B. TECH. PROJECT REPORT

On

Gender Classification using Near-Infrared Periocular Images

BY

Manyala Anirudh

DISCIPLINE OF ELECTRICAL ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY INDORE

December 2016

Gender Classification using Near-Infrared Periocular Images

A PROJECT REPORT

Submitted in partial fulfillment of the requirements for the award of the degrees

of
BACHELOR OF TECHNOLOGY
in

ELECTRICAL ENGINEERING

Submitted by:
Manyala Anirudh

Guided by:
Dr. Vivek Kanhangad (Assistant Prof, IIT INDORE)
Dr. Deepu Rajan (Associate Prof, NTU, Singapore)

INDIAN INSTITUTE OF TECHNOLOGY INDORE
December 2016

CANDIDATE'S DECLARATION

I hereby declare that the project entitled **"Gender Classification Using Near-Infrared Periocular Images"** submitted in partial fulfillment for the award of the degree of Bachelor of Technology in 'Electrical Engineering' completed under the supervision of **Dr. Vivek Kanhangad, Asst Prof, Electrical Engg, IIT INDORE and Dr. Deepu Rajan, Assoc Prof, School of Computer Science and Engg, NTU, Singapore** is an authentic work.

Further, I declare that I have not submitted this work for the award of any other degree elsewhere.

Signature and name of the student(s) with date

MANYALA ANIRUDH, 4/11/16

CERTIFICATE by BTP Guide(s)

It is certified that the above statement made by the students is correct to the best of our knowledge.

Signature of BTP Guide(s) with dates and their designation

Dr. Vivek Kanhangad
Assistant Prof Electrical Engg
IIT INDORE, 4/11/16

Dr. Deepu Rajan
Associate Prof, SCSE
NTU, Singapore, 4/11/16

Preface

This report on “Gender Classification Using Near Infrared Periocular Images" is prepared under the guidance of Dr. Vivek Kanhangad and Dr. Deepu Rajan. Through this report I tried to give a detailed (theoretical and experimental) results on employing different features for the gender classification task on three different databases. This is also the first work to explore convolutional neural networks in the area of gender classification using near-infrared periocular images.

I have tried to best of my abilities and knowledge to explain the content in more informative, illustrative and lucid manner. I also added appropriate figures and the methods used in the experimental setup for easy understanding of the reader.

Manyala Anirudh
B.Tech. IV Year
Discipline of Electrical Engineering
IIT Indore

Acknowledgements

I am obliged to Dr. Vivek Kanhangad and Dr. Deepu Rajan for their kind support and valuable guidance. It is their help and support, due to which I became able to complete the design and technical report.

I would like to thank my family and Dr. Chowdari, Senior Executive Director, President's Office, NTU for providing me an opportunity to work in NTU, Singapore.

I would also like to acknowledge Mr. Hisham Cholakkal, PhD candidate, School of Computer Science and Engineering, NTU, Singapore, Mr. Jubin Johnson, PhD candidate, School of Computer Science and Engineering, NTU, Singapore for their valuable support and help at critical stage of the project.

I also take this opportunity to thank Miss. Sushree Sangeeta, M.Tech student at IIT INDORE for helping me in the collection of the database and Mr. Suneel Kumar Telagamsetti, PhD candidate, Electrical Engineering, IIT INDORE and my friends Chaitanya, Manideep, Prem, Ramlakhan, Veerendra and Yeshwanth for helping me out with the application and other procedures for the international internship.

Manyala Anirudh
B.Tech. IV Year
Discipline of Electrical Engineering
IIT Indore

Abstract

Periocular region has emerged as a key biometric trait with potential applications in the forensics domain.

In this report, I present an approach for gender classification using near-infrared images of the periocular region. The proposed approach involves detection and extraction of left and right periocular regions. This is followed by extraction of features using a deep convolutional neural network (CNN). A trained support vector machine (SVM) classifier utilizes these features to predict the gender information. Performance evaluations have been carried out on three databases, which includes an in-house and two public databases. Local binary pattern and histogram of oriented gradient based methods have been used as baseline methods to ascertain the effectiveness of the proposed approach.

Our results indicate that the proposed approach achieves higher classification accuracy than the two baseline methods on a database which contains a large number of non-ideal images. On the other two databases, the proposed approach compares favourably with the baseline methods. Further, accuracy of the proposed approach is consistently higher than the existing eyebrow feature based method. As part of the additional experiments, I also evaluated our proposed method on periocular images without considering eyebrows in it. Our experiment results shows that eyebrows play a crucial role in predicting the gender from periocular images.

Table of Contents

List of Figures

List of Tables

CHAPTER 1

1. INTRODUCTION

Periocular biometrics has been studied extensively in recent years. In addition to civilian applications (e.g., access control), it plays an indispensable role in surveillance applications that
aid law enforcement agencies. Apart from utilizing the periocular region as a primary characteristic for biometric recognition, researchers (Merkow et al., 2010; Lyle et al., 2012) have shown that gender information can be extracted from periocular images acquired in visible (VIS) and near-infrared (NIR) spectra. Researchers (Jain et al., 2004) have shown that soft biometrics such as gender, age and ethnicity complement primary biometric traits and incorporating this ancillary information leads to significant improvement in recognition performance of the biometric system.

In particular, gender is a key soft biometric and gender classification using NIR images of the periocular region has several potential applications, including crime scene investigation by the law enforcement agencies. Examination of surveillance video from the crime scene is an essential part of the investigation to identify the offender. However, offenders often wear mask or other face covering devices to hide their identities leaving only a small area around the eyes exposed, as shown in Fig. 1.

In such scenarios, investigation agencies can rely on the periocular region to identify the offender as face and iris based identification are likely to fail. The gender information extracted from NIR periocular images can be utilized in the identification process in two ways. Firstly, it can be fused with primary characteristics to enhance the recognition performance

Fig. 1. Offenders wearing mask to hide their identity

and secondly, gender information can be used for indexing large biometric databases to reduce the computational burden and thereby speeding up the identification process. Despite its significance and potential applications, there has been very little work on gender classification using NIR images of the periocular region. The goal of this work is to address this issue in periocular biometrics. I have focused on NIR images because cameras used for night time surveillance employ NIR illuminators to capture images of the scene in low-light and no-light conditions.

Contributions of this work can be summarized as follows. Firstly, I have explored a deep feature-based approach for gender classification using NIR images of the periocular region.
Secondly, I have studied the performance of our approach in a scenario where eyebrow information is not available for gender classification. Thirdly, I have evaluated the performance of the proposed approach on an in-house and two publicly available databases and compared the results with the state-of-the-art methods. I also plan to release our database to the research community.

CHAPTER 2

2. Literature Review

In the literature, several methods have been proposed for extraction of gender information from primary biometric characteristics such as face (Wang et al., 2010), iris (Thomas et al., 2007), voice (Shafey et al., 2014), gait (Lee and Grimson, 2002), fingerprint (Rattani et al., 2014) and hand (Amayeh et al., 2008). A recent survey on soft biometrics can be found in (Dantcheva et al., 2015).

The problem of gender estimation from face images has received a lot of attention as humans naturally rely on facial features for gender recognition. Apart from the images captured in the visible spectrum, researchers have explored gender estimation from thermal and NIR face images. Ross and Chen (2011) investigated the feasibility of using NIR face images for gender classification. Their approach employs principal component analysis (PCA) based features and SVM classifier. Their experimental results indicate that NIR face images contain discriminatory information for gender classification. However, the performance is found to be inferior to VIS face image based gender classification. The authors also explored local binary pattern (LBP) based features for gender classification using NIR and thermal face images (Chen and Ross, 2011).

Ever since Park et al. (2009) explored the use of periocular region as a biometric trait, periocular biometrics has received increased attention from researchers. A detailed survey on periocular biometrics can be found in (Alonso-Fernandez and Bigun, 2015). Much of the work in this area has focused on personal recognition using the primary characteristic. Specifically, there has been very little work on estimating gender from periocular images. Merkow et al. (2010) explored the feasibility of performing gender classification using VIS images of the periocular region. The authors employed LBP for feature extraction and their approach achieves classification accuracy of more than 85% on a database of 936 VIS face images retrieved from the web. Their results suggest that periocular region carries discriminatory information for gender classification and that the performance is slightly lower than gender classification using full face images. Gender classification using VIS images of the periocular region has also been explored by Lyle et al. (2012).

Pixel intensities and LBP features are utilized for discrimination in their approach, which has been evaluated on a set of periocular images obtained from FRGC face database (Phillips et al., 2005). Another interesting work in the area of gender recognition using VIS periocular images is reported in (Castrill´on-Santana et al., 2015). The authors have studied the performance of the periocular region for gender classification on a challenging dataset, which contains 14385 face images in the wild . A set of local descriptors generated using

histogram of oriented gradients (HOG), LBP, local ternary pattern (LTP) and Weber local descriptor (WLD) have been used for feature representation.

The authors have also proposed a score-level fusion of local descriptors for enhancing the classification performance.

Gender classification using NIR images of the periocular region has been explored in (Lyle et al., 2012; Dong and Woodard, 2011). Lyle et al. (2012) have investigated the effectiveness of LBP, HOG, discrete cosine transform (DCT) and local color histogram based features. The authors have also explored artificial neural network (ANN) and SVM for classification. The performance of periocular (with eye region masked out) and eye regions have been evaluated separately. Subsequently, fusion of periocular and eye regions has been performed to improve the classification performance. Performance evaluations have been carried out on a dataset of 350 NIR periocular images, which are obtained by excluding non-ideal frames in MBGC portal video recordings (Phillips et al., 2009).

Dong and Woodard (2011) have proposed an approach for gender classification using eyebrow shape features. The focus of their work is on images captured in non-ideal conditions. The feature vector consists of a set of geometric features extracted from the eyebrow region. For performance evaluation, a dataset of 922 NIR images is extracted from MBGC portal video recordings (Phillips et al., 2009). Their approach achieves 96% accuracy for gender classification using NIR periocular images. The method involves manual segmentation of the eyebrow region and is expected to work only when the periocular image contains the entire eyebrow.

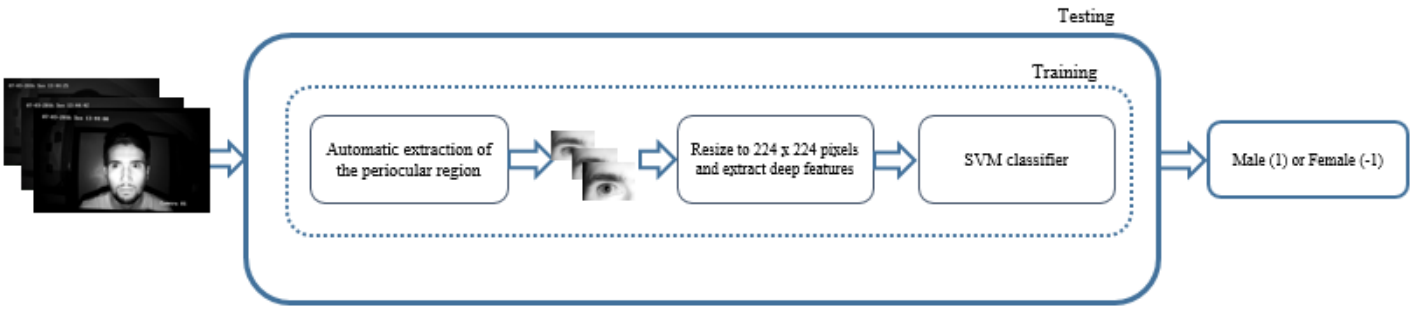

Fig. 2 Block diagram of the proposed approach

CHAPTER 3

3. System Overview

3.1 Pre-Requisites

This work is performed on Matlab 2014 software and some partial work is done Matlab 2015b version. For employing CNN's, I installed the required MatConvNet. It is a Matlab tool box useful for implementing CNNs for computer vision application. The required VGG-Face and VGG-Very deep 16 (VD) pre-trained models can be found in the MatConvNet webpage. All the tutorials and basic codes for the CNN can be found in their website.

3.2 Approach

Block diagram of the proposed approach for gender classification is shown in Fig. 2. Our approach involves automated extraction of the periocular region from the input image, which generally contains a full face and other parts of the body. Therefore, the extraction of periocular region is performed in a hierarchical manner. Specifically, the input image is processed to detect a face, which is followed by detection of eye-pair in the detected face. This hierarchical process is expected to reduce the number of false positives in eye-pair detection. In this work, we have employed Viola-Jones algorithm (Viola and Jones, 2004) (VJ)for face and eye-pair detection. Once an eye-pair is detected, a rectangular region-of-interest (ROI) around the eyes is extracted as shown in Fig. 3. The Viola-Jones algorithm can be implemented in Matlab using vision.CascadeObjectDetecotr() function. One can obtain the bounding box for almost any image by changing the value of merge threshold in the function.

Fig. 3 Bounding box obtained from the VJ algorithm

The ROI size is determined based on the dimensions of the bounding box produced by the eye-pair detector. Specifically, it is (w+ w/a; h+ h/b) pixels, where w and h are the width and the height of the bounding box, respectively. The parameters a and b in the fractions can be set empirically so that extracted periocular

images contain eyebrows. The resultant periocular image, which contains both eyes as well as the surrounding area, is separated into left and right periocular images by splitting the image vertically into two halves. The steps involved in the automated extraction of periocular images is illustrated in Fig. 4.
Feature extraction from the left or right periocular image is performed using a deep CNN. The resulting feature vector is fed to a SVM, which classifies the periocular region as belonging
to male or female class. The following section provides details of the deep CNN employed for feature extraction in this work.

3.3 Pre-trained CNNs

Generally, while employing a deep CNN for a certain application, one can build a network from scratch, fine-tune an existing model or employ a pre-trained model. Sharif Razavian et al. (2014) explored effectiveness of generic image descriptors extracted from a pre-trained CNN for diverse recognition tasks. Interestingly, extensive experimental results presented in (Sharif Razavian et al., 2014) indicate that features extracted from an off-the-shelf deep CNN provides state-of-the-art performance for all the tasks. Ozbulak et al. (2016) compared a domain specific CNN with another model, which was trained specifically for the task of gender classification from face images.
Their results suggest that domain specific model outperforms the task specific CNN, mainly due to limited amount of labelled data available for training the latter.
The results discussed above led us to choose a pre-trained model namely, VGG-Face (Parkhi et al., 2015) for feature extraction in this work. Since the periocular region is a part of the human face, VGG-Face, which has been trained on face images, can be considered as a domain specific model for our application. Before going into the details of the architecture of VGG-Face model, the following chapter describes about the basic terminologies in CNNs area and about the other features used in this work.

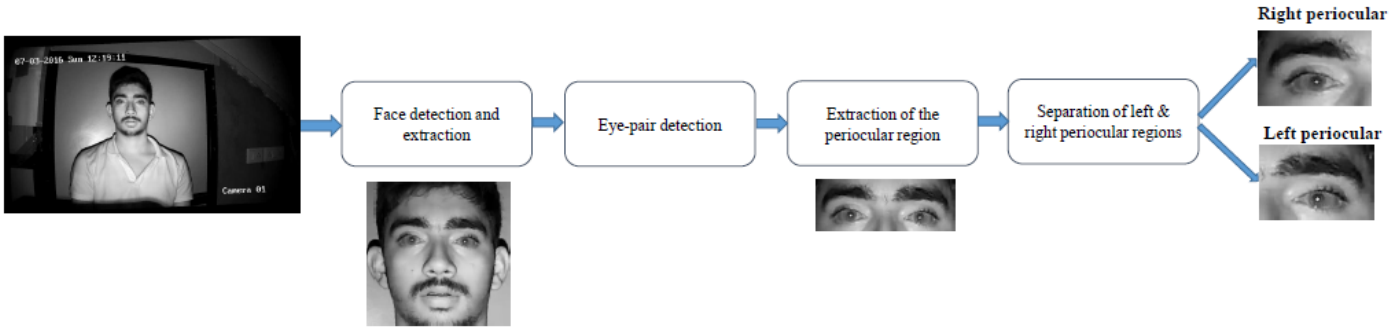

Fig. 4 Automated extraction of left and right periocular images

CHAPTER 4

4. Overview of CNN structure

Generally CNNs takes the input as an image and passes it through a series of convolutional, nonlinear, pooling (downsampling), and fully connected layers to get an output. The output can be a single class or a probability of classes that best describes the image.

4.1 Convolutional Layer

The first layer in a CNN is always a Convolutional (conv) Layer. The input to the first convolutional layer is the input image. If we consider an input image of 32 x 32 x 3 array of pixel values. The conv layer can be explained by imagining a flashlight that is shining over the top left of the image. Let's say that the light this flashlight shines covers a 5 x 5 area. And now, let's imagine this flashlight sliding across all the areas of the input image. In machine learning terms, this flashlight is called a filter (or sometimes referred to as a neuron or a kernel) and the region that it is shining over is called the **receptive field.** Now this filter is also an array of numbers (the numbers are called weights or parameters). A very important note is that the depth of this filter has to be the same as the depth of the input so the dimensions of this filter is 5 x 5 x 3. Now, let's take the first position the filter is in for example. It would be the top left corner. As the filter is sliding, or convolving, around the input image, it is multiplying the values in the filter with the original pixel values of the image.These multiplications are all summed up as we have a spatial resolution of 5x5 and a depth of 3 we will obtain 5x5x3= 75 multiplications. This number is just representative of when the filter is at the top left of the image. This process is repeated for every location on the input volume by sliding the filter with a stride. Every unique location on the input volume produces a number. After sliding the filter over all the locations, we will find out that what we're left with is a 28 x 28 x 1 array of numbers, which we call an **activation map or feature map.** The reason you get a 28 x 28 array is that there are 784 different locations that a 5 x 5 filter can fit on a 32 x 32 input image. These 784 numbers are mapped to a 28 x 28 array.

Let's say now we use two 5 x 5 x 3 filters instead of one. Then our output volume would be 28 x 28 x 2. By using more filters, we are able to preserve the spatial dimensions better. Mathematically, this is the overview of a convolutional layer.

Now in a traditional convolutional neural network architecture, there are other layers that are interspersed between these conv layers. In general sense, they provide nonlinearities and preservation of dimension that

help to improve the robustness of the network and control overfitting. A classic CNN architecture would look like this.

Input -> Conv -> ReLU -> Conv -> ReLU -> Pool -> ReLU -> Conv -> ReLU -> Pool ->Fully Connected

The filters in the first conv layer are designed to detect. They detect low level features such as edges and curves. As one would imagine, in order to predict whether an image is a type of object, we need the network to be able to recognize higher level features such as hands or paws or ears for a dog. Therefore, let's think about what the output of the network is after the first conv layer. It would be a 28 x 28 x 3 volume (assuming we use three 5 x 5 x 3 filters). When we go through another conv layer, the output of the first conv layer becomes the input of the 2nd conv layer. When we were talking about the first layer, the input was just the original image. However, when we're talking about the 2nd conv layer, the input is the activation map(s) that result from the first layer. Hence, each layer of the input is basically describing the locations in the original image for where certain low level features appear. Now when we apply a set of filters on top of that (pass it through the 2nd conv layer), the output will be activations that represent higher level features. Types of these features could be semicircles (combination of a curve and straight edge) or squares (combination of several straight edges). As we go through the network and go through more conv layers, we get activation maps that represent more and more complex features. By the end of the network, you may have some filters that activate when there is handwriting in the image, filters that activate when they see pink objects, etc. Another interesting thing to note is that as we go deeper into the network, the filters begin to have a larger and larger receptive field, which means that they are able to consider information from a larger area of the original input volume.

4.2 Fully Connected Layer

This layer basically takes an input volume (whatever the output is of the conv or relu or pool layer preceding it) and outputs an N dimensional vector where N is the number of classes that the program has to choose from. For example, in a digit classification program, N would be 10 since there are 10 digits. Each number in this N dimensional vector represents the probability of a certain class. For example, if the resulting vector for a digit classification program is [0 .1 .1 .75 0 0 0 0 0 .05], then this represents a 10% probability that the image is a 1, a 10% probability that the image is a 2, a 75% probability that the image is a 3, and a 5% probability that the image is a 9. This is the soft-max approach. The way this fully connected layer works is that it looks at the output of the previous layer (which as we remember should represent the activation maps of high level features) and determines which features most correlate to a particular class. For example, if the program is predicting that some image is a dog, it will have high values in the activation maps that represent high level features like a paw or 4 legs, etc. Similarly, if the program is predicting that some image is a bird, it will have high values

in the activation maps that represent high level features like wings or a beak, etc. Basically, a FC layer looks at what high level features most strongly correlate to a particular class and has particular weights so that when you compute the products between the weights and the previous layer, you get the correct probabilities for the different classes.

4.3 Back Propagation

The network is trained using back-propagation. It is nothing but determining the weights that are required in the network. Backpropagation can be separated into 4 distinct sections, the forward pass, the loss function, the backward pass, and the weight update. During the **forward pass**, we take a training image which is a 32 x 32 x 3 array of numbers and pass it through the whole network. Initially, since all of the weights or filter values were randomly initialized, the output will probably be something like [.1 .1 .1 .1 .1 .1 .1 .1 .1 .1], basically an output that doesn't give preference to any number in particular. The network, with its current weights, isn't able to look for those low-level features or thus isn't able to make any reasonable conclusion about what the classification might be. This goes to the **loss function** part of backpropagation. Note that we are doing back propagation only using the training data. This data has both an image and a label. If we consider digit classification, and the first training image inputted was a 3. The label for the image would be [0 0 0 1 0 0 0 0 0 0]. A loss function can be defined in many different ways but a common one is MSE (mean squared error), which is ½ times (actual - predicted) squared.

$$E_{total} = \sum \tfrac{1}{2}(target - output)^2 \qquad (1)$$

Let's say the variable L is equal to that value. The loss will be extremely high for the first couple of training images. We want to get to a point where the predicted label (output of the ConvNet) is the same as the training label (This means that our network got its prediction right). In order to get there, we want to minimize the amount of loss we have. Visualizing this as just an optimization problem in calculus, we need to find out which inputs (weights in our case) most directly contributed to the loss (or error) of the network.

This is the mathematical equivalent of a **dL/dW** where W are the weights at a particular layer. Now, what we want to do is perform a **backward pass** through the network, which is determining which weights contributed most to the loss and finding ways to adjust them so that the loss decreases. Once we compute this derivative, we then go to the last step which is the **weight update**. This is where we take all the weights of the filters and update them so that they change in the direction of the gradient.

$$w = w_i - \eta \frac{dL}{dW} \qquad (2)$$

w = Weight
w_i = Initial Weight
η = Learning Rate

The **learning rate** is a parameter that is chosen by the programmer. A high learning rate means that bigger steps are taken in the weight updates and thus, it may take less time for the model to converge on an optimal set of weights. However, a learning rate that is too high could result in jumps that are too large and not precise enough to reach the optimal point.

The process of forward pass, loss function, backward pass, and parameter update is generally called one **epoch**. The program will repeat this process for a fixed number of epochs for each training image. Once we finish the parameter update on the last training example, hopefully the network should be trained well enough so that the weights of the layers are tuned correctly.

4.4 Stride

Stride controls how the filter convolves around the input volume. The amount by which the filter shifts is the **stride**. Stride is normally set in a way so that the output volume is an integer and not a fraction. Let's look at an example. Let's imagine a 7 x 7 input volume, a 3 x 3 filter (Disregard the 3rd dimension for simplicity), and a stride of 1. The following Fig. 5 shows the output when a stride of 1 is applied.

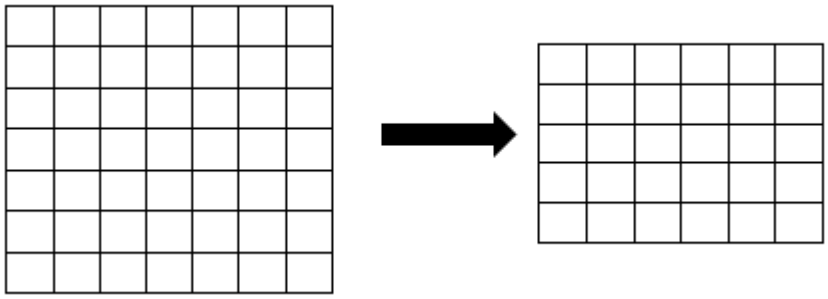

Fig. 5 Output of 5x5 volume is obtained when a stride of 1 is applied to a 7x7 input.

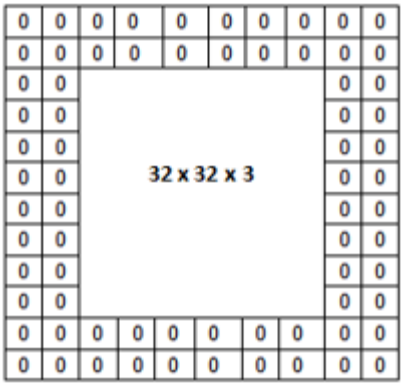

Fig. 6 With zero padding we are able to preserve the dimension.

Notice that if we tried to set our stride to 3, then we'd have issues with spacing and making sure the receptive fields fit on the input volume. Normally, programmers will increase the stride if they want receptive fields to overlap less and if they want smaller spatial dimensions.

4.5 Padding

When we apply three 5 x 5 x 3 filters to a 32 x 32 x 3 input volume, the output volume would be 28 x 28 x 3. Notice that the spatial dimensions decrease. As we keep applying conv layers, the size of the volume will decrease faster than we would like. In the early layers of our network, we want to preserve as much information about the original input volume so that we can extract those low level features. Let's say we want to apply the same conv layer but we want the output volume to remain 32 x 32 x 3. To do this, we can apply a zero padding of size 2 to that layer. Zero padding pads the input volume with zeros around the border. If we think about a zero padding of two, then this would result in a 36 x 36 x 3 input volume. Fig. 6 describes an overview on how padding is done.

4.6 ReLU (Rectified Linear Units) Layer

After each conv layer, it is convention to apply a nonlinear layer (or **activation layer**) immediately afterwards. The purpose of this layer is to introduce nonlinearity to a system that basically has just been computing linear operations during the conv layers (just element wise multiplications and summations). In the past, nonlinear functions like tanh and sigmoid were used, but researchers found out that **ReLU layers** work far better because the network is able to train a lot faster (because of the computational efficiency) without making a significant difference to the accuracy. The ReLU layer applies the function f(x) = max(0, x) to all of the values in the input volume. In basic terms, this layer just changes all the negative activations to 0.This layer increases the

nonlinear properties of the model and the overall network without affecting the receptive fields of the conv layer.

4.7 Pooling Layers

After some relu layers, programmers may choose to apply a **pooling layer**. It is also referred to as a down sampling layer. In this category, there are also several layer options, with max-pooling being the most popular. This basically takes a filter (normally of size 2x2) and a stride of the same length. It then applies it to the input volume and outputs the maximum number in every sub region that the filter convolves around.

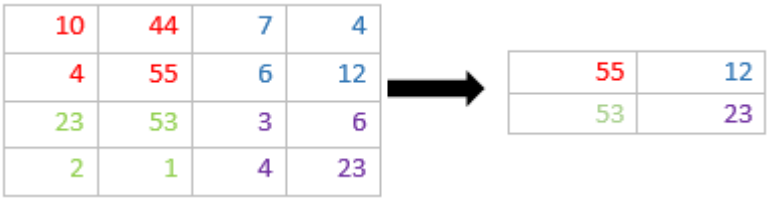

Fig. 7 Pooling with a filter of 2x2 of stride 2

Other options for pooling layers are average pooling and L2-norm pooling. The reasoning behind this layer is that once we know that a specific feature is in the original input volume (there will be a high activation value), its exact location is not as important as its relative location to the other features. This layer drastically reduces the spatial dimension (the length and the width change but not the depth) of the input volume. This serves two main purposes. The first is that amount of parameters or weights is reduced by 75%, thus lessening the computation cost. The second is that it will control **overfitting**. This term refers to when a model is so tuned to the training examples that it is not able to generalize well for the validation and test sets. A symptom of overfitting is having a model that gets 100% or 99% on the training set, but only 50% on the test data.
Fig. 7 describes about how pooling takes place.

Now, as you got introduced to few basic terminologies in this area, the next section is dedicated to explaining the pre-trained model which we employed in our work.

4.8 VGG-Face

The VGG-Face architecture consists of 16 weight layers out of which 13 are convolutional (conv) and remaining 3 are fully connected (fc) layers. All the convolution filters have a uniform receptive field of 3X3 pixels and each convolutional layer is followed by a rectified linear unit (relu) layer, which introduces non-linearity to the network by replacing all the negative values in the convolution output with zeros. The number of filters in the first convolutional layer is 64 and after each max-pooling layer (mpool), it is increased by a factor of 2. Therefore, the number of filters in the convolutional layer that follows the third mpool (pool3) is 512. There is no further increase in the number of the filters between pool4 and fc6. Spatial pooling is carried out by the five max-pooling layers over 2X2 pixels window, with a stride of 2. The spatial resolution of feature maps are reduced by half at each convolutional layer due to the spatial max-pooling. Therefore, the feature map resolution is reduced to 1/ 16 of the input image resolution at fifth convolution layer (Cholakkal et al., 2016). The three convolutional layers and relu layers preceding the fifth max-pooling (pool5) are numbered as conv5_1, conv5_2, conv5_3, relu5_1, relu5_2, and relu5_3, respectively. Similar naming convention is used for other layers.

CHAPTER 5

5. Features and Classifier

5.1 CNN based feature extraction

The VGG-Face has been trained on face images and hence, the fc layers encode the spatial information from the entire face image. Therefore, to extract features from images of the periocular region, I have excluded the fc layers and used only the convolution layers which are shared across spatial regions for feature extraction. For an input image of size 224X224 pixels, the relu5_2 feature map is of size 14X14X512. Here, 14X14 is the spatial resolution at relu5_2 and 512 is the number of channels in conv5_2. I have introduced a 14X14 spatial max-pooling at the relu5_2 (28^{th} layer) feature map to generate a CNN feature vector containing 512 elements. Each element in this feature vector represents the maximum response of the corresponding convolution channel.

Before feeding the input periocular image to the VGG-Face model for feature extraction, it is re-sized to 224X224 pixels.

The following sub sections briefly describes about the other features used in our work.

5.2 LBP

It is a texture descriptor initially used in different texture classification problems. From past few years it is also being used in face recognition and classification purpose. It is generally applied to gray-scale images. The binary code for a pixel is calculated based on the intensity values of the surrounding pixels. It generally encodes edges, corners etc. For a classic LBP, the feature size is generally 256. It is represented by symbol $LBP_{P,R}$

Every pixel will have its corresponding binary code. Suppose if we consider a 102x102 pixel image we will have 100x100 binary codes (Omitting the edges as we cannot calculate LBP code for them) these are converted to decimal numbers whose value ranges from 0-255. Now for an efficient representation and for reducing the dimension, we construct a histogram which counts the frequency of the decimal numbers and assigns them to 256 unique bins in the histogram. Hence, the 100x100 code of image is reduced to a histogram of 256 bins. The size of each bin represents the number of times the decimal number appeared in that code.

For eg: if number 0 is obtained 3000 times in 100x100 code. The length or size of the 1^{st} bin of the histogram will be 3000. For an easy explanation of the method I described it on a 3x3 square grid, but researchers usually work in a circular boundary where the neighborhood points lie on the boundary and the center pixel will be the center of the circle. In our work I chose 8 neighborhood points and radius of circle is 1. L2- normalization is applied for these feature-bins before feeding it to the SVM. The features obtained are invariant to uniform gray scale changes. Fig. 8 describes the working principle of LBP.

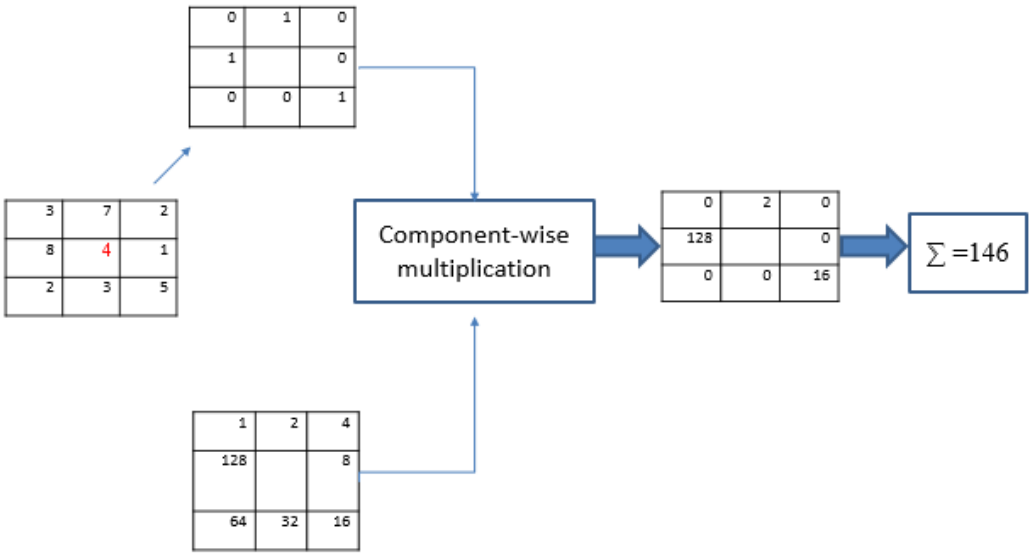

Fig. 8 Working principle of LBP

Uniform LBP is just a special case of LBP and it is usually represented as $LBP_{P,R}^{u2}$. It has 59 features instead of 256. This is bit powerful than the previous version because of its lower dimension. From the previously shown LBP codes (0-255), it only considers binary codes which contain "at-most 2" transitions ("0 to 1" or "1 to 0") taken in a circular fashion. For eg: 00000000 (0 transitions), 00000001 (2 transitions), 01100000 (2 transitions) etc will be considered and the remaining codes which have greater than 2 transitions when seen circularly are omitted. For eg: 00010100 (4 transitions), 01011011 (6 transitions). Out of 256 binary codes, 58 codes fall under this category. Therefore, this approach will have total bins of 59. 58 for the uniform codes and the rest will be assigned in a separate bin. In this work, I implemented uniform LBP in Matlab 2015b version using extractLBPFeatures function.

5.3 Histogram of Oriented Gradients (HOG)

HOG was introduced by Dalal and Triggs as a technique to represent the gradient orientations in a regular area of the image, called cell. The input image is divided into a rectangular grid of cells, representing each cell by a histogram, and the whole image by the concatenation of the respective cell histograms.

HOG reduces the illumination influence normalizing each cell histogram taking into account the cell neighborhood, that is known as the block.

In this work, we extracted HOG features using extractHOGFeatures() function in Matlab 2014. I made use of a cell size of 64X64 pixels. The block size is 2X2 cells and there is an overlapping of 50%. I considered the number of orientation bins as 12. When we consider an image of 224X224 pixels, we would be having 4 blocks in an image. Therefore, our feature vector will be of size 4(blocks) x 4(cells) x 12 bins = 192.

5.4 Dong's features

5.4.1 Global Shape Features

They extracted three global shape features from each eyebrow independent of scale and rotation. They are Rectangularity: It reflects how rectangular an eyebrow shape is. It is defined as the area of the eyebrow divided by the area of the minimum bounding rectangle.

Eccentricity: It specifies the eccentricity of the ellipse that has the same second-moments as the eyebrow region. It is defined as the ratio of the distance between the foci of the ellipse and its major axis length.

Isoperimetric quotient: It represents how a shape is similar to a circle. It is defined as the ratio of its area and that of the circle having the same perimeter.

Fig 9 describes the way in which they have extracted their features.

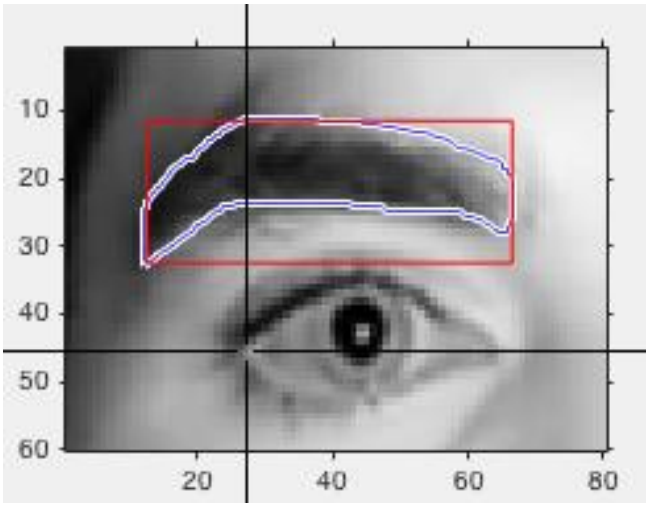

Fig. 9 Drawing the contour of the eyebrow shape and choosing various points

Fig. 10 Extracting local area features

5.4.2 Local Area Features

They extracted a total of eight local features by calculating the local area percentage of the eyebrow as illustrated in Figure. 10. First, they divided the minimum bounding rectangle around the eyebrow region into four sub-regions of the same width. Assuming the whole eyebrow area is A and the area of eyebrow in each sub-region is AI, A2, A3, and A4 respectively, they calculated the area percentage for each sub-region.

Later they divided the minimum bounding rectangle around the eyebrow region into four sub-regions of the same height and got another four area percentage features. Hence, we have total of 8 features.

5.4.3 Critical Point Features

There are several critical points on the eyebrow like top-most point, left most point, right most point and centroid of the eyebrow shape. The locations or distances of these points are calculated after shifting the origin to endocathion (inner corner of the eye) or exocathion (outer corner of the eye). These critical features are 8 in number.

Combining all the features, we have a total of 19 feature vector. I implemented their work by following all the guidelines mentioned in their work.

5.5 CLASSIFIER

In our approach, the 512-dimensional feature vector generated using the deep CNN is fed to a SVM for classification. Specifically, I have used linear soft-margin formulation of SVM, which finds an optimal linear separating hyper-plane while performing a trade-off between maximization of the margin and minimization of the training error (Vapnik, 1998). The training data consists of a set of feature vectors

$x_i \in R^d$ and the associated class information $y_i \in \{1,-1\}$.

A hyperplane separating the two classes is given by:

$$w^T x + b = 0 \qquad (3)$$

where the parameters w and b are determined using the training data. The SVM finds an optimal hyperplane by solving the following optimization problem (Vapnik, 1998):

$$\begin{aligned} \underset{\mathbf{w},b,\zeta}{\text{minimize}} \quad & \mathbf{w}^T\mathbf{w} + C\sum_i \zeta_i \\ \text{subject to} \quad & y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \zeta_i, \\ & \zeta_i \geq 0. \end{aligned} \qquad (4)$$

Where $delta_i$ are the slack variables, and C is the regularization parameter, which controls overfitting. Based on the trained weight vector, the SVM determines the class of the test feature vector v_i using the following equation:

$$class(v_i) = sgn(w^T v_i + b) \qquad (5)$$

class (v_i) has two possible values, which indicate the class of the test periocular image.

In our experiments, SVM implementation has been done using the LIBSVM package (Chang and Lin, 2011). The parameters C and gamma are chosen from the grid search. I varied C from [0.5, 1, 2, 4, 8, 16, 32] and gamma from [0.0625, 0.125, 0.25, 0.5, 1, 2]. The best C and gamma are chosen from these values based on dago-protocol.

CHAPTER 6

Experimental setup

In this work, performance evaluations have been carried out using a standard five-fold cross validation procedure proposed in (Dago-Casas et al., 2011). Previous work (Castrill´on-Santana et al., 2015; Dong and Woodard, 2011; Lyle et al., 2012) for gender classification from periocular images has also reported five-fold cross validation results. The protocol presented in (Dago-Casas et al., 2011) for gender classification specifies how the dataset should be divided into 5 folds, maintaining a balanced distribution of images from male and female classes in each fold.

Datasets used for performance evaluation have been partitioned into 5 folds in such a way that all images of a subject are contained in one fold, which ensures that images of a subject do not appear in both training and test sets. Since there is no overlap of a subject's images between two folds, I refer to it as non-overlapping case in this work. As performance measures, I reported accuracy (ACC), sensitivity (SEN) and specificity (SPE) in this work. For calculating SEN and SPE, I have considered images belonging to male subjects as positive samples and those belonging to female subjects as negative samples. Therefore, SEN and SPE provide measures of the system's ability to correctly classify images from male and female subjects, respectively. These performance measures are computed as follows:

$$ACC = (TP + TN)/ (TP + TN + FP + FN) \quad (6)$$

$$SEN = (TP)/ (TP + FN) \quad (7)$$

$$SPE = (TN)/ (TN + FP) \quad (8)$$

where TP, TN, FP and FN denote number of true positive, true negative, false positive and false negative samples, respectively.

To obtain better estimates of performance, I have repeated the experiments 10 times. Therefore, I report the average and standard deviation of the performance measures obtained in each of the evaluations.

6.1 Why do we require sensitivity and specificity?

Suppose, if we consider 1000 persons (900 male and 100 female) for gender classification task. If our model is predicting all the persons in the test as shown in Table 1

Table 1. An experimental table to demonstrate the use of SEN and SPF apart from ACC

	Predicted Female	Predicted Male
Ground truth Female	0(True Negative)	100(False Positives)
Ground truth Male	0(False Negatives)	900(True Positives)

From the above table

ACC = 90%

SEN = 1

SPF = 0.

Even though we are able to obtain an accuracy of 90% the model is biased and it's not performing well because specificity is 0 for this model which implies that it is unable to predict a woman properly. A good model will have specificity and sensitivity close to 1.

Therefore, for better estimates of performance, we should also include SEN and SPF along with ACC

6.2 Database

I have used three databases for performance evaluation, details of which are presented in the following subsections.

6.2.1 IITI database

This database, collected in-house, consists of head and shoulder NIR images of 96 subjects, out of which 49 are male and 47 are female subjects. Images were captured using a Hikvision (model number: DS-2CD2420F-I(W)) surveillance camera (see Fig. 11) in an indoor office environment under no-light conditions. I captured images of the subjects at two standoff distances by instructing the subjects to keep their head close to the camera (about 2.5 feet) and far from the camera (about 5 feet). In this manner, a total of six images were captured from each subject, with three images at the shorter and three images at the longer standoff distance, as shown in Fig. 11.

Fig. 11. (a) Hikvision surveillance camera used in IITI database collection. (b) Images of a subject in the database, with top row and bottom row showing images captured at closer standoff distance and longer standoff distance, respectively

Images thus acquired have been processed to generate a set of left and right periocular images.
The detailed composition of the IITI periocular image dataset is shown in Table 2. As can be seen, this dataset has nearly balanced classes. The size of the acquired head and shoulder images is 1920 X1080 pixels, while the extracted periocular images vary in size from person to person and also because of the scale changes. The average size of a periocular image in our dataset is 127.6±26.58 165.1±34.3 pixels.

Table. 2. Composition of IITI NIR periocular image dataset

	Left Periocular			Right Periocular		
	Male	Female	Total	Male	Female	Total
Subjects	49	47	96	49	47	96
Images	294	282	576	294	282	576

6.2.2 MBGC portal database:

MBGC portal database (Phillips et al., 2009) consists of 149 facial video recordings of 114 subjects walking through a portal. The size of frames in the video is 2048X2048 pixels. This database has been used to evaluate the performance of our approach in non-ideal conditions like motion blur and varying illumination. Fig. 12 shows few images of this challenging database. In the literature, Dong and Woodard (2011) have evaluated their eyebrow shape based gender classification approach on a set of periocular images (which included a large number of non-ideal images) extracted from videos in MBGC portal database. Therefore, I have followed the procedure described in (Dong and Woodard, 2011) to extract images from the video

Fig. 12. Few images from the challenging non-ideal images of MBGC

recordings. Specifically, I have extracted frames containing motion-blur, variations in illumination, pose and expressions. Also, I have selected only the subjects whose eyebrows are fully-visible and excluded frames containing partial or no-eyebrows. While Dong and Woodard (2011) have reported a total of 922 such images, I have obtained a set of 909 images, which includes 497 left and 412 right periocular images. In this work, we will refer to this set of images as MBGC periocular image dataset. Table 3 shows the composition of this dataset. As is the case with IITI dataset, these periocular images have been extracted automatically. Fig. 13 shows sample images from this dataset.

Table 3. Composition of the MBGC periocular image dataset

	Left Periocular			Right Periocular		
	Male	Female	Total	Male	Female	Total
Subjects	52	37	89	46	31	77
Images	294	203	497	216	196	412

Fig. 13. Sample NIR images of the left and right periocular regions, which are automatically extracted from the MBGC database. The first and second rows show images of male and female subjets, respectively.

6.2.3 IIITD Multi-spectral periocular database

IIITD Multi-spectral periocular (IMP) database (Sharma et al., 2014) contains periocular images captured in VIS, NIR and night vision spectra. In this work, I have used only the subset (of the database) containing NIR periocular images. This subset contains a total of 620 images of 62 subjects. Since the ground-truth gender information has not been\ provided as part of the dataset, I have made use of visible periocular images in the database to manually assign gender labels to all the subjects.

The ground-truth gender information used in our experiments has been generated by correlating the gender assignments (independently) made by five members of our research group. Out of 62 subjects, our ground-truth information indicates that 35 are male subjects and 27 are female subjects. Sample images are shown in Fig. 14 and table 4 presents details of the IMP database.

Fig. 14 Sample NIR periocular images from the IMP database

Table. 4 Composition of NIR subset of the IMP database

	Left Periocular			Right Periocular		
	Male	Female	Total	Male	Female	Total
Subjects	35	27	62	35	27	62
Images	175	135	310	175	135	310

CHAPTER 7

RESULTS AND DISCUSSION

7.1 Experimental Results

In this chapter, I present results from performance evaluations carried out in our study. For a comparative evaluation, I have implemented three baseline algorithms based on state-of the-art texture and shape descriptors, which have been widely used for gender classification in face and periocular images.

Specifically, Baseline 1 and Baseline 2 are based on global and concatenated local histograms of uniform LBP (Ojala et al., 2002), respectively. To generate local histogram features, I divided the input periocular image into 2X2 regions. The third baseline approach (Baseline 3) employs HOG (Dalal and Triggs, 2005) features for gender classification. The parameters involved in computation of the HOG descriptor are determined empirically. Specifically, I have set the cell size and block size to 64X64 pixels and 2X2 cells, respectively. In addition to the aforementioned baseline algorithms, I have compared our approach with the gender classification approach proposed in (Dong and Woodard, 2011), since it has also been evaluated on periocular images extracted from the MBGC portal database.

For this purpose, I have re-implemented their approach. Parameters of the non-linear SVM in their approach, specifically, the regularization parameter C and gamma of the radial basis function

(RBF) kernel are set based on the protocol in (Dago-Casas et al., 2011) as described before in Chapter 5.

7.2 Dago Protocol

In a five-fold cross validation we use 4 folds for training and 1 for testing. In the training stage, in order to select the optimum values of the parameter C of the SVM (when using SVMs classifier) they used the following approach. One of the training folds was selected as validation fold and, using the three remaining folds, they trained a classifier for each value (or pair of values in the SVM case) of the parameters we wanted to test and tested them in the validation fold. This was repeated using the four training folds as validation fold and results were averaged. Then, those values of the parameters that obtained the best average classification accuracy were selected.

After validation, the optimum parameters were used to train a classifier using the whole training set (4 folds), and its performance was tested in the fifth fold. This whole procedure was repeated for the 5 possible combinations of training and test folds, and the results were averaged.

7.3 VGG-Face vs VGG-VD

As described in Chapter 5, the proposed approach employs a deep CNN namely, VGG-Face for feature extraction. The architecture of VGG-Face is similar to that of VGG-VD (Simonyan and Zisserman, 2014). Therefore, in the first set of experiments, I have compared the performances of these two pre-trained CNN models for gender classification. Specifically, I have used VGG-VD with 16 weight layers. In contrast to VGG-Face, this is a generic model as it has been trained on diverse ImageNet database (Russakovsky et al., 2015) for object localization and classification tasks. Table 5 presents 5-fold classification accuracies obtained using features extracted from different intermediate layers of the two CNNs on left periocular images of the MBGC dataset. As can be observed, features extracted from VGG-Face as well as VGG-VD are effective for gender classification. Overall, the domain specific VGG-Face model provides better performance. However, it can be observed that the drop of performance in the case of VGG-VD is not too significant. This may be due to the generic nature of the image representation obtained from the pre-trained CNN models. Generally, these intermediate layers respond to low-level features such as edges and corners in the image, while the top layers capture features, which are specific to the task it is trained for (Zheng et al., 2016). Since the VGG-Face yields marginally higher classification accuracy for most of the cases, I extracted features from the output of relu5_2 layer of this

model as the feature vector for further analysis.

Table 5. Performance comparison of the VGG-Face and VGG-VD on left periocular images of the MBGC dataset

Layer	VGG-face (%)	VGG-VD
25(conv5_1)	86.39	86.63
26(relu5_1)	86.37	87.03
27(conv5_2)	88.97	84.91
28(relu5_2)	89.87	85.42
29(conv5_3)	86.76	85.97
30(relu5_3)	86.37	86.01
31(pool5)	86.54	85.52

7.4 Non-overlapping case

As previously mentioned, 5-fold cross validations have been carried out in a non-overlapping scenario, in which I have made sure that all the images of a subject appear in only one fold. Table 6 summarizes results from this set of experiments.

Table 6 Performance on three different database for non-overlapping case

METHOD	LEFT PERIOCULAR			RIGHT PERIOCULAR		
	MBGC Dataset					
	ACC	SEN	SPE	ACC	SEN	SPF
Proposed method	**89.87±0.96**	0.942±0.013	0.833±0.01	**84.89±0.74**	0.894±0.01	0.803±0.025
(Dong and Woodard, 2011)	86.5±1.45	0.87±0.013	0.85±0.02	82.95±2.19	0.87±0.02	0.78±0.029
Baseline 1	75.8±0.95	0.874±0.007	0.594±0.02	65.3±1.65	0.81±0.01	0.481±0.025
Baseline 2	85.73±0.98	0.92±0.01	0.768±0.01	75.18±1.39	0.78±0.01	0.71±0.03
Baseline 3	76.4±1.79	0.81±0.015	0.687±0.034	63.7±1.01	0.722±0.019	0.54±0.022
	IITI Dataset					
	ACC	SEN	SPE	ACC	SEN	SPF
Proposed method	**95.4±0.84**	0.95±0.005	0.952±0.009	94.2±0.62	0.946±0.01	0.93±0.006
(Dong and Woodard, 2011)	85.8±1.02	0.87±0.009	0.838±0.017	83.31±1.15	0.817±0.02	0.849±0.008
Baseline 1	81.7±0.98	0.77±0.01	0.857±0.01	88.4±0.867	0.845±0.015	0.92±0.008
Baseline 2	92.7±0.64	0.91±0.012	0.93±0.01	94.2±0.68	0.93±0.01	0.94±0.009
Baseline 3	91.2±1.33	0.91±0.01	0.90±0.01	**95.8±0.98**	0.952±0.009	0.96±0.011
	MBGC Dataset					
	ACC	SEN	SPE	ACC	SEN	SPF
Proposed method	**86.11±.5**	0.96±0.013	0.72±0.028	84.77±1.42	0.905±0.017	0.771±0.033
(Dong and Woodard, 2011)	N.A	N.A	N.A	N.A	N.A	N.A
Baseline 1	74.7±3.47	0.81±0.027	0.65±0.06	73.5±2.57	0.74±0.02	0.71±0.055
Baseline 2	84.6±1.3	0.88±0.013	0.79±0.03	90.4±1.32	0.93±0.02	0.86±0.023
Baseline 3	82.6±2.31	0.89±0.012	0.74±0.05	80.8±2.31	0.83±0.02	0.76±0.024

As can be observed, the proposed approach achieves the highest classification accuracy on left periocular images of the three datasets. On the other hand, there is no clear winner in the case of right periocular images.

On the MBGC dataset, the proposed approach provides the highest accuracy for gender classification in both the left and the right periocular images. Among the methods selected for comparison, the method of Dong and Woodard (2011) is clearly better than the baseline methods. As previously mentioned, the MBGC dataset contains a large number of non-ideal images and descriptors such as LBP are known to be sensitive to image blur and local variations in illumination. This explains why baseline algorithms perform relatively poorly on the MBGC dataset.
On the other hand, the shape features employed in (Dong and Woodard, 2011) are less likely to be affected by the quality of periocular images.

On the IITI dataset, it is noteworthy that local descriptor based baseline algorithms (Baseline 2 and Baseline 3) outperform eyebrow feature based method. Periocular images in IITI dataset do not contain blurred regions or significant variations in illumination. Therefore, local descriptors in baseline algorithms are effective in capturing the discriminatory information, leading to improved performance for gender classification.
On the IMP dataset, the performance of the proposed approach is quite comparable with that of Baseline 2. Note that I have not evaluated the performance of the eyebrow shape based method (Dong and Woodard, 2011) on the IMP database. This is because, several of the images in the IMP database contain only partial eyebrows.

Overall, our experimental results suggest that eyebrow shape features are best suited for gender classification in non-ideal images, while local texture and shape features are likely to yield better performance in systems that use constrained imaging set up. More importantly, the performance of the proposed approach is more consistent across the databases considered in this study.

7.5 Overlapping case

Previous works (on gender classification from NIR periocular images), including the one (Dong and Woodard, 2011) that used MBGC dataset for performance evaluation have not indicated whether the 5-fold cross validations have been performed for the non-overlapping case. Therefore, I have also performed the standard 5-fold cross validation on all the images of the MBGC dataset. In this case, there is a high possibility that different images of a subject may appear in the training as well as the testing folds. Therefore, the classification performance is expected to be higher as compared with the non-overlapping case.
Table 7 presents the performance measures for this case. As can be observed, the proposed approach achieves the highest classification performance for this case as well. At this point, it must be noted that our implementation of the method in (Dong and Woodard, 2011) has not achieved the accuracy (96% for left as well as right periocular images) reported in their paper. This may be due to the following reasons. As mentioned in Chapter 5, although I have followed their procedure for extraction of periocular images from the MBGC database, our dataset falls short by 13 images. Therefore, our dataset is not exactly the same as the one used in (Dong and Woodard, 2011) for performance evaluation. In addition, since the eyebrow region is segmented manually, the variability (in the eyebrow contour) arising from this manual process might affect the performance of the approach.

Table 7. Performance on MBGC dataset for the overlapping case

METHOD	LEFT PERIOCULAR			RIGHT PERIOCULAR		
	ACC	SEN	SPE	ACC	SEN	SPF
Proposed method	**97.3±0.35**	0.98±0.005	0.95±0.006	**96.4±0.89**	0.97±0.009	0.95±0.01
(Dong and Woodard, 2011)	91.2±0.96	0.91±0.013	0.91±0.008	92.6±0.66	0.94±0.007	0.90±0.015
Baseline 1	78.6±0.67	0.89±0.008	0.62±0.01	72.4±0.764	0.84±0.008	0.59±0.018
Baseline 2	90.5±0.36	0.95±0.004	0.83±0.009	85.7±1.2	0.89±0.014	0.81±0.013
Baseline 3	84.8±0.76	0.87±0.007	0.8±0.012	78.4±1.13	0.83±0.013	0.72±0.013

Table 8. Performance on IITI dataset without eyebrow information

METHOD	LEFT PERIOCULAR			RIGHT PERIOCULAR		
	ACC	SEN	SPE	ACC	SEN	SPF
Proposed method	78.83±1.45	0.79±0.01	0.77±0.015	**78.3±0.84**	0.79±0.013	0.77±0.0161
Baseline 1	76.9±1.5	0.74±0.019	0.79±0.016	75.37±1.4	0.731±0.015	0.776±0.016
Baseline 2	80.5±1.49	0.8±0.025	0.8±0.013	75.4±1.69	0.76±0.017	0.74±0.028
Baseline 3	75.06±2.15	0.72±0.02	0.77±0.022	75.33±0.95	0.76±0.019	0.73±0.016

7.6 Effect of absence of eyebrows on the performance

The goal of this set of experiments is to obtain performance estimates of the proposed approach in scenarios where eyebrow information may not be available for gender recognition. For this purpose, I have manually cropped the periocular region from images in the IITI database, while making sure the eyebrow region is excluded. Specifically, I have marked the outer corners of the eyes manually for each image. With the help of these points, images have been rotated such that the line connecting the two corner points becomes horizontal. This is followed by manual extraction of a rectangular periocular region without the eyebrow. Fig. 15 shows a few sample images generated for our experiments. As can be seen, gender recognition from such images can be quite challenging.

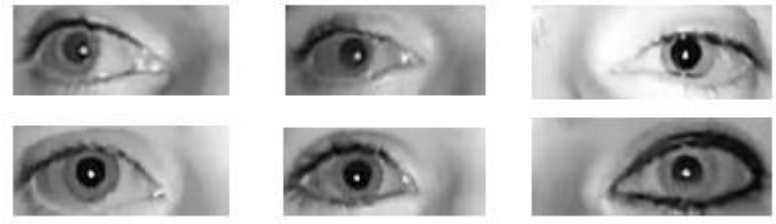

Fig. 15 Periocular images without eyebrow information. The first and second rows show images of male and female subjects in the IITI database.

The results from this set of experiments are presented in Table 8. In the absence of eyebrow in the periocular images, our approach yields classification accuracy of 78.83% and 78.38% on the left and the right periocular images, respectively. These results indicate that, in the proposed approach, features from eyebrow region contribute considerably to gender classification.

This observation is consistent with the previous work (Dong and Woodard, 2011) that has established that eyebrows play a key role in recognizing the gender.

CHAPTER 8

Conclusion and future work

In this work, I have investigated the effectiveness of features extracted from a deep CNN for NIR periocular image based gender classification. The proposed approach employs a pre-trained CNN namely, VGG-Face and extracts features for gender classification from a convolutional layer in this net.

The 512- dimensional descriptor thus extracted is fed to a linear SVM for classification. Three databases of NIR periocular images have been used to ascertain the performance of the proposed approach. Our 5-fold cross validation results indicate that the generic descriptor extracted from the VGG-Face model carries significant discriminatory information for gender classification. Overall, the performance of the proposed approach is quite comparable with the state-of-the-art hand-crafted descriptors.

More importantly, it achieves the state-of-the-art performance on a set of periocular images extracted from the MBGC database, which contains several non-ideal frames.

The results from additional experiments carried out to study the effect of absence of eyebrow in periocular images on the performance of the proposed approach suggest that eyebrow features contribute considerably to gender classification. With the drop in performance due to the lack of eyebrow information, the proposed approach achieves nearly 79% classification accuracy on our dataset.

As part of our future work, I would like to expand our dataset by including more challenging images of the periocular region. Specifically, I plan to augment our dataset with images collected using a set up that emulates surveillance scenarios. When such a large dataset of NIR periocular images are available, I plan to explore whether fine-tuning of the CNN leads to further improvement in gender classification performance. I also plan to develop an effective technique for fusion of information in left and right periocular images. Such a fusion scheme may be employed to enhance the performance when both periocular images of an individual are available for gender classification.

REFERENCES:

1. Alonso-Fernandez, F., Bigun, J., 2015. A survey on periocular biometrics research. PatternRecognition Letters .

2. Amayeh, G., Bebis, G., Nicolescu, M., 2008. Gender classification from hand shape, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, pp. 1–7. doi:10.1109/CVPRW.2008. 4563122.

3. Castrill´on-Santana, M., Lorenzo-Navarro, J., Ram´on-Balmaseda, E., 2015. On using periocular biometric for gender classification in the wild. Pattern Recognition Letters .

4. Chang, C.C., Lin, C.J., 2011. Libsvm: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology (TIST) 2, 27.

5. Chen, C., Ross, A., 2011. Evaluation of gender classification methods on thermal and near-infrared face images, in: International Joint Conference on Biometrics (IJCB), pp. 1–8. doi:10.1109/IJCB.2011.6117544.

6. Cholakkal, H., Johnson, J., Rajan, D., 2016. Weakly supervised top-down salient object detection. arXiv preprint arXiv:1611.05345.

7. Dago-Casas, P., Gonz´alez-Jim´enez, D., Yu, L.L., Alba-Castro, J.L., 2011. Single-and cross-database benchmarks for gender classification under unconstrained settings, in: Computer Vision workshops (ICCV Workshops), 2011 IEEE international conference on, IEEE. pp. 2152–2159.

8. Dalal, N., Triggs, B., 2005. Histograms of oriented gradients for human detection, in: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), IEEE. pp. 886–893.

9. Dantcheva, A., Elia, P., Ross, A., 2015. What else does your biometric data reveal? a survey on soft biometrics. IEEE Transactions on Information Forensics and Security, 1–26.

10. Dong, Y., Woodard, D.L., 2011. Eyebrow shape-based features for biometric recognition and gender classification: A feasibility study, in: Biometrics (IJCB), 2011 International Joint Conference on, IEEE. pp. 1–8.

11. Jain, A.K., Dass, S.C., Nandakumar, K., 2004. Can soft biometric traits assist user recognition?, in: Defense and Security, International Society for Optics and Photonics. pp. 561–572.

12. Lee, L., Grimson, W.E.L., 2002. Gait analysis for recognition and classification, in: Automatic Face and Gesture Recognition, 2002. Proceedings. Fifth IEEE International Conference on, IEEE. pp. 148–155.

13. Lyle, J.R., Miller, P.E., Pundlik, S.J.,Woodard, D.L., 2012. Soft biometric classification using local appearance periocular region features. Pattern Recognition 45, 3877–3885.

14. Merkow, J., Jou, B., Savvides, M., 2010. An exploration of gender identification using only the periocular region, in: Biometrics: Theory Applications and Systems (BTAS), 2010 Fourth IEEE International Conference on, IEEE. pp. 1–5.

15. Ojala, T., Pietikainen, M., Maenpaa, T., 2002. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Transactions on pattern analysis and machine intelligence 24, 971–987.

16. Ozbulak, G., Aytar, Y., Ekenel, H.K., 2016. How transferable are cnn-based features for age and gender classification?, in: Biometrics Special Interest Group (BIOSIG), 2016 International Conference of the, IEEE. pp. 1–6.

17. Park, U., Ross, A., Jain, A.K., 2009. Periocular biometrics in the visible spectrum: A feasibility study, in: Biometrics: Theory, Applications, and Systems, 2009. BTAS'09. IEEE 3rd International Conference on, IEEE. pp. 1–6.

18. Parkhi, O.M., Vedaldi, A., Zisserman, A., 2015. Deep face recognition, in: British Machine Vision Conference, p. 6.

19. Phillips, P.J., Flynn, P.J., Beveridge, J.R., Scruggs, W.T., Otoole, A.J., Bolme, D., Bowyer, K.W., Draper, B.A., Givens, G.H., Lui, Y.M., et al., 2009. Overview of the multiple biometrics grand challenge, in: International Conference on Biometrics, Springer. pp. 705–714.

20. Phillips, P.J., Flynn, P.J., Scruggs, T., Bowyer, K.W., Chang, J., Homan, K.,Marques, J., Min, J., Worek, W., 2005. Overview of the face recognition grand challenge, in: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), IEEE. pp. 947–954.

21. Rattani, A., Chen, C., Ross, A., 2014. Evaluation of texture descriptors for automated gender estimation from fingerprints, in: European Conference on Computer Vision, Springer. pp. 764–777.

22. Ross, A., Chen, C., 2011. Can gender be predicted from near-infrared face images?, in: International Conference Image Analysis and Recognition, Springer. pp. 120–129.

23. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z.,Karpathy, A., Khosla, A., Bernstein, M., et al., 2015. Imagenet large scale visual recognition challenge. International Journal of Computer Vision 115, 211–252.

24. Shafey, L.E., Khoury, E., Marcel, S., 2014. Audio-visual gender recognition in uncontrolled environment using variability modeling techniques, in: IEEE International Joint Conference on Biometrics, pp. 1–8. doi:10.1109/BTAS. 2014.6996271.

25. Sharif Razavian, A., Azizpour, H., Sullivan, J., Carlsson, S., 2014. Cnn features off-the-shelf: an astounding baseline for recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 806–813.

26. Sharma, A., Verma, S., Vatsa, M., Singh, R., 2014. On cross spectral periocular recognition, in: 2014 IEEE International Conference on Image Processing (ICIP), IEEE. pp. 5007–5011.

27. Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 .

28. Thomas, V., Chawla, N.V., Bowyer, K.W., Flynn, P.J., 2007. Learning to predict gender from iris images, in: First IEEE International Conference on Biometrics: Theory, Applications, and Systems, pp. 1–5. doi:10.1109/ BTAS.2007.4401911.

29. Vapnik, V., 1998. Statistical learning theory. 1998.

30. Viola, P., Jones, M.J., 2004. Robust real-time face detection. International Journal of Computer Vision 57, 137–154.

31. Wang, J.G., Li, J., Yau, W.Y., Sung, E., 2010. Boosting dense sift descriptors and shape contexts of face images for gender recognition, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition -Workshops, pp. 96–102. doi:10.1109/CVPRW.2010.5543238.

32. Zheng, L., Zhao, Y.,Wang, S.,Wang, J., Tian, Q., 2016. Good practice in CNN feature transfer. CoRR abs/1604.00133. URL: http://arxiv.org/abs/ 1604.00133.

33. https://adeshpande3.github.io/adeshpande3.github.io/The-9-Deep-Learning-Papers-You-Need-To-Know-About.html

34. http://www.scholarpedia.org/article/Local_Binary_Patterns