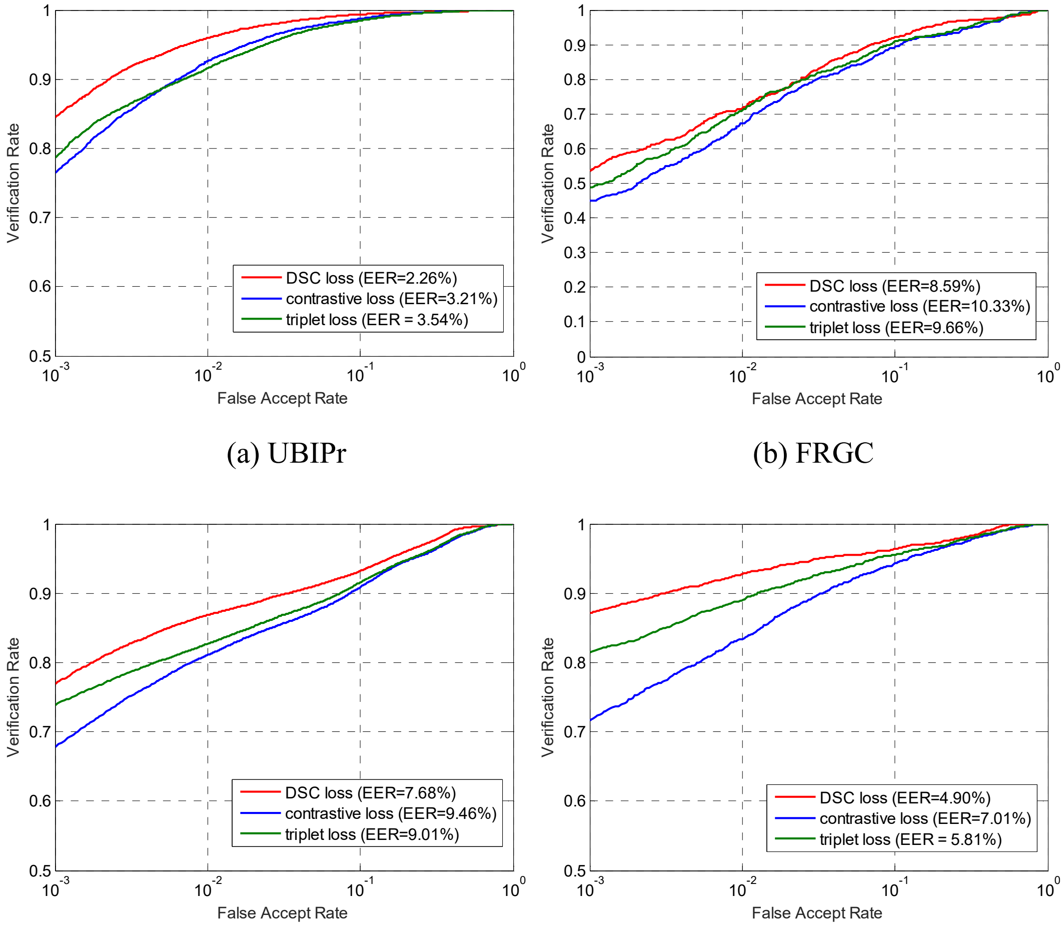
For experiments carried out under open-world configuration, it is important to clarify the reasonable difference of training mechanisms for the four methods: a) For our method and [48], the visible models are trained on UBIPr database and tested on

UBIPr and FRGC databases; the NIR models are trained on FOCS and tested on FOCS and CASIA.v4-distance datasets. In other words, experiments on FRGC and CASIA.v4 are under *cross-database* scenarios. Such a training/test configuration is identical to the original one in [48], which therefore provides a fair comparison. Nevertheless, while our model only uses the left periocular images for training, the model from [48] has employed both the left and the corresponding right periocular images which are required for training its semantical branch CNN. The result will be that [48] potentially benefits from two times more training samples during the comparison. b) For methods [10] and [47], the required training efforts are less, and it is observed that the *within-database* training and testing manner offers better results for these two methods. Therefore the training and testing are performed on the same dataset for them. Aforementioned experimental configuration is also the same as used in [48], and justification has been provided to incorporate the best possible performance from these two baseline methods and ensure fairness in the performance comparisons.

5.5.4 Open-World Performance

A. Effectiveness of *DSC* Loss Function

The performance of the proposed *DSC* loss function is firstly examined. We compare it with conventional contrastive loss and triplet loss, which are also designed for verification tasks. The experiment is performed on all the employed databases. Three *AttNet* models with identical structures are trained with *DSC* loss, contrastive loss and triplet loss respectively. When training with contrastive loss and triplet loss, the margins are discretely tuned from $\{1, 2, 3, 4\}$, and the ones providing best performance are used for comparison. The receiver operating characteristic (ROC) curves are shown in Figure 5.11.



(a) UBIPr

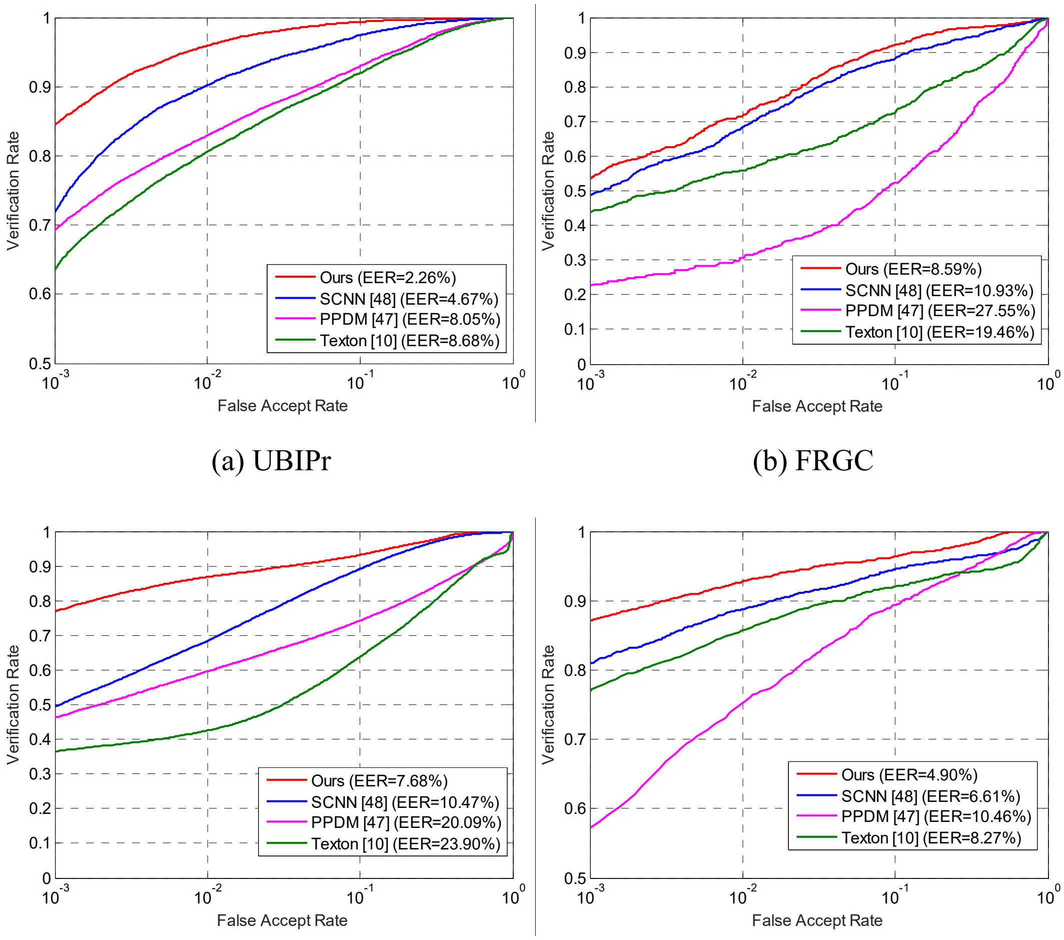
(b) FRGC

(c) FOCS

(d) CASIA.v4-distance

Figure 5.11: ROCs of training *AttNet* with *DSC* loss and conventional losses on four employed databases. The parameters of DSC loss are empirically set to a=10 and b=5; margins for contrastive loss and triplet loss are tuned among {1, 2, 3, 4} and the best performing ones are used here for comparison, which are m=3 and m'=4.

It can be observed that *DSC* loss delivers noticeable and consistent improvements over the other two losses, especially for lower false acceptance rates (FAR). The performance at low FAR is regarded as more important for biometric verification systems, and the key factor to this metric is the ability to verify challenging cases, *i.e.*, highly dissimilar genuine pairs and similar imposter pairs. The superiority of *DSC* loss is mainly attributed to the marginal effects for both positive and negative pair samples during the feature learning process, such that more training efforts can be put into challenging cases.



(a) UBIPr

(b) FRGC

(c) FOCS

(d) CASIA.v4-distance

Figure 5.12: ROC curves of the periocular verification using our method and comparison with other state-of-the-art methods on different databases.

B. Comparison with State-of-the-art Works

As discussed earlier, the performance of the proposed approach has been comparatively evaluated with state-of-the-art methods [10], [47] and [48] in the literature for the open-world setting. The resulting ROC curves are provided in Figure 5.12. We can observe from these results that our method consistently outperforms the other three baseline methods on all of the four employed databases. It is important to note that the advancements from our method are particularly significant at lower FAR, which indicates the outstanding capability of our method for verifying

Table 5.4: Results of significance test for comparison of our method and [48]. *p*-value indicates the probability of the null hypothesis, *i.e.*, two sets of data do not differ significantly.

Comparison with SCNN	UBIPr	FRGC	FOCS	CASIA.v4-distance
z-statistic	14.323	3.859	25.259	8.829
p-value*	$<10^{-4}$	1.14×10^{-4}	$<10^{-4}$	$<10^{-4}$

* *p* will be denoted as $<10^{-4}$ if the computed *z* is too large such that the corresponding *p* is too small for the computer to return the exact value.

challenging periocular samples. Even under the challenging cross-database training and test protocol, the proposed method has exhibited high level of robustness. The promising results from the proposed attention based model have further validated the importance of eyebrow and eye regions for the periocular recognition. We would like to specially clarify that the performance from one of the baselines PPDM [47] on FRGC is lower than what was reported in their original paper due to the difference in experimental settings. In the original setup in [47], the periocular regions were *manually* cropped from the face images in FRGC, and only single session data was used for the experiments. In our experiments, as introduced earlier, the periocular regions were automatically segmented and data from multiple sessions was selected for matching. Such configuration is highly desirable, closer to reality but introduces noticeable scale variation and misalighment for the data, which violates the patch correspondce and is the main reason for the performance degradation of PPDM. Furthermore, we adopted all-to-all verification protocol which is believed to be more challenging than their employed gallery-probe protocol.

We have also performed significance tests to ascertain the statistical significance of the improvements from our method. The method for the significance test is described in [83], which is based on the area under the curve (AUC) of the ROC statistics. Comparison has been made with [48] only, as this method delivers the best performance among the three baselines. The results from the tests are provided in Table

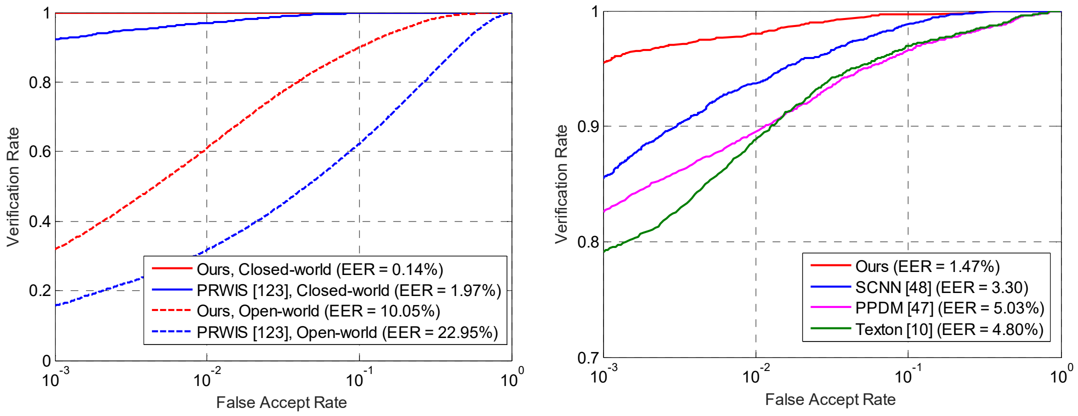
5.4. It can be inferred that, with widely used confidence level of 95%, the improvements from our method are statistically significant over its competitors.

5.5.5 Closed-World Performance

As discussed earlier, the proposed approach is mainly designed for open-world verification problem. However, some recent methods/competitions also adopt or focus on closed-world setting, in which all the subjects to be recognized are known during training/development phase, and it is usually allowed to use the gallery set for the training process. Typical examples include [118][119][123]. Despite the closed-world setting is less challenging, it is feasible for some applications to know all the interested subjects in advance during training phase, such as watchlist system. Hence, we supplement experiments under the closed-world configuration, which were conducted on UBIRIS.v2 and VISOB databases.

Under the closed-world setting, we maintained the architecture of *AttNet* but trained it in a different way. Similar to [123], we added a softmax layer after the feature layer (fc5 in Figure 5.2) with N_C output neurons, where N_C is the number of classes (subjects) to be recognized. As closed-world setting is applied, N_C is consistent during training and test phases. Each output neuron at the softmax layer is regarded as the probability that the current sample belongs to a specific subject, and therefore is used as a verification score. Figure 5.13 provides ROCs for the verification results on UBIRIS.v2 with comparison to [123], and on VISOB with comparison to [10], [47] and [48]. Note that for experiments on UBIRIS.v2, we also attached open-world results for comparative study. To obtain the comparative open-world results from [123], we used the l^2-norm distance between the feature vectors from fc7 layer as suggested in their paper.

From the results on UBIRIS.v2, we can observe that our approach consistently



(a) UBIRIS.v2 (open-/closed-world)

(b) VISOB (closed-world)

Figure 5.13: ROC curves on UBIRIS.v2 database and VISOB database (iphone-day-light-short subset). Note that the *AttNet* result under closed-world setting on UBIRIS.v2 is close to line $y = 1$.

outperforms the recently published state-of-the-art method [123]. Under the closed-world settings, our results have scored significantly high accuracy (0.14% EER), due to reason that class-specific recognition has been learned with softmax loss function for given and fixed set of subjects (and same for the baseline method). In contrast, when switched to open-world setting, both [123] and our method suffer from obvious performance degradation, which reflects that open-world problem inherently brings more challenges compared with the closed-world problem. However, our appraoch can still achieve superior results over that from [123].

The results on VISOB dataset reveal that our method still consistently outperforms other methods investigated in this paper. It should be noted that the eye images in this database do not include the eyebrow region, and the eye region occupies most the image area (Figure 5.10f). This implies that the proposed visual attention mechanism may not benefit much the recognition performance. Figure 5.14 visulizes the intermediate features learned by *AttNet* on such data, from which we can observe that enhancing attention within the eye region does not affect much the feature contents. In this case, *AttNet* can serve as a common CNN for backing up the perfomance even if

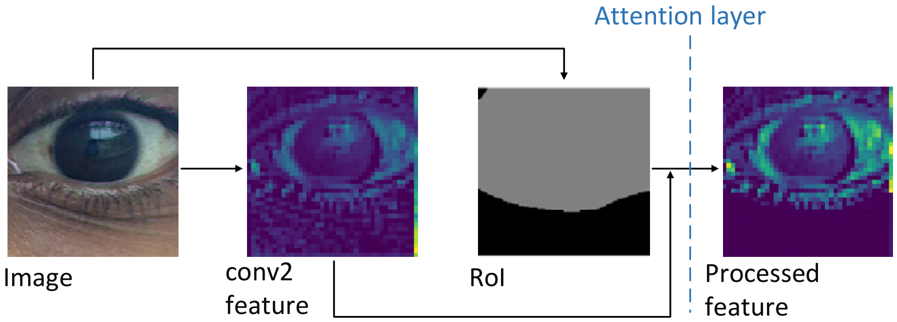


Figure 5.14: Visualization of convolutional features on a VISOB image which does not contain eyebrow. In this case the attention mechanism does not much impact on the feature distribution, and *AttNet* will basically act like a common CNN to guarantee fundamental performance.

desired regions are absent or can not be correctly segmented. Another aspect to notice is that, as mentioned before, we have only acquired the Visit-1 subset (550 subjects) for this dataset rather than the Visit-2 (290 subjects) which was used for benchmarking in [118], therefore it would be unfair to directly compare the results provided in this paper with those in [118].

5.6 Summary

This chapter has developed an attention based CNN architecture for more accurate and robust periocular recognition. The proposed framework includes *FCN-Peri*, which can accurately detect eyebrow and eye regions as key regions of interest, and *AttNet*, which makes use of the RoI information for more discriminative feature learning. A newly developed verification oriented loss function, referred to as *DSC* loss, has also been introduced in this work. The new loss function has shown to provide marginal effects for both positive and negative training samples during learning, which contributes to more robust feature representation for matching challenging periocular image pairs. Extensive experiments on four publicly available databases presented in Chapter 5.3 indicate that, the proposed attention based framework achieves significantly better results than several state-of-the-art methods for the periocular recognition. The effectiveness of the newly designed *DSC* loss function was also separately validated through comparison with conventional contrastive loss and triplet loss. The experimental results provide strong support to our assumption that, information within eyebrow and eye regions are critical to periocular recognition, and deserve more attention during feature learning and matching. The trained models and source for reproducing our experimental results are made publicly available via [116].

Despite success in simulating human visual attention model for the automated periocular recognition, as illustrated from promising results on multiple databases in this work, a lot more work needs to be done, *e.g.* to develop on-the-fly and more intelligent RoI learning through the feedback from the feature learning process, on the basis of pre-trained *FCN-Peri*. More robust and adequate visual attention mechanisms, in addition to the currently used feature adjustment strategy, is also expected to further improve the performance and therefore pursued in the future extension of this work. Last but not least, the separate impact from each of eyebrow and eye regions is another interesting and important aspect to investigate.

CHAPTER 6
Conclusions

This thesis has presented details of my research work on developing novel algorithms for accurate and reliable iris and periocular recognition, which aims at addressing the problem of less constrained human recognition. In this chapter we will draw conclusions on the contributions made by my research work, followed by contemporary limitations of the approaches developed in this thesis as well as future work for further facilitating the work on less constrained iris and periocular recognition.

6.1 Contributions

We firstly looked at the problem of accurate iris segmentation under relaxing conditions with visible and/or NIR spectrums, which is a critical primary step for the subsequent iris matching process. The key challenges lie in the existence of noise, occlusion, source reflection and other artifacts like glasses which severely degrade image quality. We established a novel relative total variation model with l^1 norm regularization, named as RTV-L^1, to remove the aforementioned degrading factors. The key advantage of RTV-L^1 is that it can suppress the noise and texture from the acquired eye images while preserving the salient structures, which is highly desired for accurate preliminary segmentation. We also developed a series of robust post-processing to refine the segmentation contours. The proposed approach significantly outperforms other state-of-the-art iris segmentation methods, especially for degraded eye images acquired under less constrained environments. In addition, the newly proposed RTV-L^1 is also expected to be useful for general computer vision tasks that involve noise removal and/or structure analysis.

We further stepped forward to investigating more effective and generalizable iris feature representation for matching irises more accurately. Having observed that traditional hand-crafted iris features suffer from heavy parameter tuning and low generalizability, we exploited deep neural networks for the iris feature learning. Based on our analysis that the discriminating information of iris pattern comes from local intensity distributions, we originally adopted fully convolutional network (FCN), instead of widely used convolutional neural network (CNN), for learning spatially corresponding iris features. We also designed a problem-specific loss function, i.e., extended triplet loss (ETL), to accommodate frequently observed occlusion and spatial translation for the iris matching. The proposed framework has delivered superior performance over popular and state-of-the-art methods in terms of matching accuracy and data generalizability on four publicly available database, including those for at-a-distance and non-ideal imaging scenarios.

Beside at-a-distance iris recognition, we also devoted significant research effort into periocular recognition with both visible and NIR spectrum, which is considered as another promising approach for addressing the focused problem of less constrained recognition. We firstly developed the semantics-assisted convolutional neural network (SCNN) which was inspired by human inference mechanism that combined high-level semantic information in the periocular images (*e.g.*, gender, side) for more comprehensive deep feature learning. Experimental results indicated that the supplement of such semantic information can help to recover more discriminative features than a usual CNN, especially when the training data is less sufficient. The proposed method has gained superior performances over state-of-the-art periocular recognition methods especially under cross-dataset evaluation.

Further beyond exploitation of semantic information, we proposed an attention based deep architecture for periocular recognition to simulate the visual classification system of human. Motivated by earlier studies that eye and eyebrow are of critical importance for identifying periocular, regional visual attention was incorporated into CNNs by emphasizing convolutional responses within detected eye and eyebrow regions. The learned features through such attention mechanism were observed to be more discriminative and stable. We also developed a verification oriented loss function, distance-driven sigmoid cross-entropy (DSC) loss, which provides better regularization for the training data than the traditional loss functions. This approach further boosted state-of-the-art performance dramatically for periocular recognition under varying less constrained situations.

An overview of the contributions from my research work has been illustrated in Figure 6.1. In summary, my research work on less-constrained at-a-distance iris segmentation, iris feature generation and periocular recognition has achieved significant superiority over traditional approaches and largely advanced state-of-the-art performance in these areas, therefore making solid contributions to achieving least-constrained human recognition. The key novelties of my research outcome lie in the more suitable formulation for the l^1-norm regularized relative total variation (RTV) model in dealing with noisy and degraded data, and exploitation of deep learning techniques facilitated with fully convolutional spatially corresponding features, semantics information as well as the visual attention based feature extraction mechanism. The above contributions are also expected to be generalizable to other common computer vision tasks which involve noise removal, local texture analysis, deep feature enhancement and so on.

6.2 Limitations and Future Work

As discussed earlier, solid contributions have been made in my research work to

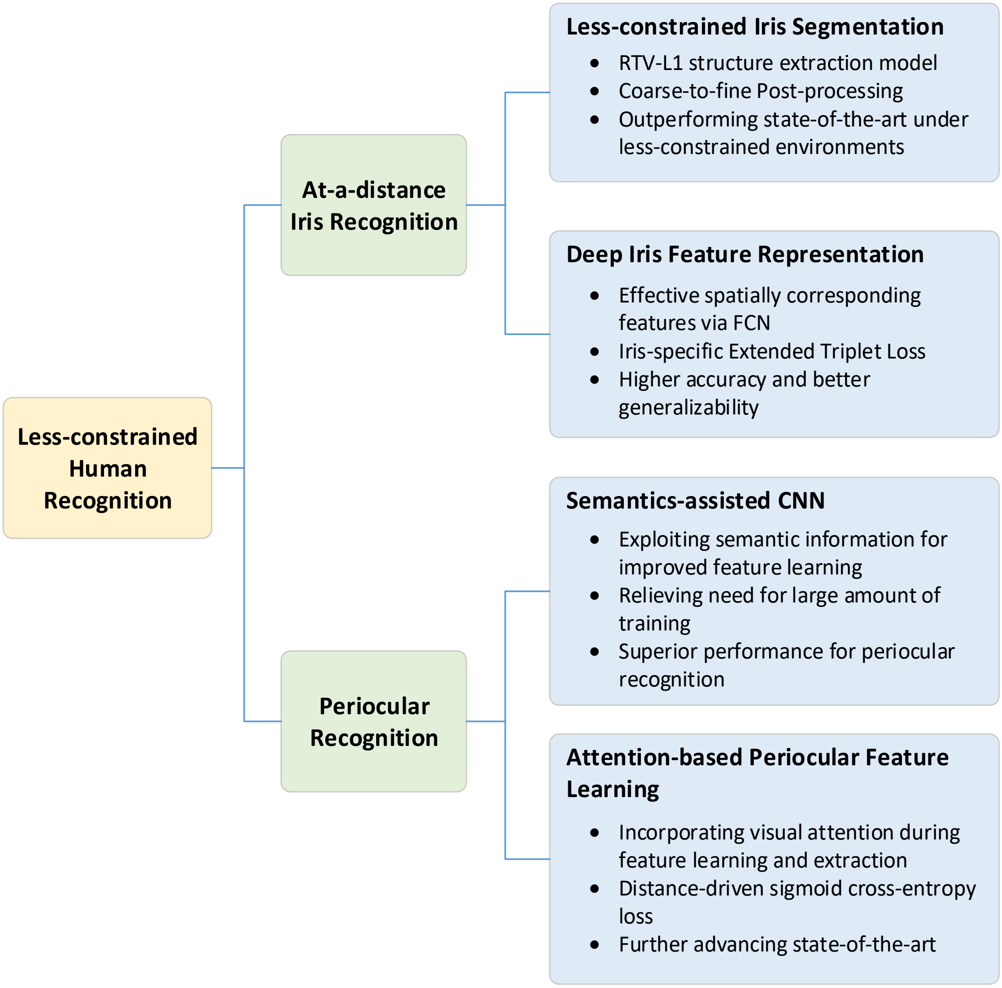

Less-constrained Human Recognition
At-a-distance Iris Recognition
Periocular Recognition
Less-constrained Iris Segmentation
• RTV-L1 structure extraction model
• Coarse-to-fine Post-processing
• Outperforming state-of-the-art under less-constrained environments
Deep Iris Feature Representation
• Effective spatially corresponding features via FCN
• Iris-specific Extended Triplet Loss
• Higher accuracy and better generalizability
Semantics-assisted CNN
• Exploiting semantic information for improved feature learning
• Relieving need for large amount of training
• Superior performance for periocular recognition
Attention-based Periocular Feature Learning
• Incorporating visual attention during feature learning and extraction
• Distance-driven sigmoid cross-entropy loss
• Further advancing state-of-the-art

Figure 6.1: Overview of research contributions delivered from this thesis.

advancing state-of-the-art for at-a-distance iris and periocular recognition in addressing least-constrained human recognition. However, there are still a number of limitations at the current stage of my research, which should be further addressed in the future extension of the work presented in this thesis. In particular, critical limitations and future directions for overcoming such bottlenecks are concluded as follows:

- The RTV-L^1 based iris segmentation framework described in Chapter 2 is much hand-crafted, relying on several parameters and involves certain ad-hoc operations. While such properties can save the applications from heavy training, they also decrease the generalizability and may lead to certain fragility for the approach. In the future we will extend the segmentation framework by unifying the hand-crafted parameters and ad-hoc operations into learnable architectures, such as FCN, to increase the robustness and adaptiveness of the iris segmentation. Special attention should be paid to control and model complexity and optimization algorithm so that it can learn from least amount of data for saving training efforts.

- The deeply learned spatially corresponding features for iris recognition presented in Chapter 3 are not fully end-to-end in terms of non-iris region detection (MaskNet) and the final binary feature encoding. It is indicated in the literature that end-to-end training is more desired to achieve higher performance and adaptiveness to the data for deep learning approaches [120]. We plan to develop end-to-end version of the spatial iris features by jointly optimizing FeatNet and MaskNet, where specific supervision mechanism for identifying effective iris region is required, and by designing learnable binary features on top of FeatNet, which may be similar to supervised discrete Hashing (SDH) [121]. Such measures are expected to further increase the feature performance and reliability for the more accurate and generalizable iris matching.
- The exploitation of explicit semantic information for learning more comprehensive periocular features as described in Chapter 4 is proven quite useful, but currently the mechanism for integrating identity-relevant semantic information, i.e., fusing features from two separate networks by joint-Bayesian, is quite trivial and less effective. This not only requires additional training efforts but may also lead to lower generalizability of the fused features. In the following work, we will investigate learning semantic information and features for identification simultaneously, in a fashion of multi-label learning [107] and/or

reinforcement learning [122], which should be able to further improve the level of integrity of the network and robustness of the learned features.

- The last important aspect to address lies in the attention-based periocular feature learning framework presented in Chapter 5. Currently the critical regions which attract more attention during feature learning is predefined as eye and eyebrow regions based on previous studies and human perception, and the corresponding detection model pre-trained separately. This may not be optimal and can also limit the network adaptiveness to the data. In the future we will explore more intelligent visual attention mechanism which automatically learns to discover salient regions to pay attention to via the feedback from output features, as well as improving attention implication, *e.g.,* learnable weighted sum of foreground/background features, for more adaptive and effective periocular feature learning.

As concluded above, despite the encouraging progresses gained by the research work introduced in this thesis, there are still considerably significant challenges towards more accurate at-a-distance iris and periocular recognition in addressing least-constrained human identification. We believe that with continuous effort devoted into developing more intelligent learning algorithms on the basis of my research studies, and with more insights into least-constrained human recognition, the research problem can be significantly solved in the near future.

BIBLIOGRAPHY

[1] A. K. Jain, A. Ross and S. Prabhakar, “An introduction to biometric recognition”, *IEEE Transactions on circuits and systems for video technology*, vol. 14, no. 1, pp. 4-20, 2004.

[2] A. K. Jain, P. Flynn and A. Ross, *Handbook of Biometrics*, Springer, 2007.

[3] A. K. Jain, K. Nandakumar and A. Ross, “50 years of Biometric Research: Accomplishments, Challenges, and Opportunities”, *Pattern Recognition Letters*, vol. 79, pp. 80-105, 2016.

[4] A. K. Jain, A. Ross, and K. Nandakumar, *Introduction to Biometrics*, Springer, 2011.

[5] M. J. Burge and K. W. Bowyer. *Handbook of Iris Recognition*. Springer, 2013.

[6] P. J. Grother, G. W. Quinn, J. R. Matey, M. L. Ngan, W. J. Salamon, G. P. Fiumara and C. I. Watson, “IREX III-Performance of iris identification algorithms”, *NIST Interagency/Internal Report (NISTIR)-7836*, pr. 2012.

[7] J. Daugman, “How iris recognition works”, *IEEE Transactions on circuits and systems for video technology*, vol. 14, no. 1, pp. 21–30, 2004.

[8] L. Masek, "Recognition of Human Iris Patterns for Biometric Identification." *The University of Western Australia* 2, 2003.

[9] Y. H. Li and M. Savvides, “An automatic iris occlusion estimation method based on high-dimensional density estimation,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 4, pp. 784-796, 2013.

[10] C. W. Tan and A. Kumar, “Towards online iris and periocular recognition under relaxed imaging constraints,” *IEEE Trans. Image Process.,* vol. 22, no. 10, pp. 3751-3765, 2013.

[11] H, Proença, “Iris recognition: On the segmentation of degraded images acquired in the visible wavelength,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 8, pp. 1502-1516, 2010.

[12] T. Tan, Z. He and Z. Sun, “Efficient and robust segmentation of noisy iris images for non-cooperative iris recognition,” *Image Vision Computing*, vol. 28, no. 2, pp. 223–230, 2010.

[13] C. W. Tan and A. Kumar, “Unified framework for automated iris segmentation using distantly acquired face images,” *IEEE Trans. Image Process*., vol. 21, no. 9, pp. 4068-4079, 2012.

[14] D. Monro, S. Rakshit and D. Zhang, "DCT-Based Iris Recognition", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 4, pp. 586-595, 2007.

[15] K. Miyazawa, K. Ito, T. Aoki, K. Kobayashi and H. Nakajima, "An Effective Approach for Iris Recognition Using Phase-Based Image Matching", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 10, pp. 1741-1756, 2008.

[16] Z. Sun and T. Tan, "Ordinal Measures for Iris Recognition", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 12, pp. 2211-2226, 2009.

[17] J. K. Pillai, V. M. Patel, R. Chellappa and N. K. Ratha, “Secure and robust iris recognition using random projections and sparse representations,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 9, pp. 1877-1893, 2011.

[18] H. Proenca, S. Filipe, R. Santos, J. Oliveira, and L. A. Alexandre, “The ubiris. v2: A database of visible wavelength iris images captured on-the-move and at-a-distance,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 8, pp. 1529-1535, 2010.

[19] E. R. Davies. “Circle and ellipse detection” in *Computer and Machine Vision, Fourth Edition: Theory, Algorithms, Practicalities*, Waltham: Academic Press, 2012.

[20] L. Xu, Q. Yan, Y. Xia and J. Jia, “Structure extraction from texture via relative total variation,” *ACM Trans. Graphics*, vol. 31, no. 6 (139), 2012.

[21] H. Proença and L. A. Alexandre, "The NICE. I: Noisy iris challenge evaluation-part I", in *Biometrics: Theory, Applications, and Systems (BTAS) 2007 First IEEE International Conference on*, 2007, pp. 1-4.

[22] H. Hofbauer, F. Alonso-Fernandez, P. Wild, J. Bigun and A. Uhl, “A ground truth for iris segmentation”. in *2014 IEEE 22nd International Conference on Pattern Recognition (ICPR)*, 2014, pp. 527-532.

[23] D. H. Brainard and B. A. Wandell, “Analysis of the retinex theory of color vision,” *J. Optical Soc. Am. A*, vol. 3, no. 10, pp. 1651-1661, 1986.

[24] L. Grady, “Random walks for image segmentation,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 11, pp. 1768-1783, 2006.

[25] L. I. Rudin, S Osher and E. Fatemi, “Nonlinear total variation based noise removal algorithms,” *Phys. D,* vol. 60, no. 1, pp. 259-268, 1992.

[26] A. Kumar and A. Passi, “Comparison and combination of iris matchers for reliable personal authentication,” *Pattern Recognition*, vol. 43, no. 3, pp. 1016-1026, 2010.

[27] Biometrics Ideal Test, CASIA.v4 database: http://www.idealtest.org/dbDetailForUser.do?id=4

[28] Face Recognition Grand Challenge, FRGC database: http://www.nist.gov/itl/iad/ig/frgc.cfm

[29] J. Zhang, R. Lai and C. C. Kuo,”Adaptive directional total-variation model for latent fingerprint segmentation,” *IEEE Trans. Info. Forensics & Security*, vol. 8, pp. 1261-1273, 2013.

[30] X. Bresson, S. Esedoglu, P Vandergheynst, J. P. Thiran and S. Osher, “Fast global minimization of the active contour/snake model,” *J. Math. Imaging & Vision*, vol. 28, no. 2, pp. 151-167, 2007.

[31] S. Alliney, “A property of the minimum vectors of a regularizing functional defined by means of the absolute norm,” *IEEE Trans. Sig. Process.*, vol. 45, no. 4, pp. 913-917, 1997.

[32] A. Kumar, T.-S. Chan, “Iris recognition using quaternionic sparse orientation code (QSOC),” *Proc. CVPR 2012*, pp. 59-64, CVPRW 2012, Providence, June 2012.

[33] A. Kumar, T.-S. Chan, C. W. Tan, “Human identification from at-a-distance face images using sparse representation of local iris features,” *Proc. ICB 2012*, pp. 303-309, Apr. 2012.

[34] Weblink to download implementation codes for the RTV-L^1 based iris segmentation framework, http://www.comp.polyu.edu.hk/~csajaykr/tvmiris.htm, 2015.

[35] M. Nokolova, “A variational approach to remove outliers and impulse noise,” *J. Math. Imaging & Vision*, vol. 20, pp. 99-120, 2004.

[36] T. H. Min and R. H. Park, “Comparison of eyelid and eyelash detection algorithms for performance improvement of iris recognition,” *Proc. ICIP*, pp. 257-260, San Diego, Oct. 2008.

[37] D. Menotti, G. Chiachia, A. Pinto, W. Robson Schwartz, H. Pedrini, A. Xavier Falcao and A. Rocha, "Deep Representations for Iris, Face, and Fingerprint Spoofing Detection", *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 4, pp. 864-879, 2015.

[38] A. Gangwar and A. Joshi, "DeepIrisNet: Deep Iris Representation with Applications in Iris Recognition and Cross-Sensor Iris Recognition", in *Image Processing (ICIP), 2016 IEEE International Conference on*, 2016, pp. 2301-2305.

[39] F. Alonso-Fernandez and J. Bigun, "A survey on periocular biometrics research", *Pattern Recognition Letters*, vol. 82, pp. 92-105, 2016, DOI: j.patrec.2015.08.026.

[40] A. Rattani and R. Derakhshani, "Ocular biometrics in the visible spectrum: A survey", *Image and Vision Computing*, vol. 59, pp. 1-16, 2017. DOI: 10.1016/j.imavis.2016.11.019.

[41] U. Park, A. Ross, and A. K. Jain, "Periocular biometrics in the visible spectrum: A feasibility study", in *2009 IEEE 3rd International Conference on Biometrics: Theory, Applications, and Systems (BTAS)*, 2009, pp. 1-6, DOI: 10.1109/BTAS.2009.5339068.

[42] S. Bharadwaj, H. S. Bhatt, M. Vatsa and R. Singh, "Periocular biometrics: When iris recognition fails", in *2010 IEEE 4th IEEE International Conference on Biometrics: Theory, Applications, and Systems (BTAS)*, 2010, pp. 1-6, DOI: 10.1109/BTAS.2010.5634498.

[43] C. N. Padole and H. Proenca. "Periocular recognition: Analysis of performance degradation factors." in *2012 5th IAPR International Conference on Biometrics (ICB)*, 2012, pp. 439-445, DOI: 10.1109/ICB.2012.6199790.

[44] G. Santos and H. Proenca, "Periocular biometrics: An emerging technology for unconstrained scenarios", in *2013 IEEE Workshop on Computational Intelligence in Biometrics and Identity Management (CIBIM)*, 2013, pp. 14-21, DOI: 10.1109/CIBIM.2013.6607908.

[45] L. Nie, A. Kumar and S. Zhan, "Periocular recognition using unsupervised convolutional RBM feature learning", in *2014 22nd International Conference on Pattern Recognition (ICPR)* , pp. 399-404, 2014, DOI: 10.1109/ICPR.2014.77.

[46] A. Sharma, S. Verma, M. Vatsa and R. Singh, "On cross spectral periocular recognition", in *2014 IEEE International Conference on Image Processing (ICIP)*, 2014, pp. 5007-5011, DOI: 10.1109/ICIP.2014.7026014

[47] J. Smereka, V. Boddeti and B. Vijaya Kumar, "Probabilistic deformation models for challenging periocular image verification", *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 9, pp. 1875-1890, 2015, DOI: 10.1109/TIFS.2015.2434271.

[48] Z. Zhao and A. Kumar, "Accurate periocular recognition under less constrained environment using semantics-assisted convolutional neural network", *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 5, pp. 1017-1030, 2017, DOI: 10.1109/TIFS.2016.2636093.

[49] J. Smereka, B. Vijaya Kumar and A. Rodriguez, “Selecting discriminative regions for periocular verification”, in *2016 IEEE International Conference on Identity, Security and Behavior Analysis (ISBA)*, 2016, pp. 1-8, DOI: 10.1109/ISBA.2016.7477247.

[50] P. Grother, G.W. Quinn, M. L. Ngan and J. R. Matey, “IREX IV: Part 1, Evaluation of Iris Identification Algorithms”, *NIST Interagency/Internal Report (NISTIR)-7949*, Jul. 2013.

[51] Z. Zhao and A. Kumar, "An Accurate Iris Segmentation Framework under Relaxed Imaging Constraints Using Total Variation Model." in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 3828-3836.

[52] Z. Zhao and A. Kumar, “Towards More Accurate Iris Recognition Using Deeply Learned Spatially Corresponding Features”, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV),* 2017, pp. 22-29.

[53] K. Hollingsworth, K. Bowyer and P. Flynn, "The Best Bits in an Iris Code", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 6, pp. 964-973, 2009.

[54] N. Othman, B. Dorizzi and S. Garcia-Salicetti, "OSIRIS: An Open Source Iris Recognition Software", *Pattern Recognition Letters*, vol. 82, pp. 124-131, 2016.

[55] VeriEye SDK 9.0: http://www.neurotechnology.com/verieye.html

[56] J. Long, E. Shelhamer and T. Darrell, "Fully Convolutional Networks for Semantic Segmentation", in *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, 2015, pp. 3431-3440.

[57] F. Schroff, K. Dmitry and P. James, "Facenet: A Unified Embedding for Face Recognition and Clustering", in *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, 2015.

[58] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going Deeper with Convolutions", in *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, 2015, pp. 1-9.

[59] R. Girshick, J. Donahue, T. Darrell and J. Malik, "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation", in *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, 2014, pp. 580-587.

[60] Y. Sun, X. Wang and X. Tang, "Deeply Learned Face Representations are Sparse, Selective, and Robust." in *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, 2015, pp. 2892-2900.

[61] Y. Sun, X. Wang and X. Tang, "Deep Learning Face Representation from Predicting 10,000 Classes", in *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, 2014, pp. 1891-1898.

[62] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep Face Recognition", in *British Machine Vision Conference (BMVC)*, vol. 1, no. 3, p. 6, 2015.

[63] A. Kumar and C. Kwong, “Towards contactless, low-cost and accurate 3D fingerprint identification”, in *2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3438-3443, 2013.

[64] X. Wu, Q. Zhao and W. Bu, “A SIFT-based contactless palmprint verification approach using iterative RANSAC and local palmprint descriptors”, *Pattern Recognition*, vol. 47, no. 10, pp.3314-3326, 2014.

[65] J. Daugman, "Iris Recognition Border-Crossing System in the UAE." *International Airport Review* 8, no. 2, 2004.

[66] J. Daugman, “600 Million Citizens of India are Now Enrolled with Biometric Id,” SPIE newsroom 7, 2014.

[67] NIST Presentation, “Forensic Data for Face & Iris”, http://biometrics.nist.gov/cs_links/standard/ansi-overview_2010/presentations/Forensic_data_for_Face_Iris.pdf

[68] H. King, "Galaxy Note 7 is First Samsung Device with Iris Scanner", *CNNMoney*, 2016. [Online]. Available: http://money.cnn.com/2016/08/02/technology/samsung-note-7/index.html. [Accessed: 09- Oct- 2016].

[69] L. Ma, T. Tan, Y. Wang and D. Zhang, "Efficient Iris Recognition by Characterizing Key Local Variations", *IEEE Transactions on Image Processing*, vol. 13, no. 6, pp. 739-750, 2004.

[70] ISO/IEC 19794-6:2011 (2011). Information technology -- Biometric Data Interchange Formats -- Part 6: Iris Image Data. Standard, International Organization for Standardization, Geneva, CH.

[71] Z. He, Z. Sun, T. Tan, X. Qiu, C. Zhong and W. Dong, "Boosting Ordinal Features for Accurate and Fast Iris Recognition", in *Computer Vision and Pattern Recognition (CVPR), 2008 IEEE Conference on*, 2008, pp. 1-8.

[72] J. Daugman, "The Importance of Being Random: Statistical Principles of Iris Recognition", *Pattern Recognition*, vol. 36, no. 2, pp. 279-291, 2003.

[73] K. Bowyer, and P. Flynn, "The ND-IRIS-0405 Iris Image Dataset", Notre Dame CVRL Technical Report, 2009.

[74] IITD Iris Database: http://www.comp.polyu.edu.hk/~csajaykr/IITD/Database_Iris.htm

[75] S. Crihalmeanu, A. Ross, S. Schuckers, L. Hornak, “A Protocol for Multibiometric Data Acquisition, Storage and Dissemination”, Technical Report, WVU, Lane Department of Computer Science and Electrical Engineering, 2007.

[76] OpenCV based face and eye detector: http://docs.opencv.org/trunk/d7/d8b/tutorial_py_face_detection.html

[77] F. He, Y. Han, H. Wang, J. Ji, Y. Liu and Z. Ma, "Deep Learning Architecture for Iris Recognition Based on Optimal Gabor Filters and Deep Belief Network", *Journal of Electronic Imaging*, vol. 26, no. 2, p. 023005, 2017.

[78] Web link to download the source code and executable files for the deep spatially corresponding iris feature descriptor:

http://www.comp.polyu.edu.hk/~csajaykr/deepiris.htm

[79] V. Nair and G. Hinton, "Rectified linear units improve restricted boltzmann machines", *Proc. 27th Intl. Conf. Machine Learning (ICML)*, 2010, pp. 807-814, 2010.

[80] Y. Lecun, L. Bottou, Y. Bengio and P. Haffner, "Gradient-based learning applied to document recognition", *Proc. IEEE*, vol. 86, no. 11, pp. 2278-2324, 1998.

[81] A. Krizhevsky, I. Sutskever, and G. Hinton, “Imagenet classification with deep convolutional neural networks”, In *Advances in neural information processing systems,* 2012, pp. 1097-1105.

[82] F. Juefei-Xu, K. Luu, M. Savvides, T. D. Bui and C. Y. Suen, "Investigating age invariant face recognition based on periocular biometrics." In *Biometrics (IJCB), 2011 International Joint Conference on*, 2011, pp. 1-7.

[83] E. DeLong, D. DeLong and D. Clarke-Pearson, "Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach", *Biometrics*, vol. 44, no. 3, p. 837, 1988.

[84] D. L. Woodard, S. Pundlik, P. Miller, R. Jillela, and A. Ross, "On the fusion of periocular and iris biometrics in non-ideal imagery", *Pattern Recognition (ICPR), 2010 20th IEEE International Conference on*, 2010, pp. 201-204.

[85] A. Kumar, "Neural network based defection of local textile defects," *Pattern Recognition,* vol. 36, pp. 1645-1659, July 2003

[86] R. Jillela and A. Ross. "Mitigating effects of plastic surgery: Fusing face and ocular biometrics", in *Biometrics: Theory, Applications and Systems (BTAS), 2012 IEEE Fifth International Conference on*, 2012, pp. 402-411.

[87] Y. Taigman, M. Yang, M. A. Ranzato and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification", in *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, 2014, pp. 1701-1708.

[88] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting", *The Journal of Machine Learning Research,* vol. 15, no. 1 pp. 1929-1958, 2014.

[89] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition", *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June, 2016.

[90] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database", in *Computer Vision and Pattern Recognition (CVPR), 2009 IEEE Conference on*, 2009, pp. 248-255.

[91] L. Wan, M. Zeiler, S. Zhang, Y. Lecun, and R. Fergus, “Regularization of neural networks using dropconnect”, *Proc. 30th Intl. Conf. on Machine Learning*, ICML 2013, 1058-1066, 2013.

[92] K. He, X. Zhang, S. Ren and J. Sun, "Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 9, pp. 1904-1916, 2015.

[93] ImageNet Large Scale Visual Recognition Challenge (ILSVRC): http://www.image-net.org/challenges/LSVRC/

[94] G. B. Huang, M. Ramesh, T. Berg and E. Learned-Miller, *Labeled faces in the wild: A database for studying face recognition in unconstrained environments*. vol. 1. no. 2, *Technical Report 07-49*, University of Massachusetts, Amherst, 2007.

[95] L. Wolf, T. Hassner and I. Maoz, “Face recognition in unconstrained videos with matched background similarity”, in *2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 529-534, 2011.

[96] G. Bradski, “The OpenCV Library,” Dr. Dobb’s Journal of Software Tools, 2000.

[97] P. Viola and M. Jones, “Rapid object detection using a boosted cascade of simple features”, in *Computer Vision and Pattern Recognition (CVPR), Proceedings of the 2001 IEEE Computer Society Conference on*, 2001, vol. 1, pp. I-511.

[98] D. Chen, X. Cao, L. Wang, F. Wen and J. Sun, "Bayesian face revisited: A joint formulation", in *Computer Vision-ECCV 2012*, Springer Berlin Heidelberg, 2012, pp. 566-579.

[99] "FOCS dataset", 2016. [Online]. Available: http://www.nist.gov/itl/iad/ig/focs.cfm. [Accessed: 29- Mar- 2016].

[100] "UBIpr dataset", 2016. [Online]. Available: http://socia-lab.di.ubi.pt/~ubipr/. [Accessed: 29- Mar- 2016].

[101] "CIFAR-10 dataset", 2016. [Online]. Available: https://www.cs.toronto.edu/~kriz/cifar.html. [Accessed: 29- Mar- 2016].

[102] "cuda-convnet - High-performance C++/CUDA implementation of convolutional neural networks - Google Project Hosting", Code.google.com, 2016. [Online]. Available: https://code.google.com/p/cuda-convnet/. [Accessed: 29- Mar- 2016].

[103] "Caffe | CIFAR-10 tutorial", Caffe.berkeleyvision.org, 2016. [Online]. Available: http://caffe.berkeleyvision.org/gathered/examples/cifar10.html. [Accessed: 29- Mar- 2016].

[104] Weblink to download codes for SCNN for periocular recognition, http://www.comp.polyu.edu.hk/~csajaykr/scnn.rar

[105] "Biometric Evaluations Homepage", *Nist.gov*, 2016. [Online]. Available: http://www.nist.gov/itl/iad/ig/biometric_evaluations.cfm. [Accessed: 29-May-2016].

[106] "Stock Photo - epa01034158 Demonstrators cover their faces during clashes at the end of the far leftist 'No Bush-No War' rally", *Alamy*, 2016. [Online]. Available: http://www.alamy.com/stock-photo-epa01034158-demonstrators-cover-their-faces-during-clashes-at-the-97583185.html. [Accessed: 30-May-2016].

[107] G. Tsoumakas and I. Katakis, "Multi-Label Classification: An Overview", *International Journal of Data Warehousing and Mining*, vol. 3, no. 3, pp. 1-13, 2007.

[108] Z. Zhao and A. Kumar, “Improving Periocular Recognition by Explicit Attention to Critical Regions in Deep Neural Network”, *IEEE Transactions on Information Forensics and Security (T-IFS)*, vol. 13, no.12, pp. 2937-2952, 2017.

[109] S. Ren, K. He, R. Girshick and J. Sun, “Faster R-CNN: Towards real-time object detection with region proposal networks”, in *Advances in neural information processing systems (NIPS) 2015,* 2015, pp. 91-99.

[110] G. Huang, Z. Liu, M. Laurens, W. Kilian Q, “Densely connected convolutional networks”, in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* 2017, pp. 4700-4708, DOI: 10.1109/CVPR.2017.243.

[111] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel and

[112] Bengio, “Show, attend and tell: Neural image caption generation with visual attention”, in *Proceedings of the 32nd International Conference on Machine Learning* (*ICML*), 2015, pp. 2048-2057.

[112] V. Mnih, N. Heess and A. Graves, “Recurrent models of visual attention”, in *Advances in Neural Information Processing Systems (NIPS) 2014,* 2014, pp. 2204-2212.

[113] T. Xiao, Y. Xu, K. Yang, J. Zhang, Y. Peng and Z. Zhang, “The application of two-level attention models in deep convolutional neural network for fine-grained image classification”, in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* 2015*,* pp. 842-850, DOI: 10.1109/CVPR.2015.7298685.

[114] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang and X. Tang, "Residual attention network for image classification", in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 6450-6458, DOI: 10.1109/CVPR.2017.683.

[115] S. Chopra, R. Hadsell and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification", in *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005, vol. 1 pp. 539-546, DOI: 10.1109/CVPR.2005.202.

[116] Weblink to download source codes and trained models for attention based network for periocular recognition, http://www.comp.polyu.edu.hk/~csajaykr/attnet.htm.

[117] S. Bakshi, P. K. Sa, H. Wang, S. S. Barpanda and B. Majhi, "Fast periocular authentication in handheld devices with reduced phase intensive local pattern," *Multimedia Tools and Applications*, 2017, DOI: 10.1007/s11042-017-4965-6.

[118] A. Rattani, R Derakhshani, S. K. Saripalle and V. Gottemukkula, "ICIP 2016 competition on mobile ocular biometric recognition", in *2016 IEEE International Conference on Image Processing (ICIP)*, 2016, pp. 320-324, DOI: DOI: 10.1109/ICIP.2016.7532371.

[119] R. Raghavendra and C. Busch, "Learning deeply coupled autoencoders for smartphone based robust periocular verification", in *2016 IEEE International Conference on Image Processing (ICIP),* 2016, pp. 325-329, DOI: 10.1109/ICIP.2016.7532372.

[120] S. Sukhbaatar, J. Weston and R. Fergus, "End-to-end memory networks", in *Advances in Neural Information Processing Systems (NIPS) 2015,* 2015, pp. 2440-2448.

[121] F. Shen, C. Shen, W. Liu and H. Tao Shen, "Supervised discrete hashing", in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* 2015, pp. 37-45.

[122] V. Mnih, A. P. Badia, M. Mirza, A. Graves, T. Lillicrap, T. Harley, D. Silver and K. Kavukcuoglu, "Asynchronous methods for deep reinforcement learning", in *International Conference on Machine Learning (ICML)*, 2016, pp. 1928-1937.

[123] H. Proença and J. C. Neves, "Deep-PRWIS: Periocular Recognition Without the Iris and Sclera Using Deep Learning Frameworks", *IEEE Transactions on Information Forensics and Security,* vol. 13, no. 4, pp. 888-896, 2018, DOI: 10.1109/TIFS.2017.2771230.

[124] J. M. Smereka, and B. V. Kumar, “What Is a ‘Good’ Periocular Region for Recognition?” In *2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW),* 2013, pp. 117-124, DOI: 10.1109/CVPRW.2013.25.

[125] H. Zhu, M. Long, J. Wang and Y. Cao, “Deep Hashing Network for Efficient Similarity Retrieval”, In *2016 Association for the Advancement of Artificial Intelligence Conference* (*AAAI),* 2016, pp. 2415-2421.

[126] Z. Cao, M. Long, J. Wang and P. S. Yu, “Hashnet: Deep learning to hash by continuation”, in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017.

[127] P. E. Miller, A. W. Rawls, S. J. Pundlik and D. L. Woodard, "Personal identification using periocular skin texture." In *Proceedings of the 2010 ACM Symposium on Applied Computing*, 2010, pp. 1496-1500.

[128] D. L. Woodard, S. J. Pundlik, J. R. Lyle and P. E. Miller. "Periocular region appearance cues for biometric identification." In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2010, pp. 162-169.

[129] J. R. Lyle, P. E. Miller, S. J. Pundlik and D. L. Woodard, "Soft biometric classification using periocular region features." In *2010 Fourth IEEE International Conference on Biometrics: Theory Applications and Systems (BTAS)*, 2010, pp. 1-7, 2010.