university of groningen

faculty of science and engineering

MASTER'S THESIS

Fusion of trainable features for gender recognition from face images

FRANS JUANDA SIMANJUNTAK

Supervisor: Dr. George Azzopardi
Co Supervisor: prof.Dr. Dimka Karastoyanova

Department of Computing Science
Faculty of Science and Engineering
University of Groningen
June 2018

Frans Juanda Simanjuntak: *Fusion of trainable features for gender recognition from face images,,* Master's in Computing Science, © June 2018

SUPERVISORS:
Dr. George Azzopardi
prof.Dr. Dimka Karastoyanova

LOCATION:
Groningen

ABSTRACT

Gender recognition from face images has been a well-studied topic in computer vision for years and its popularity arises due to the fact that nowadays it has been applied in many different fields such as surveillance and security systems, demographic information collection, marketing research, real time electronic marketing, criminology, and augmented reality. One of the most common problems with gender recognition is the pose variations of face images, and in addition, the partial occlusion of the face, age variation, different races and expression are also several challenges that we might encounter. In order to address the challenges, many attempts have been made and several approaches have been proposed, however a fixed solution has not been found yet.

Recently, VGGFace has become the golden standard for face recognition and turns out its performance is quite impressive by achieving 97% classification rate in recognizing gender on the GENDER-FERET (GF) data set. Another approach is by using trainable filters called Combination of Shifted Filter Responses (COSFIRE). The COSFIRE-based approach has already been evaluated for the detection of gender from face images, and it achieved an accuracy of 93.85 % on the GF and 99.19% on the Labeled Faces in the Wild (LFW) data set.

In this thesis, we propose fuse the trainable features from VGGNet and COSFIRE. In practice, we propose two ways of fusion; concatenating the VGGFace- and COSFIRE-based features and learning a stacked classifier for both feature sets. The results of the study show that both proposed techniques are demonstrated to be effective in classifying gender by achieving their best performance with an accuracy of 98.94% and 99.38% validated on the constrained data set GF and unconstrained one LFW using SVM classifier, respectively.

ACKNOWLEDGMENTS

First and foremost, I would like to thank God for giving me the strength, knowledge, and ability to undertake and complete this research study as my final journey of being a master student at the University of Groningen.

I would like to give my gratitude to both of my supervisors, Dr. George Azzopardi and prof.Dr. Dimka Karastoyanova for investing the time and also for sharing the knowledge and research skills for my thesis. I value all the hard work and patience you gave to me during the supervision which surely will be beneficial for my future.

Also, many thanks to my dad and my family in Indonesia who always support and gave me countless love and prayers. My beloved friends and my class mates, thank you for our love hate relationship, support, laugh, tears, and experience we shared for the last two years. My life would not be that meaningful and colorful in Groningen without all of you by my side.

Lastly and the second foremost, thanks to LPDP (Lembaga Pengelola Dana Pendidikan) who gave me opportunity to pursue my dream to study abroad in the Netherlands by awarding me a scholarship and lighting my financial burden which allows me to focus more on the most important aspect of school, learning. Thank you so much Indonesia.

CONTENTS

I MASTER THESIS

1 INTRODUCTION 3
- 1.1 Research Questions . . . 6
- 1.2 Methodology . . . 6
- 1.3 Thesis Organization . . . 7

2 RELATED WORK 9
- 2.1 Geometric-based . . . 9
- 2.2 Appearance-based . . . 10
- 2.3 Geometric and Appearance-based . . . 12
- 2.4 State-of-the-art Summary . . . 13

3 METHODOLOGY 15
- 3.1 Face Detection and Alignment . . . 15
- 3.2 Neural Networks-based Classifier . . . 17
  - 3.2.1 Convolutional Neural Networks (CNN) . . . 18
  - 3.2.2 VGGFace . . . 20
  - 3.2.3 VGGFace Classification Model . . . 22
- 3.3 COSFIRE-based Classifier . . . 22
  - 3.3.1 COSFIRE Method and 2D Gabor filters . . . 23
  - 3.3.2 COSFIRE Filter Configuration . . . 24
  - 3.3.3 COSFIRE Filter Response . . . 24
  - 3.3.4 Face Descriptor . . . 25
  - 3.3.5 COSFIRE Classification Model . . . 26
- 3.4 The Proposed Methods . . . 27
  - 3.4.1 Fusion of CNN and COSFIRE features by concatenation approach . . . 27
  - 3.4.2 Fusion of CNN and COSFIRE features by a stacking approach . . . 29

4 EXPERIMENTAL RESULTS 31
- 4.1 Datasets . . . 31
- 4.2 Pre-Processing . . . 34
- 4.3 Experiments . . . 35
  - 4.3.1 Result with COSFIRE-based Method . . . 35
  - 4.3.2 Result with VGGFace CNNs-based Method . . . 35
  - 4.3.3 Result with the fusion of COSFIRE and VGGFace-based methods . . . 36
  - 4.3.4 Comparison with other methods . . . 37

5 DISCUSSION 39
- 5.1 The effectiveness of the proposed methods . . . 39
- 5.2 The performances of the proposed methods on the constrained and unconstrained data sets . . . 40
- 5.3 The effectiveness of SVM and XGBoost classifiers . . . 40

6 CONCLUSION AND FUTURE WORK 43
6.1 Future Work 43

II APPENDIX

A DATA SETS 47
A.1 Gender Feret 47
A.2 Labeled Faces in The Wild 47
B RESULTS 49
B.1 SVM 49
B.2 XGBoost 50

BIBLIOGRAPHY 53

LIST OF FIGURES

Figure 1.1 Typical problems of gender recognition 4
Figure 1.2 The average face images of male and female . 5
Figure 3.1 A high level diagram of the proposed methods. 15
Figure 3.2 Representation of the face alignment algorithm. 16
Figure 3.3 Pre-processing using Viola-Jones face detector on LFW data set. 17
Figure 3.4 The process of convolution on images. 19
Figure 3.5 An example of max pooling. 20
Figure 3.6 An example of fully connected layer. 20
Figure 3.7 The architecture of CNNs 20
Figure 3.8 The architecture of VGGFace 21
Figure 3.9 An overview of VGGFace as feature extractor in gender classification. 22
Figure 3.10 Configuration of COSFIRE filters using a training male face GF. 25
Figure 3.11 Application of the COSFIRE filters on a face image 26
Figure 3.12 Fusion of CNN and COSFIRE features by concatenation approach. 28
Figure 3.13 Fusion of CNN and COSFIRE features by a stacking approach. 30
Figure 4.1 Example male and female images from the GF data set. 32
Figure 4.2 Example male faces and female faces from the LFW data set. 33
Figure 4.3 Example images which were discarded from the LFW data set. 34
Figure B.1 The performance of concatenation and stacking approach on the GF data set using SVM. . 49
Figure B.2 The performance of concatenation and stacking approach on the LFW data set using SVM. 49
Figure B.3 The performance of COSFIRE on the GF data set using XGBoost. 50
Figure B.4 The performance of VGGFace on the GF data set using XGBoost. 50
Figure B.5 The performance of concatenation approach on the GF data set using XGBoost. 51
Figure B.6 The performance of COSFIRE on the LFW data set using XGBoost. 51
Figure B.7 The performance of VGGFace on the LFW data set using XGBoost. 52

Figure B.8 The performance of concatenation approach on the LFW data set using XGBoost. 52

LIST OF TABLES

Table 2.1 Literature review summary. 14
Table 4.1 Results of the COSFIRE and VGGFace on the GF and LFW data sets using SVM. 36
Table 4.2 Results of the COSFIRE and VGGFace CNNs on the GF and LFW data sets using XGBoost. . 36
Table 4.3 Results of the fusion of COSFIRE and VGGFace on the GF and LFW data sets using SVM. 37
Table 4.4 Comparison of the results on the GF data set. . 37
Table 4.5 Comparison of the results on the LFW dataset. 38
Table A.1 The division of training and test set of the GF data set. 47
Table A.2 The division of training and test set of the LFW data set. 47
Table B.1 The performance of concatenation and stacking approach on the GF and LFW data set using SVM classifier. 49
Table B.2 The performance of concatenation approach on the GF and LFW data set using XGBoost classifier. 50

ACRONYMS

CNNs Convolutional Neural Networks

COSFIRE Combination of Shifted Filter Responses

LFW Labeled Faces in the Wild

SVM Support Vector Machines

SURF Speeded Up Robust Features

PCA Principle Component Analysis

ICA Independent Component Analysis

LBP Local Binary Pattern

GF GENDER-FERET

Part I

MASTER THESIS

1 INTRODUCTION

The face is a remarkable part of the human body and considered as one of the most important one since it has some distinctive physical and expressive features which allow the identification of certain properties [13]. The miraculous variety of facial features helps humans recognize each other that leads to the formation of complex societies. Since human face provides important biometric features regarding gender, age, ethnicity, and identity, therefore it has been extensively studied in computer vision.

Gender classification is one of the studies that makes use of human faces to distinguish whether a person is a man or a woman. This task is the basis for all advanced applications such as biometric authentication, surveillance and security system, demographic information collection, marketing research, real time electronic marketing, criminology, and augmented reality.

For instance, surveillance system implements gender recognition in order to investigate allegations of illegal behavior. This system process input image from recorded videos in real time. The main challenge of this system is the computational time needed for searching a match between the input face image and the thousand of samples stored in a reference database [13]. One way to reduce the computation time of matching the identity is to first detect the gender. After that, the system can easily match these parameter (input image and gender) with the samples in database. This approach reduces the computation time resulting in faster identification.

Another application of gender recognition is real time electronic marketing. This system aims at showing advertisements on a billboard to the majority of people approaching towards the billboard screen. If the majority are male, the screen shows an advertisement for male. Conversely, if the majority are female the advertisement are adjusted accordingly.

The systems mentioned above are only a few of the implementations of gender recognition. However, recognizing gender from facial images is not an easy task either for the computer or even for the human itself. For instance, a quick study of human behavior in gender recognition was performed by Mahmoud Afifi and Abdelrahman Abdekhamed. From the experiment, they notice that, beside the high importance of isolated facial features, the visual information from the general look of persons also possesses an important role in the classification process regardless of the visibility of facial features [1]. This results was also noticed in prior work by Liand Lu [37].

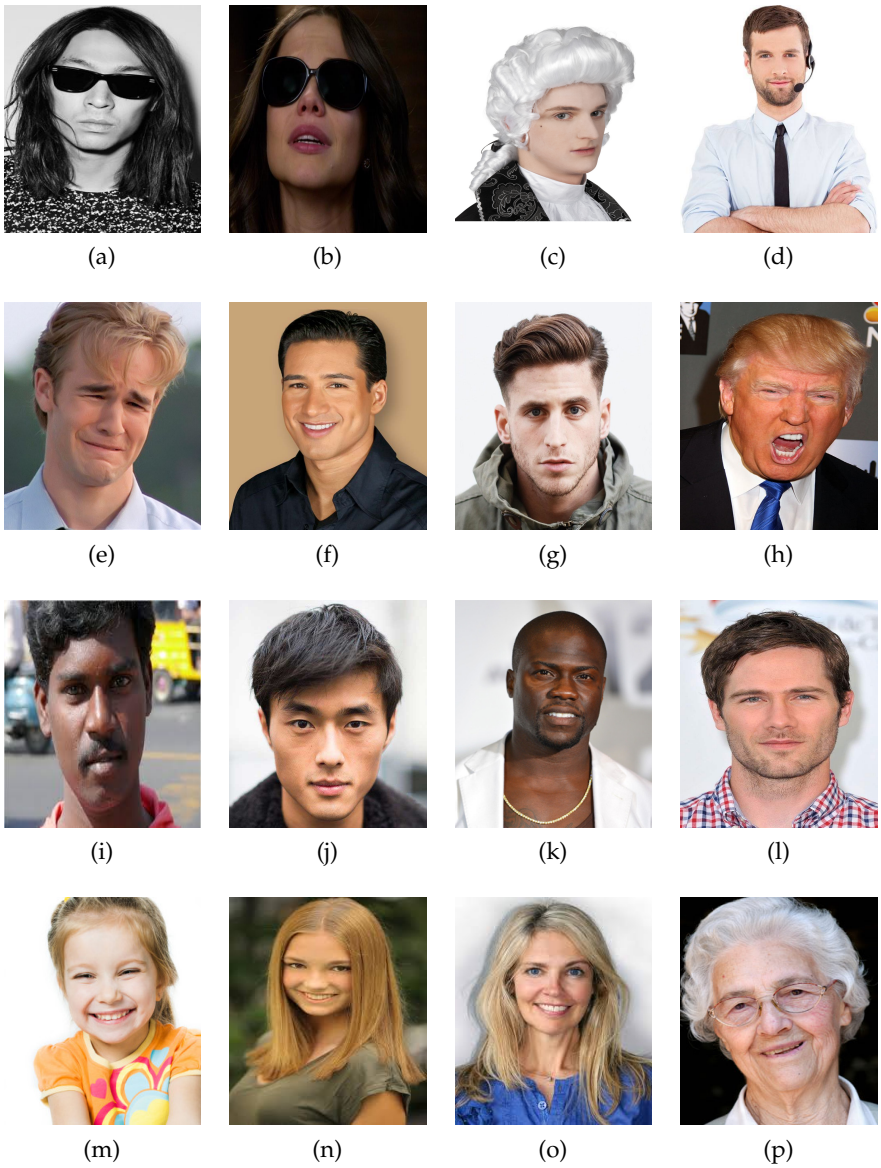


Figure 1.1: Typical problems of gender recognition due to the occlusion with spectacles: (a) (b), wig (c), microphone (d), different expressions: sad (e) happy (f) neutral (g) angry (h), different races: austroloid (i) mongoloid (j) negroid (k) caucasoid (l), and age variation: children (m) adolescent (n) adult (o) eldery (p).

Figure 1.1 shows typical problems of gender classification due to the occlusion with spectacles, wig, and microphone (a,b,c,d), different facial expressions (e,f,g,h), different races (i,j,k,l), and age variation (m,n,o,p). The examples in figure 1.1 (a) and (b) illustrate two images that are somehow difficult to distinguish their gender because the presence of spectacles and hairdo. For some people, the gender of the first model is not easy to infer because both of them are equipped with sun glasses and they also have a long hair. Their eyes are covered as well as some parts of the eye brows which prevent us to discrimi-

nate the differences using eyes and eyebrows. At a glance, we might guess that both of them are female which in fact it is wrong. However, the gender might be obvious and distinguishable to others because it can be differentiated by looking at the facial shape, cheekbones, and jawline. Generally speaking, the shape of the male face is longer and larger than female, the cheek bone is sharp, and the jawline is a bit rougher and prominent which lead to a conclusion that figure 1.1 (a) is definitely male and (b) is female.

Moreover, we also might encounter another problem such as dealing with blurred face images. This is a situation when we can barely see the shape or the area in the frame because it has no distinct outline. As an example of this particular situation, figure 1.2 shows the blurred version of male face and female face of the GF data set. In this case, it is difficult to infer the gender from these images because the intensity distribution of the hair as well as the eyes regions are almost similar. One might decide they are male because of the short hair.

(a) (b)

Figure 1.2: The average face images of (a) male and (b) female from the GF data set.

We have mentioned several applications and some challenges we might encounter in gender classification, and at the time of writing this field has been a well-studied topic in computer vision for many years which motivated researches to create a better algorithm for extracting features from facial images. Recently, Convolutional Neural Networks (CNNs) [14] have become the golden standard for object recognition [33] [47]. CNNs are often used to recognize objects and scenes, and perform object detection and segmentation. They learn directly from image data, eliminating the need for manual feature extraction in which the features are learned directly by the CNNs. In general, the performance of CNNs when dealing with gender classification problems is around 95 - 97% [3] [40].

Another method to extract features from facial image by using trainable filters for key detection is called COSFIRE. This method is automatically configured to be selective for a local contour pattern specified by a prototype [6]. The configuration comprises selecting given channels of a bank of Gabor Filters and determining certain blurs an

shift selected Gabor filters. The performances of this trainable filter are quite promising by achieving approximately 95% accuracy tested on three different cases: the detection of vascular bifurcations in retinal fundus images [9] [11], the recognition of handwritten digits, and the detection of traffic signs in complex scenes [6].

Since both CNNs and COSFIRE are demonstrated to be effective in extracting the most important features of such object, further study is required to investigate whether combining both features would improve the performance of model in gender classification.

1.1 RESEARCH QUESTIONS

The existing studies investigate the performance of extracted features either from CNNs or COSFIRE and subsequently use Support Vector Machines (SVM) as classifier. These approaches are demonstrated to be effective in dealing with classification related problems such as the detection of vascular bifurcations in retinal fundus images, handwritten digits, detection of traffic signs, gender classification, and many other cases. Also, their performances are quite promising and comparable to the state-of-the-art. These facts lead to the main research questions:

1. How will the performance change if the extracted features from both CNNs and COSFIRE are combined and trained with the same classifier (SVM) to infer gender? Will the performance increase or decrease?

2. How will the performance be affected when this approach is applied on constrained and unconstrained data sets which deal with the pose variations, partial occlusion of the face, age, race, and expression?

3. How will the performance be affected using other classifier, for instance, decision tree?

1.2 METHODOLOGY

To answer the questions in section 1.1, we first start with a literature study to discover the state of the art of existing methods which focused on gender recognition. We also review the state of the art of both CNNs and COSFIRE. Moreover, we also take into consideration the data sets that we are going to use in order to be able to compare the results with existing studies. Since we expect that our system should be able to validate both constrained and unconstrained data sets, therefore pre-processing task need to be performed on both data sets to make sure they are valid as input for CNNs and COSFIRE. After that, features will be extracted from both methods and subsequently

merged as a large feature vector. The term *"merged"* refers to appending CNNs features to COSFIRE.

In this study we proposed two techniques in classifying gender:

1. Concatenation technique
   This approach uses the fusion of features from both methods as input by appending the extracted features from CNNs to the features from COSFIRE and subsequently train and validate them using SVM classifier.

2. Stacking technique
   This approach is an extension of concatenation method by using the score vectors generated by SVM. First, we extract the features from CNNs and COSFIRE and train them separately. Then, the score vectors generated from each classifier are merged and used as input for a new SVM classifier.

We validate the performance by comparing the accuracy of CNNs, COSFIRE, and the fusion of CNNs and COSFIRE. We also conduct an experiment to investigate the performance of the concatenation approach using another classifier such as XGBoost decision tree [17].

1.3 THESIS ORGANIZATION

The rest of the thesis is structured as follows. Section 2 presents the related works that correlates with the present study. The experimental setup and its implementation regarding fusion of CNNs and COSFIRE features is described in section 3. Then, the experimental results and the discussion are explained in section 4 and 5 respectively. Finally, the conclusion is drawn in section 6.

2

RELATED WORK

Several studies in relation to gender recognition have been found. These studies are usually based on extracting features from the given images and subsequently use the features to train classifier.

In general, vision-based gender recognition methods can be grouped into three categories: geometric-based, appearance-based, and combination of geometric-based and appearance-based [1]. The geometric-based measures the distance between different key points in the facial image, the appearance-based is based on the pixel-values of facial images[42], while the latter uses the combination of the methods mentioned above.

An overview of existing studies with regard to gender recognition is elaborated in below sections.

2.1 GEOMETRIC-BASED

As mentioned earlier, the geometric-based methods extract and utilize geometric features from the given image to predict gender [1]. This method was firstly introduced by Burton et al. 15 year ago in a paper named *"What's the difference between men and women? Evidence from facial measurement"*. The total of 179 monochrome photographs (91 male and 88 females faces) were used in which all faces were of young adults in a neutral expression. In order to avoid the hair that might conceal the most important parts of the face, all volunteers were asked to wear swimming caps. Also, all males were clean shaved and females asked to not wear any makeup.

The analysis of this study relies on 73 facial points which restricted to simply computes the Euclidean distance between points. Subsequently, the discriminant analysis was used to infer the gender. The results of this experiment shows that with 12 variables, the study was able to demonstrate the level of performance at around 85 %. However, this technique might be unreliable since it attempts to use of a linear combination of variables. Such approach by analyzing the distance between key points is not sufficient to infer the gender. More information from multiple source needs to gather in order to develop a reliable technique which is able to discriminate between males and females.

Likewise, Brunelli et al. [45] also conducted a study based on geometrical features. They extracted 16 geometric features from faces such as eyebrow thickness and pupil to eyebrow separation, as input to HyperBF network to learn the differences between the genders [37].

The result of this experiment using face images of twenty males and twenty females as training set shows an average performance of 79% correct gender classification on images of new faces.

Another algorithm based on geometric technique is called COSFIRE filters. This method was inspired by the mechanism of neurons in the human visual cortex that can be used for key points detection and pattern recognition. The filters are contour based detectors in which the responses calculated as the weighted geometric mean of the shifted responses of simple orientation selective filters [6]. Since the performance of COSFIRE was really impressive in recognition of handwritten digits by achieving 99.48 percent accuracy, the authors of that study attempted to apply the filters to infer gender from the given facial images. First, they extract the features from the images and then train the features with SVM classifier [5].This study demonstrated the effectiveness of COSFIRE filters on GF dataset achieving an accuracy rate of 93.7 %.

In the following year, they extended the study by combining COSFIRE filters with domain-specific so-called Speeded Up Robust Features (SURF). SURF is scale- and rotation-invariant interest point detector and descriptor. It consists of fixing a reproducible orientation based on information from a circular region around the interest point and constructs a square region aligned to the selected orientation, and extract the SURF descriptor from it [15]. In particular, they used SURF to extract the most important features related to eyes, nose, and mouth. It turns out, the fusion of these methods achieves 94.74 % on GF and 99.4% on LFW which outperforms the state-of-the-art.

2.2 APPEARANCE-BASED

The appearance-method extracts features from either the whole face images (holistic features), regions of the face images (local features), or the combination of holistic and local features [1]. This approach is based on the pixel values of the face images. Neural networks is one of the techniques that follows this approach. It trains single or multiple neural networks with image pixels as input.

In 1990, Golomb et al. conducted a research regarding gender recognition using Neural Networks in which the networks was called *"SexNet"*. In this study, they trained a network with 90 training samples of size 30x30. The result of this study shows that the network's average error rate of 8.1% compared favorably to humans, who averaged 11.6% [28]. This study was then followed by Cottrell et al. who performed gender classification task based on face and motions using holons[19], Burton et al. who measured the facial images in three ways: (i) simple distances between key points in the pictures; (ii) ratios and angles formed between key points in the pictures; (iii) three-dimensional (3-D) distances derived by combination of full-face

and profile photographs[16], and Tamura et al. who identified gender with more than 90% accuracy from low frequency components of mosaic 8 × 8 images of the cental part of the human face, which cannot be recognized any more as human faces.[51].

In addition to neural networks, Sun et al. [50] also performed another experiment with Principle Component Analysis (PCA). PCA is a statistical procedure that transforms a number of possibly correlated variables into a smaller number of uncorrelated variables. This study considered the problem of gender classification from frontal facial images using genetic feature subset selection. They used PCA to represent each image as a feature vector in a low-dimensional space and subsequently employed Genetic algorithm to select a subset of features from the low-dimensional representation by disregarding certain eigenvectors that do not seem to encode important gender information. Then, four different classifiers were used to train and classify the features. Turns out, the best performance was obtained using the SVM classifier with an error rate of 4.7%.

Not only did PCA get the attention of researches to be applied on gender classification problem, Independent Component Analysis (ICA) was also taken into account as another approach. Jain and Huang [32] used ICA to represent each image as a feature vector in a low dimensional subspace. In this study, they used frontal facial images so-called GF which consists of 500 images (250 females and 250 males) randomly withdrawn from the facial database.Using a classifier based on linear discriminant analysis (LDA) in a lower dimensional subspace, it achieved an accuracy of 99.3%.

Another approach of appearance-based method is Local Binary Pattern (LBP). It is a type of visual descriptor used for classification in computer vision particularly the case of texture spectrum model. Hadid et al. [30] applied this approach by firstly reviewing 13 recent and popular local binary patterns variants on two different problems (gender and texture classification) using benchmark databases. From the experiments, they found out that basic LBP provides good results and generalizes well to different problems.The best results were obtained with binarized statistical image features (BSIF) however it has a downside regarding the cost of higher computational time.

Subsequently, this approach was also followed by Moeini and Mozafarri [41]. They proposed to learn separated dictionaries for male and female genders for representing the gender in facial images by using LBP to extract 64 features of the face.During the training process, they define two dictionaries to learn the defined dictionaries and then the Sparse Representation Classification (SRC) was employed for classification in the testing process. After validated using three public databases, GF, LFW and Groups databases, they obtained convincing results which were comparable to several state-of-the-arts.

Recently, the presence of deep neural networks has caught so much attention of researches to apply this method on gender classification problems because of its performance in achieving remarkable improvements on accuracy. A simple CNNs was applied by Levi and Hassner [36] to the Adiance benchmark [21] for age and gender classification in a holistic manner [1]. The results of this experiments outperforms the current state-of-the-art method by achieving around 86% accuracy. Another experiment using minimalistic CNN-based ensemble model was also conducted by Antipov et al [3]. They trained 3 instances of CNNs in which each instance was trained from scratch with a random initialization of weights. Then, those instances were combined into a single ensemble model by averaging the outputs of softmax layers. The performance of CNNs applied on LFW dataset was 97.31%.

Moreover, Mansanet et al. [40] used a local deep neural networks (Local-DNN) which is based on local features and deep architecture. By using a standard DNN, this model was trained to classify small patches extracted from images. A simple voting was carried out to the final classification by taking into account the contributions from all patches of the image. The best model of this approach was DCNN by achieving 96.25% accuracy applied on LFW dataset.

2.3 GEOMETRIC AND APPEARANCE-BASED

The Geometric and Appearance-based approach is the combination of both Geometric and Appearance method mentioned earlier.This approach was introduced by Mozaffari et al. [42] in a paper named *"Gender Classification Using Single Frontal Image Per Person: Combination of Appearance and Geometric Based Features"*. They proposed a new method which is based on single frontal image per person bu utilizing Discrete Cosine Transform (DCT) and Local Binary Pattern (LBP) from appearance-based approach and geometrical distance feature (GDF) based on physiological differences between male and female faces. The fusion of these three methods achieved around 95% accuracy which is 12% higher than the performance of when DCT and LBP were combined.

Tapia et al. also proposed a new method based on fusion of different spatial scale features selected by mutual information from histogram of LBP, intensity, and shape [52]. In order to select features, they employed four different features: minimum and maximal relevance (mRMR), normalized mutual information feature selection (NMIFS), conditional mutual information feature selection (CMIFS), and conditional mutual information maximization (CMIM).

In the first experiment, they applied the model on GF and UND dataset. The best result was obtained from GF with 99.13% accuracy using the fusion of 18,900 selected features. In the second experiment,

they tested the model on LFW dataset with the fusion of 10,400 features from 3 different spatial scales. They obtained a classification rate of 98.01%.

2.4 STATE-OF-THE-ART SUMMARY

Three methods of vision-based gender recognition were explained already in the previous sections. Thus, the existing studies can be summarized as shown in Table 2.1. The performances listed in the table show the performance of appearance-based approach is higher than the geometric one with an accuracy of above 90%. However, when both geometric- and appearance-based were combined, the performance of the model improves as explained in [40].

The existing studies also indicate that LFW and GF are the most popular dataset ever used on gender classification. The LFW dataset is used to validate the model on unconstrained environment while the GF is used to test facial images without the presence of noise in the frame. By far, the best approach is able to reach almost an accuracy of 100% which tested on GF and 98.01% on LFW dataset.

Considering the superiority of CNNs in recognizing objects and scenes and the capability of COSFIRE filters in detecting keypoints and its selectivity for a local contour pattern, therefore we aim at combining these methods to infer gender. In this study, we propose two techniques namely concatenation technique and stacking technique. The concatenation technique appends the extracted features from CNNs to the features from COSFIRE while the stacking technique makes use of the score vectors generated by SVM from each method as input features. Thus, by applying the fusion of CNNs and COSFIRE, these techniques are classified as the geometric and appearance-based.

In order to validate the proposed approaches, the LFW and GF data sets are going to be applied on our proposed methods so the results can be comparable to the existing studies. The details of the implementation will be elaborated in the following chapters.

Table 2.1: Literature review summary.

Group	Method	Dataset	Accuracy (%)
Geometric-based	Burton et al. [16]	179 monochrome photographs	85
	Brunelli et al. [45]	Face images of 20 males and 20 females	79
	COSFIRE (Azzopardi et al.) [5]	GF	93.7%
	Fusion of COSFIRE and SURF (Azzopardi et al.) [13]	GF	94.74
	Fusion of COSFIRE and SURF (Azzopardi et al.) [13]	LFW	99.4
Appearance-based	Neural Networks (Golomb et al.)	90 images	93.90
	Neural Networks (Tamura et al.) [51]	Low frequency of mosaic images	90
	PCA (Sun et al.) [50]	400 frontal images from 400 distinct people	95.3
	ICA (Jain and Huang) [32]	GF [32]	99.3
	LBP (Moeini and Mozzafari) [32]	GF	91.9
	[32]	LFW	94.9
	CNN (Antipov et al.) [3]	LFW	97.31
	CNN (Mansanet et al.) [40]	LFW	96.25
Geometric and Appearance-based	DCT+LBP+GDF (Mozaffari et al.) [42]	Ethnic and AR databases	96.5
	Fusion of different spatial scale features (Tapia et al.) [52]	GF	99.13
	[52]	LFW	98.01

3

METHODOLOGY

In this chapter, we present our proposed approach to addressing the gender classification problem. First, we start giving an overview of pre-processing steps using popular techniques namely Viola-Jones [55] and facial landmark tracking [53]. These methods are used to detect and align a face in a given image so the most relevant parts can be obtained, adjusted, and resized accordingly. Then, we describe CNNs followed by an elaboration of one of the most widely used CNNs architecures for face recognition so-called VGGFace [43].

Moreover, we also give an overview of COSFIRE filters including a detailed explanation in dealing with gender classification problems. Finally, we present the architectures of our proposed system which are grouped into concatenation and stacking architecture. The corresponding architectures show the pre-processing steps until leading to the final classification decision. Figure 3.1 shows a high level diagram of the proposed methods.

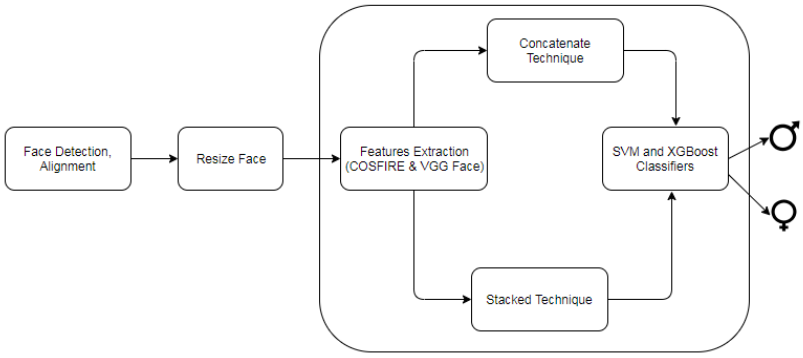


Figure 3.1: A high level diagram of the proposed methods.

3.1 FACE DETECTION AND ALIGNMENT

Face detection is a technique to identify human faces in digital images. It is commonly used in many computer systems as a pre-processing task in order to obtain the most relevant parts of an image. One of the most popular techniques that is used widely in image classification problems is Viola-Jones algorithm. It is a detection framework that is capable of processing images extremely rapidly while achieving high detection rates [55]. This framework was constructed with three important keys:

- Introducing a face detector which is able to compute very fast (Integral Image)
- A very simple and efficient classifier should be used in selecting critical visual features from a large set of potential features. This framework uses Adaboost [23] learning algorithm as classifier.
- This framework combines classifiers which allows to discard the background of the image and focus on face regions resulting in fast computation.

After performing a set of experiments and validating the algorithm on data set under a very wide range of condition including: illumination, scale, pose, and camera variation, this framework achieved its best performances comparable to the best methods from the previous studies. In this study, we use Viola-Jones algorithm in order to detect face from the given images before further processing.

Another preprocessing task that we perform is face alignment which was proposed by Uricar et al. [53]. This framework works by using Viola-Jones algorithm to detect faces in an image and subsequently applying facial landmark tracking. The purpose of this method is to detect a set of 51 facial landmarks from a given facial image. Originally this algorithm is able to find 68 fiducial points but since Viola-Jones algorithm sometimes excludes 17 points which belongs to face contour, therefore they are not taken into account as important features [13]. After obtaining the fiducial points, the average location of the two sets of eye-related landmarks is computed which gives us the opportunity to define the orientation of the line and connect the corresponding lines(the angle of the face). We can use the angle to align the face image horizontally and subsequently rotate the image around the center of the line as shown in figure 3.2.

(a) Original image (b) After applying facial landmark tracking

Figure 3.2: Representation of the face alignment algorithm. The positions of the facial landmarks are indicated by the 51 red dots and the tree blue markers indicate the left eye center, the center of the line that connects the two eye, and the right eye centers.

The output of the pre-processing after applying facial landmark tracking and Viola-Jones algorithm on the LFW data set is shown in Figure 3.3.

(a) Input image after applying facial landmark tracking

(b) After applying Viola-Jones face detector

(c) Cropped face

Figure 3.3: Pre-processing using Viola-Jones face detector on LFW data set.

3.2 NEURAL NETWORKS-BASED CLASSIFIER

Neural networks have been a field of research for more than six decades now. Throughout their history, neural networks have had a typical architecture; multiple layers of interconnected nodes, representing biological neurons.They generally have input layer, hidden layer, and output layer. By using weights and inter-node connections, the values of output nodes are calculated from layer to layer. The idea behind Neural Networks is to find optimal weights on each layer of the networks which normally adjusted using forward and backward propagation [38]. The forward propagation provides initial information to the hidden units at each layer and finally produce the output [18] while the back-propagation algorithm relies on the chain rule of differentiation for making a connection between the loss computed at the output layer and any hidden nodes. The connection of networks helps to relay the final loss of the network back to any earlier layers so that the weights of those layers may be proportionally adjusted [54].

There are several types of Neural Networks: Conventional Neural Networks, Recurrent Neural Networks [39], and the most recent one

is Convolutional Neural Networks or commonly called as CNNs. We will elaborate more about CNNs in the following subsection since this method is part of our study.

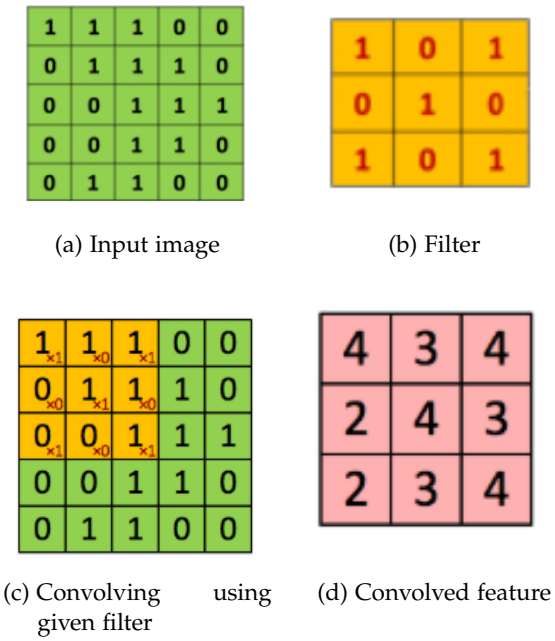
3.2.1 *Convolutional Neural Networks (CNN)*

One type of Neural Networks which so popular for object detection is CNNs. It was firstly introduced by K.Fukushima in 1983 as neocognitron [24] inspired by the feline visual processing system. In the following years it was enhanced by several researches and the most impressive work was done by Yann LeCun et al. who helped establish how we use CNNs today—as multiple layers of neurons for processing more complex features at deeper layers of the network [34]. Thus, CNNs have been applied on so many fields such as image recognition, video analysis, natural language processing, and drug discoveries.

By definition CNNs is a neural network where a signal feeds into a set of stacked convolutional pooling layer pairs, and the output of the last layer feeds into a set of stacked fully connected layers that feed into a softmax layer [54]. CNNs consist of four main operations: convolutional, non linearity (ReLU), pooling or sub sampling, and classification or fully connected layer. In order to form an architecture, these operations are divided into four forms of layer namely convolutional layer, sub-sampling layer, rectified linear unit (ReLU), and fully connected layer.

Convolutional layer This layer derives its name from the convolution operator. The primary purpose of this layer is to extract features from the given input images. Then, the spatial relationship between pixels is preserved on this layer by learning image features using small squares of input data. The weights on these connections are similar for each node in the convolutional layer. This causes the weights to have the same effect as a convolution kernel. The weight behaves like a filter in an image extracting particular information from the original image matrix [4]. Figure 3.4 shows the process of convolution on images.

Sub-sampling layer The purpose of this layer is to reduce dimensionality of each feature map without discarding the most important features. Sub-sampling or pooling is usually placed in between convolutional layers and it is done independently on each depth dimension, therefore the depth of the image remains unchanged [4]. CNNs has three types of pooling: average, max, and stochastic pooling [35], however the most commonly used is max-pooling which takes the maximum of input values as can be seen in figure 3.5.



(a) Input image (b) Filter

(c) Convolving using given filter (d) Convolved feature

Figure 3.4: The process of convolution on images taken from [2]. During the convolution process (c), the filter (b)is applied over the window of the input image (a) which finally constructs a convolved features matrix of size 3×3. (d).

Rectified linear unit (ReLU) ReLu is an activation function used by neurons which is defined as:

$$f(x) = max(0, x) \tag{3.1}$$

where x is the input to a neuron. It is preferably used in CNNs over other activation functions because it can be computed more efficient without making a significant difference to generalisation accuracy compared to conventional activation functions like the sigmoid and hyperbolic tangent. Moreover, it is also used instead of a linear activation function to add non linearity to the network so not only is the linear function computed but also the non linear.

Fully connected layer The fully connected layer implies every neuron in the previous layers is connected to every neuron in the next layer. This layer uses an activation function called "softmax" [48]. The aim of fully connected layer is to classify the input into various classes based on the training data set. An example of this layer is shown in figure 3.6.

Putting all layers together, Figure 3.7 shows an example of a complete CNNs architecture. It starts with an input image followed by subsampling, then another convolution and sub sampling layer, and finally a fully connected layer which acts as a classifier.

Nowadays there are many existing CNNs architectures available, starting from the very simplest one until the most complex design.

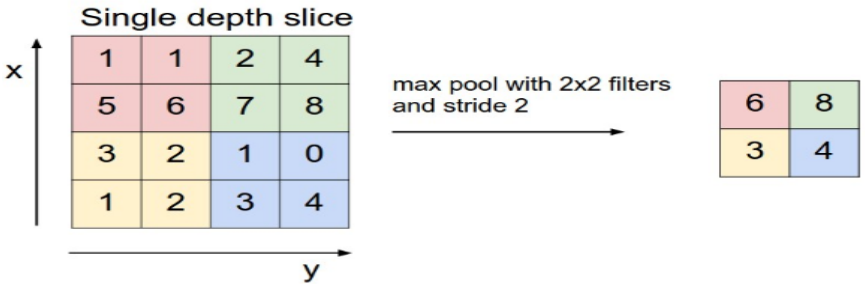


Figure 3.5: An example of max pooling taken from [2]. A pooling with a stride of 2 is applied on the convolved features which produces a matrix of size 2×2.

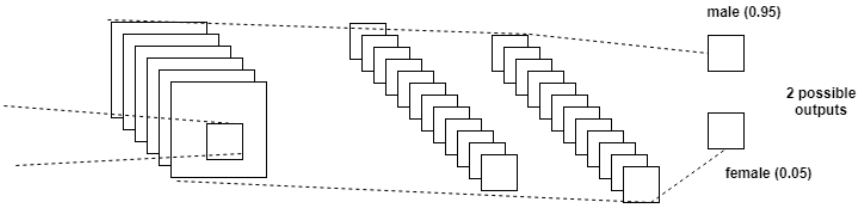


Figure 3.6: An example of fully connected layer with 2 possible outputs either male or female.

The most common architectures are Lenet, Alexnet, GoogLeNet, VGGNet, and ResNet. Since the main focus of this study is to infer gender from facial images, therefore we will not explain the details of the architectures mentioned above. In particular, we just elaborate an extension of VGGNet [47] CNNs architecture which was designed for face recognition so-called VGGFace [43]. The details of VGGFace [43] as one approach of this study is given in below section.

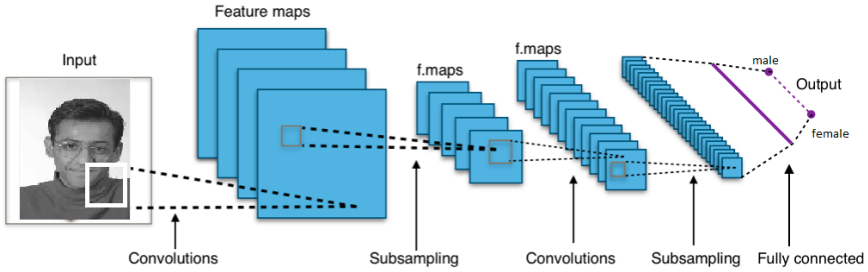


Figure 3.7: The architecture of convolutional neural networks.

3.2.2 *VGGFace*

VGGFace is an extension of VGGNet developed by Parkhi et al. from the University of Oxford [43] in 2015 which contributes in showing that the depth of the network is a critical component for good perfor-

mance. The goal of VGGFace architecture is to deal with face recognition either from a single photograph or a set of faces tracked in a video [43]. Since this work was inspired by VGGNet, therefore VGGFace architecture is almost similar to its ancestor as shown in figure 3.8.

layer	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
type	input	conv	relu	conv	relu	mpool	conv	relu	conv	relu	mpool	conv	relu	conv	relu	conv	relu	mpool	conv
name	–	conv1_1	relu1_1	conv1_2	relu1_2	pool1	conv2_1	relu2_1	conv2_2	relu2_2	pool2	conv3_1	relu3_1	conv3_2	relu3_2	conv3_3	relu3_3	pool3	conv4_1
support	–	3	1	3	1	2	3	1	3	1	2	3	1	3	1	3	1	2	3
filt dim	–	3	–	64	–	–	64	–	128	–	–	128	–	256	–	256	–	–	256
num filts	–	64	–	64	–	–	128	–	128	–	–	256	–	256	–	256	–	–	512
stride	–	1	1	1	1	2	1	1	1	1	2	1	1	1	1	1	1	2	1
pad	–	1	0	1	0	0	1	0	1	0	0	1	0	1	0	1	0	0	1

layer	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37
type	relu	conv	relu	conv	relu	mpool	conv	relu	conv	relu	conv	relu	mpool	conv	relu	conv	relu	conv	softmx
name	relu4_1	conv4_2	relu4_2	conv4_3	relu4_3	pool4	conv5_1	relu5_1	conv5_2	relu5_2	conv5_3	relu5_3	pool5	fc6	relu6	fc7	relu7	fc8	prob
support	1	3	1	3	1	2	3	1	3	1	3	1	2	7	1	1	1	1	1
filt dim	–	512	–	512	–	–	512	–	512	–	512	–	–	512	–	4096	–	4096	–
num filts	–	512	–	512	–	–	512	–	512	–	512	–	–	4096	–	4096	–	2622	–
stride	1	1	1	1	1	2	1	1	1	1	1	1	2	1	1	1	1	1	1
pad	0	1	0	1	0	0	1	0	1	0	1	0	0	0	0	0	0	0	0

Figure 3.8: The architecture of VGGFace [43].

From the above image it can be seen that VGGFace consists of 13 convolutional layers, 15 Rectified Linear Units (ReLu), 5 sub sampling (max pooling), 3 fully connected layers, and 1 softmax probability. The input to all network is a face image of size 224×224 pixels.

The convolutional layers are divided into 5 groups. The first group contains 64 filters and this number increases by similar size (64) in the second group. In the following group, the filters increases by 128, then two times this number in the fourth group, and finally it stays at 512 in the last group. Between each convolution layer and fully connected layer, a ReLu is placed.

Moreover, a max pooling layer with the stride of 2 is placed after the convolution layer on each group which makes it in total 5 max pooling layers. Three fully connected layers are also placed after the last pooling layer in which a ReLu is present in between. The first and the second fully connected layers contain 4096 features while the latter 2622 features only. The last layer of the architecture is a softmax probability which aims at classifying images into classes.

Not only does this architecture allow us to perform classification but also it supports feature extraction. Features can be extracted from each fully connected layers. Usually, the features are extracted the second fully connected layer (fc7) which contains 4096 features.

In this study, we use pre-trained VGGFace to extract features from a fully connected layer. First, we apply face detection and alignment algorithm as the pre-processing steps and then we resize the face images to the size of 224×224 pixels. After that, they are fed into networks as input and subsequently we extract the features from the second fully connected layer (fc7). The output of this layer is 4096 features from each input image which we normalize in the range between zero and one. The normalization is needed in order to avoid bias between the features of VGGFace and COSFIRE. Finally, we use

the extracted features to train SVM classifier in order to infer gender. An overview of VGGFace as a feature extractor is shown in figure 3.9.

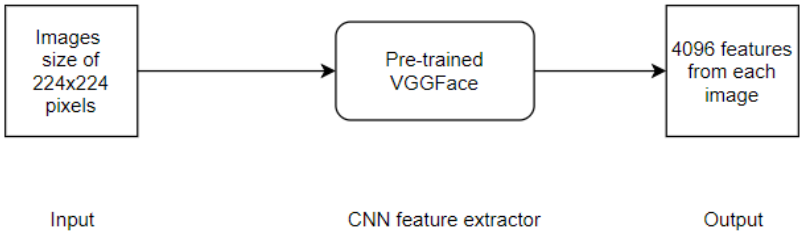


Figure 3.9: An overview of VGGFace as feature extractor in gender classification.

3.2.3 *VGGFace Classification Model*

We use the extracted features from the images in a given training set to learn an SVM classification model namely Compact Classification ECOC for support vector machines [1]. It is a compact, multiclass,and error-correcting output codes (ECOC) model which returns a compact ECOC model composed of linear classification models. This SVM model is very well known for its performance in fitting multiple class in classification tasks.

Another classifier so-called eXtreme Gradient Boosting (XGBoost) decision tree is also employed in this study. This classifier is an implementation of gradient boosted decision trees designed for speed and performance. We use XGBoost as additional classifier because it is generally fast and it dominates structured or tabular data sets on classification and regression predictive modeling problem [2].

3.3 COSFIRE-BASED CLASSIFIER

Trainable COSFIRE filters have been taking into consideration as an effective approach for key point detection and pattern recognition. This approach was proposed by Azzopardi et al. in 2013 and it has been used and developed extensively ever since in dealing with several problems on different fields such as localization and detection of traffic signs, recognition of handwritten digits [6], contour detection of Vascular Bifurnications [10] [26] [12] [49], and vessel segmentation [9] [11]. Recently, they also used this approach in dealing with gender recognition problems as mentioned in [13] and [7]. The trainable COSFIRE filters are considered to be useful because it is automatically

1 https://nl.mathworks.com/help/stats/compactclassificationecoc-class.html
2 https://machinelearningmastery.com/gentle-introduction-xgboost-applied-machine-learning/

configured to be selective for a local contour pattern specified by a prototype.

The following sub sections explain the details of how the filters works, how to configure, train, and apply them on images.

3.3.1 *COSFIRE Method and 2D Gabor filters*

As previously mentioned that COSFIRE filters are contour based detectors. This method is a trainable filter in which the pattern of such prototype is automatically analyzed. The obtained pattern is subsequently applied to images in order to localize pattern which similar to the given prototype [13]. The response of these filters are calculated as the geometric mean of the shifted response of simples orientation selective filters. In order to be able to obtain the shifted responses, the support at different locations are combined to obtain sophisticated filters. Then, geometric mean is applied on this computation in order to obtain the response of COSFIRE filters. This approach did not employ arithmetic mean while computing the responses because the geometric mean is considered to be very resistant to contrast variation and the multiplications of responses from the sub-units are sensitive to different parts of the curves. It helps the filters to generate the responses only when all elements of the pattern of interest are present.

In brief, the trainable COSFIRE filters works in three main steps. First, it applies the selected Gabor filters on image with size of 128×128 pixels whose output goes through Gaussian blurring. Then, the responses of Gaussian blurring are shifted by distinct vector. And finally the shifted responses are multiplied in order to calculate the weighted geometric mean which determines the final response.

The initial step of COSFIRE filters which considered to be important is the detection of orientation using 2D Gabor filters. A Gabor filter is made by modulating sinusoid by a Gaussian and it is considered to be important in this work because its selectivity to texture representation and discrimination which happen to be the core of COSFIRE filters. Once the Gabor filter has been applied on the images, the responses of the filters are normalized in order to keep the sum of all positive responses to 1 and negative responses to -1. Then, all the responses are thresholded t_1 of the maximum response $g_{\lambda,\theta}(x,y)$ for the combination of values (λ,θ) at every point (x,y) of the image.

In this study, the original Gabor-based type of COSFIRE filters are used because color is not considered to be a distinctive feature to recognize gender. The basic working principle of COSFIRE filters can be seen in [6].

3.3.2 *COSFIRE Filter Configuration*

After applying the selected Gabor filters on images, the responses are used as an input for COSFIRE filter. Each of these Gabor filters is defined by parameter values (λ_i,θ_i) around each of the points (ρ_i,ϕ_i) with respect to the center of COSFIRE filter. These four parameters (λ_i,θ_i,ρ_i,ϕ_i) represent the properties of the contour in the region of a given point of interest. The width is represented by $\lambda_i/2$, the orientation by θ_i, while the latter ρ_i and ϕ_i indicate the location of specified area. At each of the positions along the circle, the maximum of all responses for all the possible values of (λ,θ) that are used in the bank of filters is considered. The positions of which the values are higher than the corresponding values in the nearby positions along an angle $\phi/8$ are chosen as the most dominant points in the region of interest. Then, the polar coordinates (ρ_i,ϕ_i) are computed for all these values.

Finally, the parameters values of all points is grouped into a set of 4-tuples:

$$S_f = \{(\lambda_i, \theta_i, \rho_i, \phi_i) | i = 1...n\} \tag{3.2}$$

The subscript f represents the local prototype pattern around the region of point of interest and n denotes the number of local maximum points. Every tuple in the corresponding set determines the parameters of some contour part in f [6].

Figure 3.10 shows the configuration of COSFIRE filters by using parts of the eyebrow as prototype patterns selected from a male face images of GF.

3.3.3 *COSFIRE Filter Response*

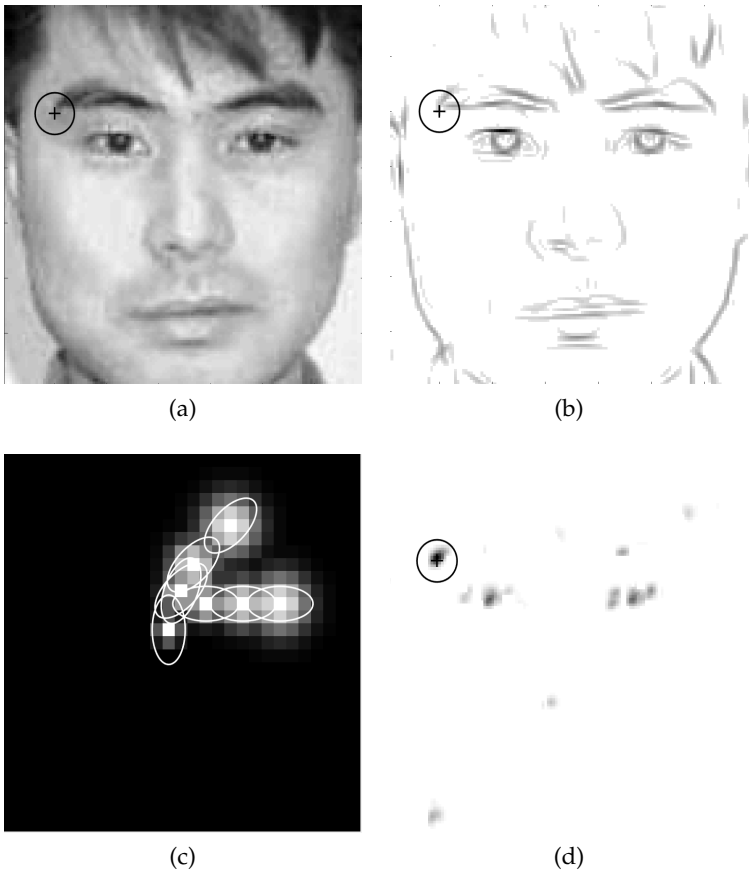
As mentioned earlier, the response of a COSFIRE filter is a geometric mean of all the responses of the thresholded Gabor filter responses. However, the responses from COSFIRE filters are first blurred in order to bring some tolerances in the position of corresponding contour parts using a Gaussian function $G_\sigma(x,y)$ as shown in equation (3.2).

$$\sigma = \sigma_0 + \alpha\rho \tag{3.3}$$

The values of σ_0 and α are constant with the default value of σ_0 = 0.67 and α = 0.1 as proposed in [6]. The orientation bandwidth can be increased by adjusting the value of α. Morover, all the responses are shifted by the polar vector (ρ_i, -ϕ_i) to bring all afferent responses towards the center of the filter.

Finally, all the blurred and shifted Gabor responses are combined using a geometric mean function described by s_f:

$$r_{sf}(x,y) = |(\prod_{i=1}^{|sf|}(s\lambda_i, \sigma_i, \rho_i, \phi_i^{(x,y)})^{w_i})^{1/\sum_{i=1}^{sf} w_i}|_{t_3} \tag{3.4}$$



(a) (b)

(c) (d)

Figure 3.10: Configuration of COSFIRE filters using a training male face GF. The encircle regions in a training image of size 128×128 (a) shows the pattern of interest that is used to configure COSFIRE filters. Figure (b) shows the reverse response of Gabor filters with 16 orientations ($\theta = \{0, \phi, ...15\phi/8\}$ and a single scale (λ=4). Figure (c) describes the structure of COSFIRE which is selective for prototype shown in figure (a). Figure (d) shows the inverted response maps of COSFIRE filters to the input image in (a)

where $|.|_{t_3}$ denotes the that the response is thresholded at a fraction t_3 of the maximum value of across all the coordinates(x,y) of the image.

In the original paper [6], it also proposed the ability of COSFIRE to tolerate the rotation, scale, and reflection by adjusting the parameter properly. However, these invariances are not necessary for this study.

3.3.4 *Face Descriptor*

After applying the COSFIRE filters on a given test image, a spatial pyramid of three levels is subsequently used to obtain the face descriptors. A face descriptor is formed using the maximum responses

of all COSFIRE filters across the entire image which are selective for different regions of a face.

In level zero, only one tile is considered but in the following levels each COSFIRE response map of COSFIRE filters is divided into $(2 \times 2 =)4$ and $(4 \times 4 =)16$ tiles respectively. From the total of 21 tiles of a spatial pyramid which are obtained from the summation of the three levels, we take the maximum value of each COSFIRE filters. That said, for n COSFIRE filters and 21 tiles, we describe a face image with a 21n-element feature vector.

Moreover, the set of n COSFIRE filter maximum responses per tile is normalized to unit length [13]. An example of the COSFIRE face descriptor using a single filter is shown in figure 3.11.

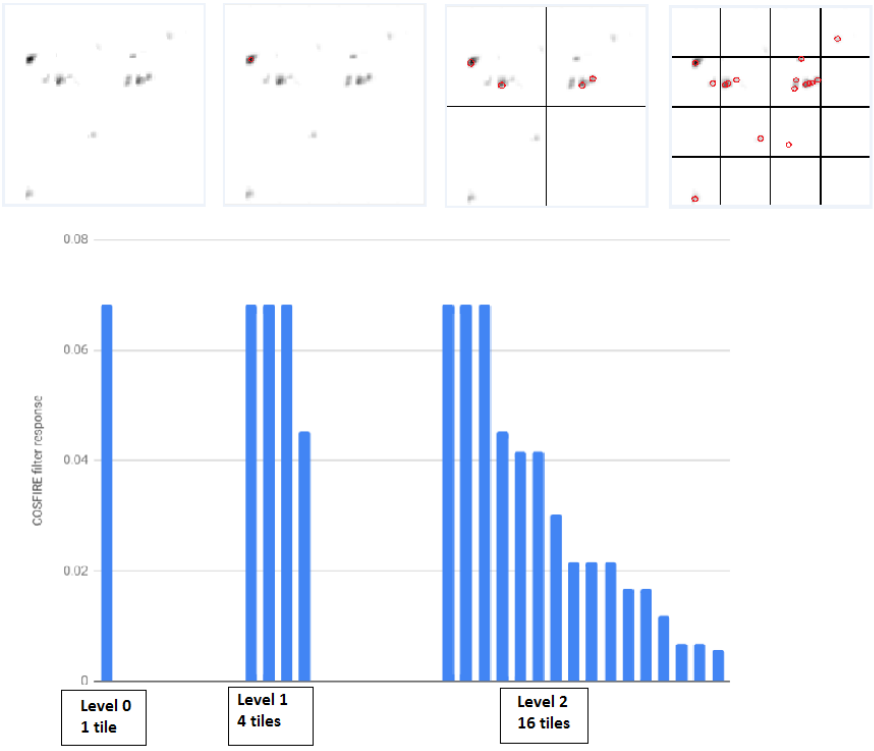


Figure 3.11: Application of the COSFIRE filters on a face image. In level zero only one tile is considered. In level one we consider four tiles in a 2×2 spatial arrangement and in level two we consider 16 tiles in a 4×4 grid.

3.3.5 *COSFIRE Classification Model*

In order to classify gender based on the extracted features from COSFIRE filters, SVM Compact ClassificationECOC for support vector machines (ECOC) is employed. It returns a compact ECOC model composed of linear classification models by fitting multiple classes. Moreover, we also employ XGBoost tree [17] to verify whether the performance of the COSFIRE filters would be affected or not.

3.4 THE PROPOSED METHODS

In this study, we propose two techniques to combine the decision made by the trainable COSFIRE filters and VGGFace CNNs using concatenation and stacking classification technique. The details of each technique are explained as follows:

3.4.1 *Fusion of CNN and COSFIRE features by concatenation approach*

The concatenation technique is an approach that uses the fusion of features from both methods as input for new classifiers by appending the extracted features from VGGFace CNNs to the features of COSFIRE. The total of 4096 features extracted from VGGFace and 5040 features from COSFIRE are merged into a large feature vectors that makes the fusion of features from both methods as much as 9136 and these features are subsequently fed as an input for both classifiers. We use SVM Compact Classification ECOC and XGBoost multi class as classifiers since they perform well in classifying multiple classes.

Furthermore, the output of each classifier is used to compare the performance of the proposed method with each individual approach (VGGFace and COSFIRE) as well as the performance of existing studies. An overview of the concatenation method architecture is shown in figure 3.12.

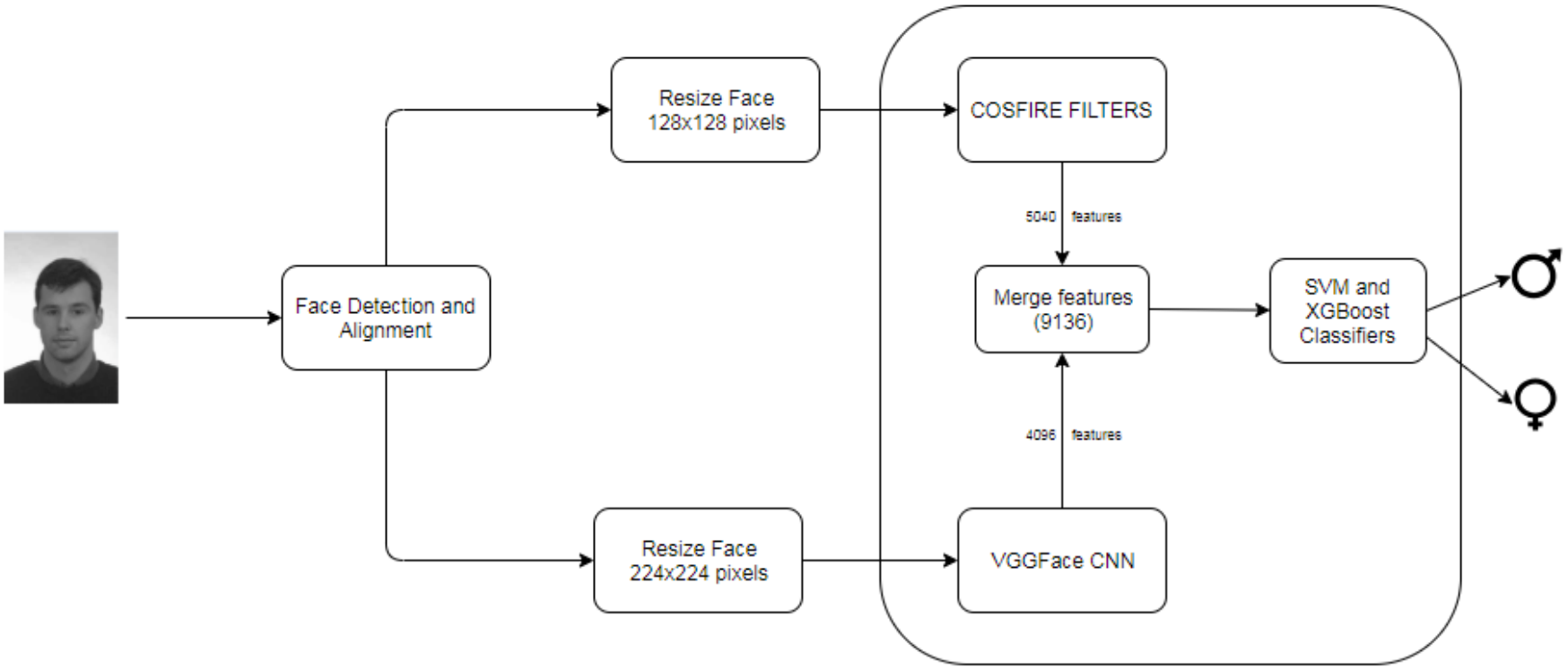


Figure 3.12: Fusion of CNN and COSFIRE features by concatenation approach.

3.4.2 *Fusion of CNN and COSFIRE features by a stacking approach*

Another approach called stacking technique is also employed in this study as one of the proposed methods. This approach is slightly different compare to the previously explained technique because it considers the score vector generated from each SVM classifier as an input for a new model instead of the merged features. First, the features generated from both COSFIRE filters and VGGFace CNNs are trained separately using SVM classifier. Then, the score of each input feature belongs to either male or female is used as input for new classifier. The size of this vector is $n \times 2$ where n denotes the total number of inputs and 2 represents the total number of features (the score as male and female). Then, the score vectors from both methods are merged which create a new feature vector of size $n \times 4$. Moreover, the new feature vector is used as input for a new SVM classifier in order to classify gender (See figure 3.13).

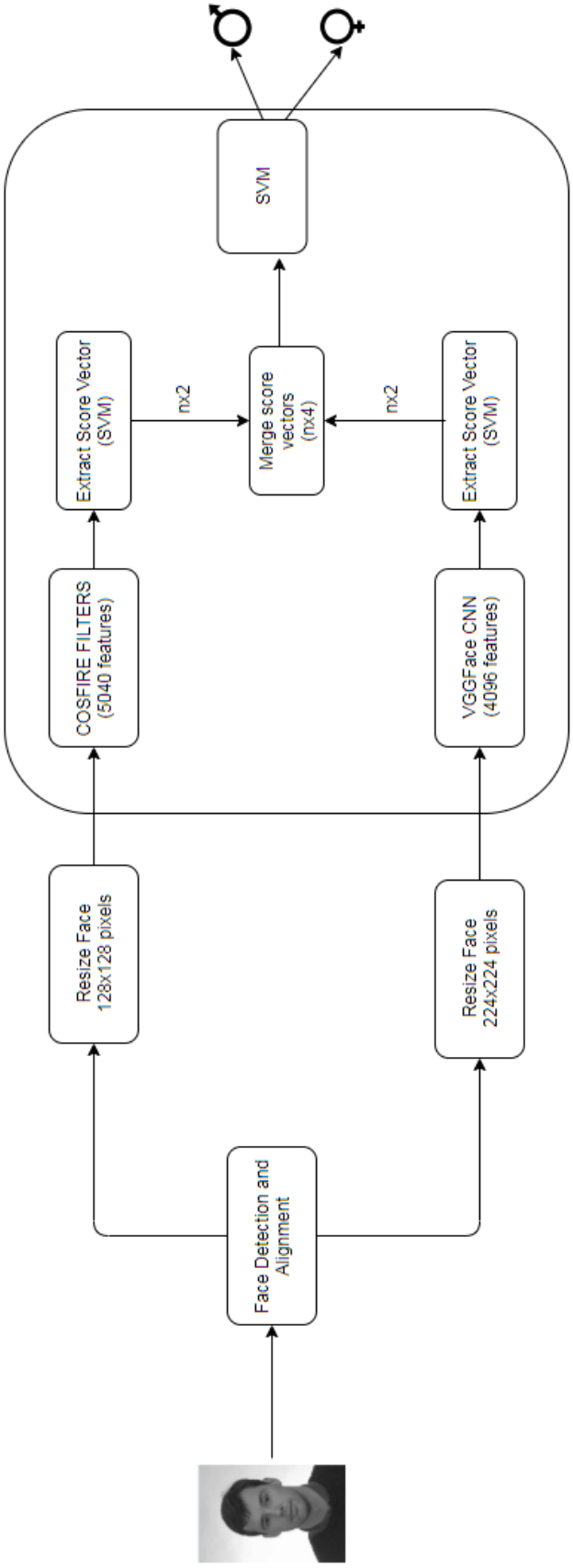


Figure 3.13: Fusion of CNN and COSFIRE features by a stacking approach.

4

EXPERIMENTAL RESULTS

In this chapter we report the experimental setup and summarize the results. First, we provide details on the data sets we used for the evaluation of the proposed methods namely GENDER-FERET (GF) and Labeled Face in the wild (LFW). Then, we briefly explain the pre-processing steps of the corresponding experiment.

Moreover, we report the results regarding the performance of each method and the fusion of CNNs and COSFIRE. We explain the results based on the proposed architectures explained in the previous section.

4.1 DATASETS

In this study we used two data sets namely GENDER-FERET [27] and Labeled Face in the wild [31]. These are two of the standard data sets used by most researchers in evaluating the performance of the proposed methods and to be able to compare the results with the state-of-the-art.

The GENDER-FERET data set (GF) was collected by Dr. Harry Wechsler at George Mason University [44]. This data set consists of the images of people with different expression, age, race, and pose in a controlled way. It is considered as a constrained data set because only one face appears in each frame.

In our experiment, we used 946 GF images which were divided into two parts: training and testing sets. A balanced number of training and test sets as published in [8], [13], and [5] were applied on the division of data sets. The training set consists of 474 images (237 males and 237 females) while the test set contains 472 images (236 males and 236 females). The examples of male and female images from GF data set can be seen in figure 4.1.

Moreover, we also used Labeled Face in the Wild data set (LFW) in order to verify one of the objectives of this study that is validating the proposed methods on images with unconstrained environment. This data set is maintained by the university of Massachusetts which was designed for studying the problem of unconstrained face recognition [1]. It consists of 13,000 images of 5,749 celebrity and politician faces collected from the websites on internet. This data set is classified as one of the most challenging ones in image classification tasks since the images were taken when the subjects doing their daily activities such as playing sports, doing a fashion show, giving a speech or cam-

1 http://vis-www.cs.umass.edu/lfw/

(c)

Figure 4.1: Example male (a,b) and female (c,d) images from the GF data set.

paining, doing an interview, and others. Looking at the facts that the environment of this data set is random, the face may appear more than one in the frame. Also, the illumination, background, age, expression, and race are varied. Figure 4.2 shows examples of LFW data set for both male and female.

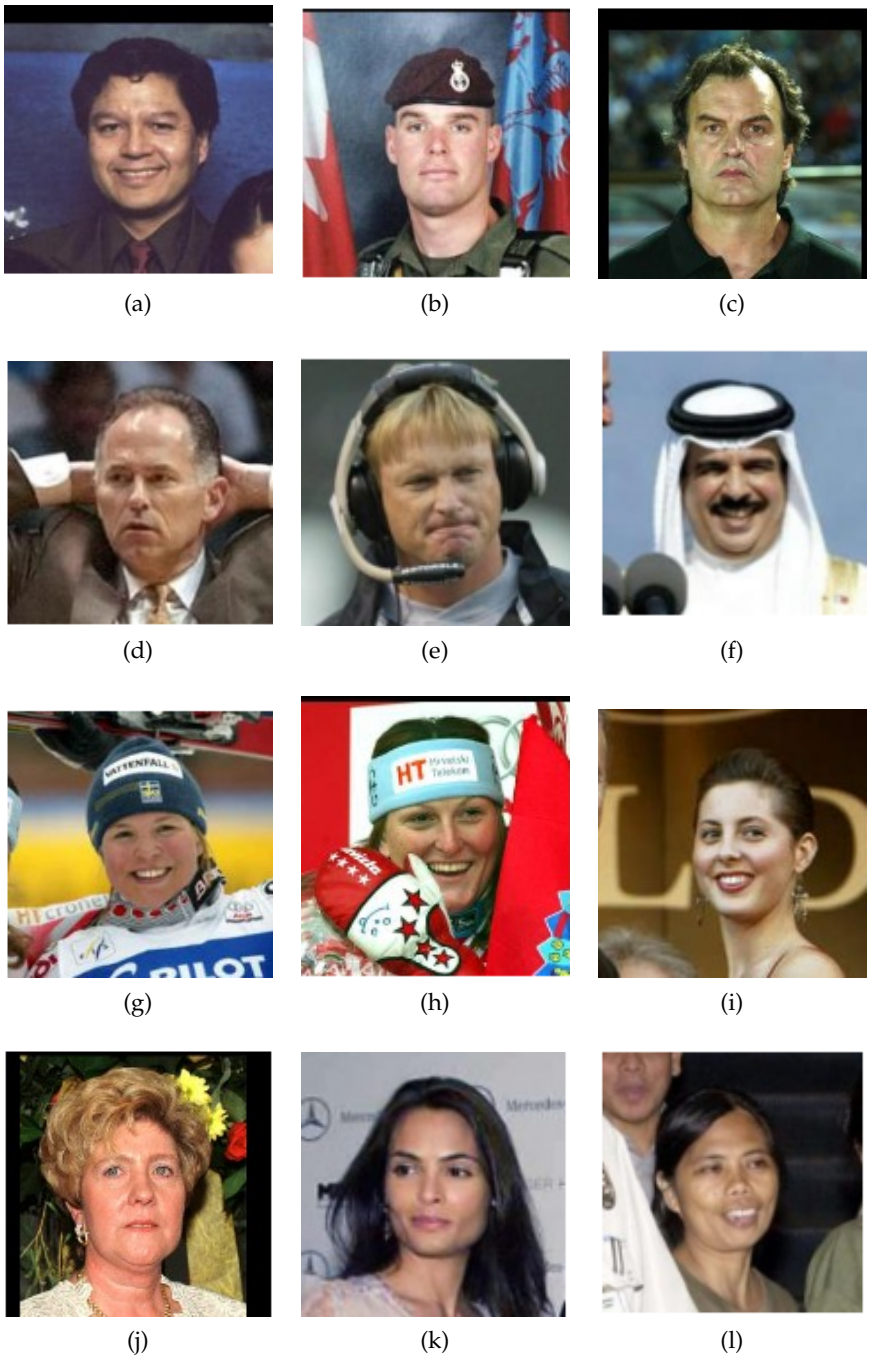


Figure 4.2: Example male faces (a,b,c,d,e,f) and female faces (g,h,i,j,k,l) from the LFW data set.

Since LFW images contains lots of noises and not all of them are suitable to be used in the experiment, therefore we discarded unsuitable ones. Following the recommendation in [5] and [13], 9,763 grayscale images were chosen for the experiment in which 2,293 are females while the rest are males. We also labeled the gender manually as suggested in [1]. All the images were aligned and the faces that are not

clear or the ground truth difficult to establish were discarded. Figure 4.3 shows the example images which are not suitable for the experiment.

Since the number of images between male and female is not balance, we applied 5-fold cross-validation by partitioning images into five subsets of similar size and keeping the same ratio between male and female [46]. Then, the accuracy was computed by taking the average of all folds.

Figure 4.3: Example images which were discarded from the LFW data set.

4.2 PRE-PROCESSING

As mentioned earlier in chapter 3, we applied Viola-Jones algorithm [55] and facial landmark tracking [53] to every image in order detect face and to align the corresponding images. After that, we cropped the face images accordingly. We also resized the cropped images to a fixed size of 128×128 pixels and 224×224 pixels as input for COSFIRE and VGGFace respectively.

Below are listed the steps of pre-processing task using facial landmark tracking and Viola-Jones algorithm:

- First, we detect the face and align all images using facial landmark tracking algorithm.
- Then, we use Viola-Jones algorithm on the aligned images in order to crop the most relevant parts of the face.

4.3 EXPERIMENTS

In this section, we report the evaluation of the proposed methods which were applied on the GF and LFW data sets. First, we report the results of the COSFIRE-based method followed by the results of the CNNs-based method and then we present the results achieved by the fusion of those methods. We also compare the results of the current study with other methods in the end.

4.3.1 *Result with COSFIRE-based Method*

Following the same procedure in [5] and [13], we conducted experiments with COSFIRE-based method on the GF and LFW data sets. However, instead of using 180 filters as suggested in the prior works, we employed 240 COSFIRE filters because these filters works pretty good on our proposed methods. In order to configure COSFIRE filters, the total of 120 COSFIRE filters are randomly selected from male training images and another 120 filters from female training images. For each of the training images, we selected a random region of 19×19 pixels in order to create a prototype to configure a COSFIRE filters. If the pattern consists of at least five features (tuples), then we considered it as a valid prototype. Conversely, if the number of features less than five, we repeated the same procedure over and over again until it meets the condition. As explained in [5], we configured the parameters of the COSFIRE filters with $t_1 = 0.1$, $t_2 = 0.75$, $\sigma_0 = 0.67$, and $\alpha = 0.1$. Moreover, we also configured the responses of the Gabor filters along the concentric circles and center point $\rho = 0,3,6,9$.

The results of the COSFIRE-based method using SVM classifier is shown in table 4.1. It can be seen from the table that the performances of COSFIRE filters applied on the GF and LFW data sets are 93.85 % and 99.19 % respectively. The performance of COSFIRE filters on the GF is 4% below VGGFace CNNs and it was able to level up its performance with VGGFace at 99% when tested on LFW. Moreover, when XGBoost classifier was employed, the COSFIRE filters is still able to maintain its accuracy to not fall over below 98% on the LFW data set, however the accuracy drops to 11% when COSFIRE was validated using LFW (See table 4.2).

4.3.2 *Result with VGGFace CNNs-based Method*

Unlike COSFIRE filters, the VGGFace CNNs-based method does not need to configure several parameters because the configuration has been pretty much set up in the architecture. The only thing that we did was to make sure the extracted features were obtained from the second fully connected layer which can be done by configuring the parameter output of the VGGFace to *fc7*. After that, the features were

Table 4.1: Results of the COSFIRE and VGGFace CNNs-based method on GF and LFW data sets using SVM classifier.

Method	Dataset	Accuracy (%)
COSFIRE filters	GF	93.85
	LFW	99.19
VGGFace CNNs	GF	97.45
	LFW	99.71

Table 4.2: Results of the COSFIRE and VGGFace CNNs-based method on GF and LFW data sets using XGBoost.

Method	Dataset	Accuracy (%)
COSFIRE filters	GF	86.86
	LFW	98.38
VGGFace CNNs	GF	96.18
	LFW	76.22

extracted by the model from each image followed by the normalization of the corresponding features to values between 0 and 1. The normalized features were then trained using SVM ECOC and XGBoost Decision Tree classifiers.

Table 4.1 and 4.2 depict the performances of VGGFace CNNs-based method validated on the GF and LFW data sets using SVM and XGBoost classifiers. The results listed in the tables show the performance of VGGFace is always constant around 96 to 97 % which outperforms the performance of the COSFIRE filters-based method when SVM classifier was employed. However, this approach does not always perform well on the LFW data set particularly when XGBoost classifier was employed. Instead, its performance drops to 76% which is one of the lowest performance during the experiments.

4.3.3 *Result with the fusion of COSFIRE and VGGFace-based methods*

We performed several experiments in order to observe the performances of the proposed methods. First, we implemented the concatenation approach by appending the features of VGGFace CNNs to the features of COSFIRE and subsequently employed SVM ECOC and XGBoost classifiers to perform the gender classification. The results of this experiment show that when the features are merged, the performance increases around 0.1 to 2 % when compared to the best performances achieved by VGGFace CNNs. As we can see in table 4.3, the performance of the model on both data sets using SVM ECOC clas-

sifier is above 98 % and it slightly decreases to not more than 1.1% when XGBoost classifier was employed.

Table 4.3: Results of the fusion of COSFIRE and VGGFace CNNs-based method on the GF and LFW data sets using SVM.

Fusion technique	Classifier	Dataset	Accuracy (%)
Concatenation	SVM ECOC	GF	98.30
	SVM ECOC	LFW	99.28
	XGBoost	GF	97.20
	XGBoost	LFW	98.46
Stacking	SVM ECOC	GF	98.94
	SVM ECOC	LFW	99.38

Moreover, we also conducted another experiment which follows the second approach, the stacking technique. Instead of merging the features, this approach considers the score vectors generated from COSFIRE and CNNs SVM classifiers as an input for a new SVM classifier.

Table 4.3 shows the performance of stacking technique using SVM ECOC classifier. From this table, we can see that the results of this technique outperforms the concatenation one when validated on the GF and LFW data sets. The stacking approach was able to achieve an accuracy of 98.94% on the GF and 99.38% on the LFW.

4.3.4 *Comparison with other methods*

As mentioned earlier that one of the objectives of this study is to prove the effectiveness of the proposed methods by comparing them with the existing studies. In this occasion, we performed comparative analysis with the ones that use the GENDER-FERET and LFW data sets. The comparison of the performance is depicted in table 4.4 and 4.5.

Table 4.4: Comparison of the results on the GF data set.

Method	Description	Accuracy (%)
Azzopardi et al. [8]	RAW LBP HOG	92.6
Azzopardi et al. [5]	COSFIRE	93.7
Azzopardi et al. [13]	COSFIRE SURF	94.7
Proposed 1 (Concatenation Technique)	COSFIRE VGGFACE	98.3
Proposed 2 (Stacking Technique)	COSFIRE VGGFACE	98.9

The results of the performances depicted in table 4.4 shows that both proposed methods outperform the performances of prior studies proposed in [8], [5], and [13] when validated on the GF data set. The stacking technique is able to achieve an accuracy of 98.9 % which is 4.2 % higher than the accuracy of the best previously state-of-the-art listed in the given table. Moreover, the concatenation technique also shows an outstanding performance which is only 0.6% below the stacking one.

Furthermore, we also compared the performances of the proposed methods using the LFW data set. As we can see from table 4.5, the best method was achieved by the stacking approach with an accuracy of 99.38 % followed by the concatenation model which is 0.1% below the stacking one. These results outperform all methods proposed in [52], [20], and [46] and level up the performance of COSFIRE SURF method proposed in [13] at 99.38%.

Table 4.5: Comparison of the results on the LFW dataset.

Method	**Description**	**Accuracy (%)**
J.E Tapia et al. [52]	LBP	92.60
Dago-Casa et al. [20]	Gabor	94.00
Shan et al. [46]	Boosted LBP	94.81
Azzopardi et al. [13]	COSFIRE SURF	99.40
Proposed 1 (Concatenation Technique)	COSFIRE VGGFACE	99.28
Proposed 2 (Stacking Technique)	COSFIRE VGGFACE	99.38

5 DISCUSSION

In this section, we discuss the results of the fusion of trainable features VGGFace and COSFIRE to infer gender from face images. We begin by discussing the effectiveness of the proposed methods (concatenation and stacking approaches) followed by the discussion of how well the performance of both approaches when applied on the constrained (GF) and unconstrained LFW data sets. We finish this chapter with a discussion about the effectiveness of SVM and XGBoost classifiers during the study.

5.1 THE EFFECTIVENESS OF THE PROPOSED METHODS

The results of the study explained in the previous chapter indicate that both proposed methods are demonstrated to be effective in recognizing gender. The concatenation approach is able to achieve its best performance with an accuracy of 98.30% on the GF and 99.20% on the LFW data set. The fusion of COSFIRE and VGGFace using this approach is able to boost the performance around 0.1 to 1 % compared to the best performance of individual approach. These findings show that the fusion of COSFIRE features and VGGFace works well in inferring gender. The rationale of this approach achieving such an outstanding performance is because each individual approach is capable of obtaining the most important features from male and female faces. COSFIRE is able to select for a local contour pattern of the facial images and VGGFace learns the pattern directly from image pixels. Since they are able to extract the most relevant parts of the image and seeing the fact that the performance increases when the features are merged, therefore the fusion of the features is considered to be linear. The linearity of the fusion of these features was validated using SVM and XGBoost classifiers.

Moreover, the results of the stacking method also confirm the effectiveness of the proposed method in classifying gender. It outperforms the previous proposed method, the concatenation technique, by achieving its best performance at 98.90% which were validate on the GF followed by the LFW at 99.38%. Again, with the help of SVM classifier, we extracted score vectors from each CNNs and COSFIRE classifier. The score vectors indicate a number which is assigned to each class male and female based on the given facial image. This number is almost similar to the concept of probability of image being assigned as male and female. This approach is one the best way to prove how well the non linearity of the features in doing classification. By eval-

uating the performances of this approach on both data sets, it shows that the non linear combination of the features from each method is demonstrated to be effective as well as the linear one.

Among these two proposed methods, the stacking approach is demonstrated to perform better than the concatenation one by having 0.1 to 0.6% higher accuracy. However, the extraction of SVM scores using the features from COSFIRE and VGGFace is needed which we consider as a downside of this approach.

5.2 THE PERFORMANCES OF THE PROPOSED METHODS ON THE CONSTRAINED AND UNCONSTRAINED DATA SETS

In this study, two data sets namely Gender Feret (GF) and Labeled Face In the Wild (LFW) were used to confirm the effectiveness of the proposed methods on different data sets. We can see from the results, the proposed methods work very well with the facial images that contain less noises by achieving a remarkable accuracy of 98%. Surprisingly, when it deals with unconstrained one which by nature contains different pose variations, partial occlusion of the face, age variations, different race, and different expression, the performance stays at 99% which is considerably effective in comparison to the constrained one. Both proposed methods were able to achieve its best performance at 99% on the most challenging data set, the LFW.

However, a disadvantage of these methods is it can not perform well on the original data sets. The pre-processing tasks called Face detection and alignment are required before any further processing. Firstly, the face from the images is needed to be captured and then the images are aligned and cropped accordingly. When these methods were validated on the original data sets without applying any pre-processing task, the performance decreases by 2 to 3 % which is below the performances of the proposed methods.

5.3 THE EFFECTIVENESS OF SVM AND XGBOOST CLASSIFIERS

Support Vector Machine (SVM) has been demonstrated to be the most effective classifier in face detection and gender recognition and SVM classifier namely ECOC was employed in this study. From the experiments we conducted, SVM ECOC is demonstrated to be effective on both proposed methods by achieving an accuracy around 98% validated on the GF and LFW data sets. This classifier supports both linear and non linear feature vectors extracted from COSFIRE and VGGFace in classifying gender.

On the other hand, XGBoost classifier also turns out to perform well by achieving 97.2% and 98.46% classification rate on the GF and LFW data sets respectively. However, a downside of this approach is the results might be different on each attempt but not significantly. It

is all because XGBoost is a stochastic algorithm which splits the data set randomly resulting in the slight variety of the performance.

Moreover, when XGBoost was employed on individual approach, there is a big gap in the performance between COSFIRE and VGGFace on each data set. When validated on the GF data set, the performance of COSFIRE decreases significantly up to 10% below the concatenation approach while the performance of VGGFace remains constant. Surprisingly, the phenomenon turns out differently on the LFW data set which is in fact the opposite of the GF. The trainable COSFIRE-based approach is able to achieve an accuracy of 98% on the LFW while VGGFace dips to 76% which is 21% below COSFIRE. That being said, XGBoost classifier is considered to be less reliable than SVM in performing gender classification.

6 CONCLUSION AND FUTURE WORK

In this study we have seen that the fusion of the trainable features from COSFIRE and VGGFace is demonstrated to be highly effective in recognizing gender by boosting the performance up to 1% when compared to the best performance of individual approach. The concatenation technique is able to deal with linear features while the stacking one can handle the non linearity of the features. Both proposed methods are able to exhibit different characteristic of the human faces which comes in handy when performing gender recognition.

The experiments were performed over two public data sets namely Gender FERET (GF) and Labeled Faces in the wild (LFW). The GF data set aims at verifying the method on constrained images while the latter deals with unconstrained ones with different pose variations, age variations, different race, different expression, and partial occlusions with spectacles, wigs, microphones. In general, the results of the proposed methods achieved remarkable performances with an accuracy above 98%. They are able to deal with constrained and unconstrained data sets.

Moreover, we also employed two classifiers: SVM and XGBoost in order to observe the performances of the methods on different classifiers. Based on the results of the experiments that we conducted, both classifiers SVM and XGBoost are demonstrated to perform well in gender classification problem by achieving almost 99% performance rate using SVM and above 97% using XGBoost classifier. However, when both classifiers were employed to train the features from each COSFIRE and VGGFace separately, XGBoost classifier is considered to be unreliable since the performance of COSFIRE drops to 86% on GF data set compared to SVM. Surprisingly, the performance of VGGFace is also getting worse by achieving an accuracy of 77% when validated using LFW data set. That said, the gradient boosting algorithm is considered to be less reliable in performing classification on the extracted features from each individual feature extractor.

6.1 FUTURE WORK

In order to validate the performance more accurately, further experiments need to be performed on more data sets such as Audience benchmark for age and gender classification [25], The Specs on Faces (SoF) [1],Images of Groups dataset [22], SCface (Surveillance Cameras Face Database) [29] and others. Also, a parallel implementation of

1 http://bit.ly/sof$_{d}ataset$

COSFIRE is required so it could run on modern GPUs as suggested in [13].

Part II

APPENDIX

A

DATA SETS

A.1 GENDER FERET

Table A.1: The division of training and test set of the GF data set.

Gender	**Training**	**Test**
Male	237	236
Female	237	236
Total	**474**	**472**

A.2 LABELED FACES IN THE WILD

Table A.2: The division of training and test set of the LFW data set.

Gender	Number of images
Male	7470
Female	2293
Total	**9763**

B

RESULTS

B.1 SVM

Table B.1: The performance of concatenation and stacking approach on GF and LFW data set using SVM classifier.

Dataset	COSFIRE (%)	VGGFACE (%)	Concatenation (%)	Stacking (%)
GF	93.85	97.45	98.30	98.94
LFW	99.19	99.71	99.28	99.38

```
datasource =

    'C:\Users\Newbie\Desktop\FinalMasterThesis\implementasi\datasource\processedrawdata\GENDER-FERET\'

Recognition Rate COSFIRE: 0.938559
Recognition Rate CNN: 0.974576
Recognition Rate Concatenation CNN+ COSFIRE: 0.983051
Recognition Rate Stacked CNN+COSFIRE: 0.989407
Elapsed time is 11.044650 seconds.
```

Figure B.1: The performance of concatenation and stacking approach on the GF data set using .

```
datasource =

    'C:\Users\Newbie\Desktop\FinalMasterThesis\implementasi\datasource\processedrawdata\LFW\cnnsurf\LFW\'

Recognition Rate COSFIRE: 0.991908
Recognition Rate CNN: 0.997132
Recognition Rate CNN+COSFIRE: 0.992830
Recognition Rate CNN+COSFIRE Stacked: 0.993854
Elapsed time is 527.584598 seconds.
```

Figure B.2: The performance of concatenation and stacking approach on the LFW data set using SVM.

B.2 XGBOOST

Table B.2: The performance of concatenation approach on the GF and LFW data set using XGBoost classifier.

Dataset	COSFIRE (%)	VGGFACE (%)	Concatenation Approach (%)
GF	86.86	96.18	97.20
LFW	98.38	76.22	98.46

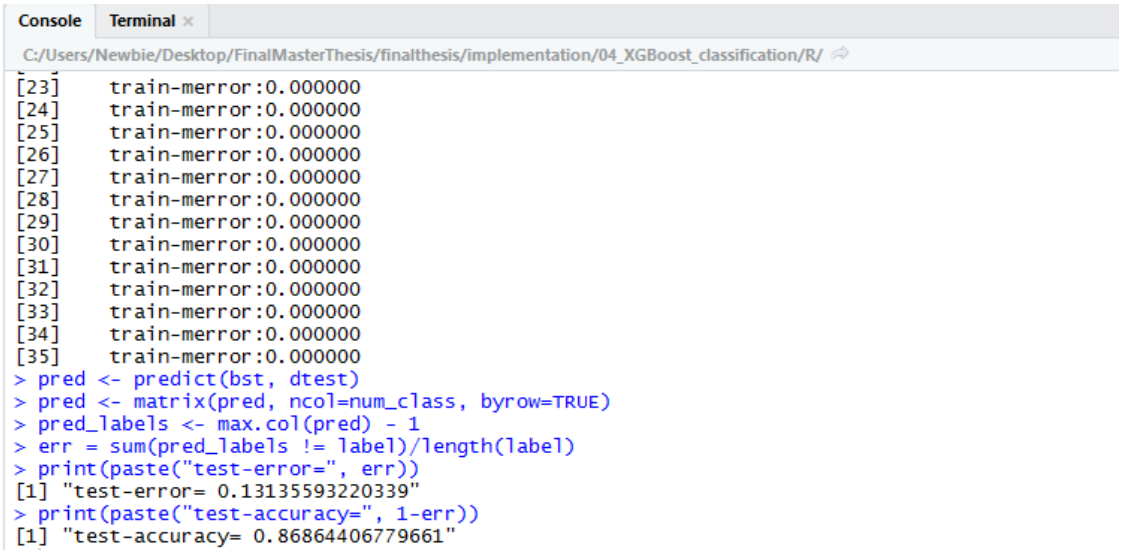


Figure B.3: The performance of COSFIRE on the GF data set using XGBoost.

Figure B.4: The performance of VGGFace on the GF data set using XGBoost.

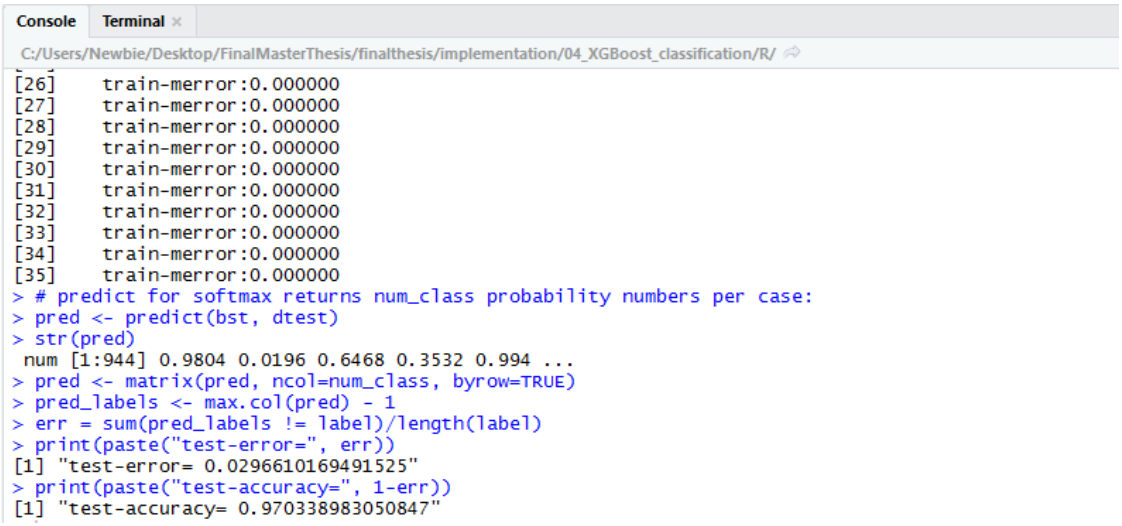


Figure B.5: The performance of concatenation approach on the GF data set using XGBoost.

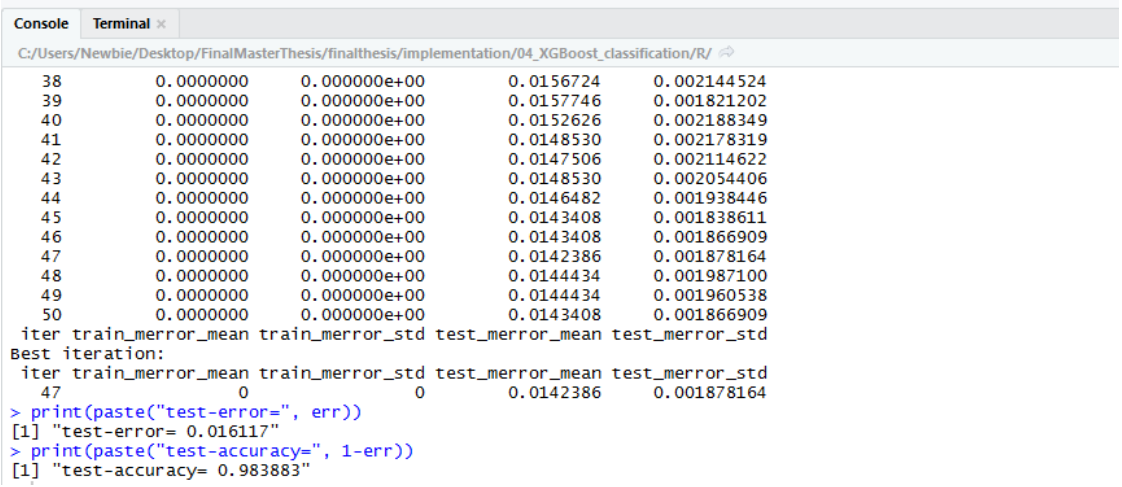


Figure B.6: The performance of COSFIRE on the LFW data set using XGBoost.

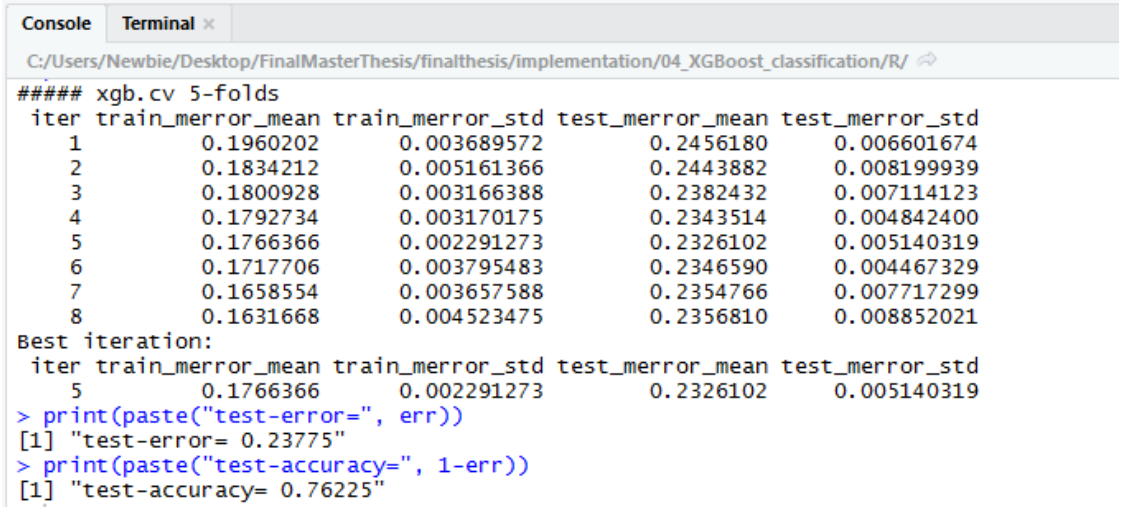


Figure B.7: The performance of VGGFace on the LFW data set using XGBoost.

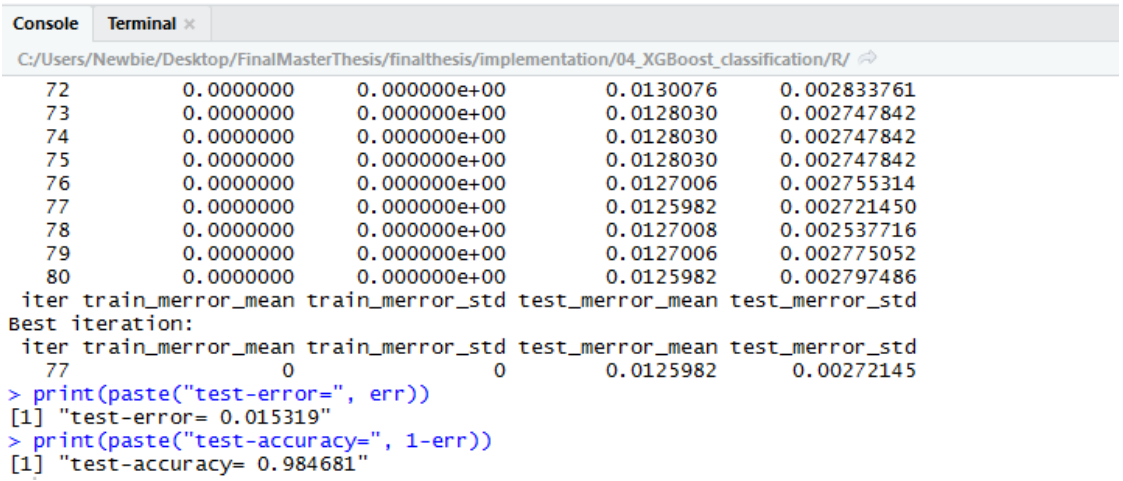


Figure B.8: The performance of concatenation approach on the LFW data set using XGBoost.

BIBLIOGRAPHY

[1] Mahmoud Afifi and Abdelrahman Abdelhamed. "AFIF4: Deep Gender Classification based on AdaBoost-based Fusion of Isolated Facial Features and Foggy Faces." In: *CoRR* abs/1706.04277 (2017).

[2] *An Intuitive Explanation of Convolutional Neural Networks*. https://ujjwalkarn.me/2016/08/11/intuitive-explanation-convnets/. Accessed: 2018-05-17.

[3] Grigory Antipov, Sid-Ahmed Berrani, and Jean-Luc Dugelay. "Minimalistic CNN-based ensemble model for gender prediction from face images." In: *Pattern Recognition Letters, 15 January 2016, Vol.70* (Jan. 2016). DOI: http://dx.doi.org/10.1016/j.patrec.2015.11.011. URL: http://www.eurecom.fr/publication/4768.

[4] *Architecture of Convolutional Neural Networks (CNNs) demystified*. https://www.analyticsvidhya.com/blog/2017/06/architecture-of-convolutional-neural-networks-simplified-demystified/. Accessed: 2018-05-17.

[5] G. Azzopardi, A. Greco, and M. Vento. "Gender recognition from face images with trainable COSFIRE filters." In: *2016 13th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. 2016, pp. 235–241. DOI: 10.1109/AVSS.2016.7738068.

[6] G. Azzopardi and N. Petkov. "Trainable COSFIRE Filters for Keypoint Detection and Pattern Recognition." In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35.2 (2013), pp. 490–503. ISSN: 0162-8828. DOI: 10.1109/TPAMI.2012.106.

[7] G. Azzopardi, A. Greco, A. Saggese, and M. Vento. "Fast gender recognition in videos using a novel descriptor based on the gradient magnitudes of facial landmarks." In: *2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. 2017, pp. 1–6. DOI: 10.1109/AVSS.2017.8078525.

[8] George Azzopardi, Antonio Greco, and Mario Vento. "Gender Recognition from Face Images Using a Fusion of SVM Classifiers." In: *Image Analysis and Recognition*. Ed. by Aurélio Campilho and Fakhri Karray. Cham: Springer International Publishing, 2016, pp. 533–538. ISBN: 978-3-319-41501-7.

[9] George Azzopardi and Nicolai Petkov. "A CORF computational model of a simple cell that relies on LGN input outperforms the Gabor function model." In: *Biological Cybernetics* 106 (2012), pp. 177–189.

[10] George Azzopardi and Nicolai Petkov. "Ventral-stream-like shape representation: from pixel intensity values to trainable object-selective COSFIRE models." In: *Frontiers in Computational Neuroscience* 8 (2014), p. 80. ISSN: 1662-5188. DOI: 10.3389/fncom.2014.00080. URL: https://www.frontiersin.org/article/10.3389/fncom.2014.00080.

[11] George Azzopardi, Antonio Jose Rodríguez-Sánchez, Justus H. Piater, and Nicolai Petkov. "A Push-Pull CORF Model of a Simple Cell with Antiphase Inhibition Improves SNR and Contour Detection." In: *PloS one*. 2014.

[12] George Azzopardi, Nicola Strisciuglio, Mario Vento, and Nicolai Petkov. "Trainable COSFIRE filters for vessel delineation with application to retinal images." In: *Medical Image Analysis* 19.1 (2015), pp. 46–57.

[13] George Azzopardi, Antonio Greco, Alessia Saggese, and Mario Vento. "Fusion of domain-specific and trainable features for gender recognition from face images." English. In: *IEEE Access* (Apr. 2018). DOI: 10.1109/ACCESS.2018.2823378.

[14] Frederic Bastien, Pascal Lamblin, Razvan Pascanu, James Bergstra, Ian Goodfellow, Arnaud Bergeron, Nicolas Bouchard, David Warde-Farley, and Y Bengio. "Theano: new features and speed improvements." In: (Nov. 2012).

[15] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool. "Speeded-Up Robust Features (SURF)." In: *Comput. Vis. Image Underst.* 110.3 (June 2008), pp. 346–359. ISSN: 1077-3142. DOI: 10.1016/j.cviu.2007.09.014. URL: http://dx.doi.org/10.1016/j.cviu.2007.09.014.

[16] A Mike Burton, Vicki Bruce, and Neal Dench. "What's the Difference between Men and Women? Evidence from Facial Measurement." In: *Perception* 22.2 (1993). PMID: 8474841, pp. 153–176. DOI: 10.1068/p220153. eprint: https://doi.org/10.1068/p220153. URL: https://doi.org/10.1068/p220153.

[17] Tianqi Chen and Carlos Guestrin. "XGBoost: A Scalable Tree Boosting System." In: *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '16. San Francisco, California, USA: ACM, 2016, pp. 785–794. ISBN: 978-1-4503-4232-2. DOI: 10.1145/2939672.2939785. URL: http://doi.acm.org/10.1145/2939672.2939785.

[18] *Coding Neural Networks - Forward Propagation and Back Propagation*. https://towardsdatascience.com/coding-neural-network-forward-propagation-and-backpropagtion-ccf8cf369f76. Accessed: 2018-05-17.

[19] Garrison W. Cottrell and Janet Metcalfe. "EMPATH: Face, Emotion, and Gender Recognition Using Holons." In: *Proceedings of the 1990 Conference on Advances in Neural Information Processing Systems 3*. NIPS-3. Denver, Colorado, USA: Morgan Kaufmann Publishers Inc., 1990, pp. 564–571. ISBN: 1-55860-184-8. URL: http://dl.acm.org/citation.cfm?id=118850.105194.

[20] P. Dago-Casas, D. González-Jiménez, Long Long Yu, and J. L. Alba-Castro. "Single- and cross- database benchmarks for gender classification under unconstrained settings." In: *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*. 2011, pp. 2152–2159. DOI: 10.1109/ICCVW.2011.6130514.

[21] Eran Eidinger, Roee Enbar, and Tal Hassner. "Age and Gender Estimation of Unfiltered Faces." In: *IEEE Trans. Information Forensics and Security* 9.12 (2014), pp. 2170–2179. URL: http://dblp.uni-trier.de/db/journals/tifs/tifs9.html#EidingerEH14.

[22] Eran Eidinger, Roee Enbar, and Tal Hassner. "Age and Gender Estimation of Unfiltered Faces." In: *Trans. Info. For. Sec.* 9.12 (Dec. 2014), pp. 2170–2179. ISSN: 1556-6013. DOI: 10.1109/TIFS.2014.2359646. URL: https://doi.org/10.1109/TIFS.2014.2359646.

[23] Yoav Freund and Robert E. Schapire. "A Short Introduction to Boosting." In: *In Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence*. Morgan Kaufmann, 1999, pp. 1401–1406.

[24] Kunihiko Fukushima. "Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position." In: *Biological Cybernetics* 36.4 (1980), pp. 193–202. ISSN: 1432-0770. DOI: 10.1007/BF00344251. URL: https://doi.org/10.1007/BF00344251.

[25] A. Gallagher and T. Chen. "Understanding Images of Groups Of People." In: *Proc. CVPR*. 2009.

[26] Baris Gecer, George Azzopardi, and Nicolai Petkov. "Color-blob-based COSFIRE filters for object recognition." In: *Image and Vision Computing* (2016).

[27] *Gender Recognition Dataset*. http://mivia.unisa.it/datasets/video-analysis-datasets/gender-recognition-dataset/. Accessed: 2018-05-28.

[28] B. A. Golomb, D. T. Lawrence, and T. J. Sejnowski. "SexNet: A Neural Network Identifies Sex from Human Faces." In: *Proceedings of the 1990 Conference on Advances in Neural Information Processing Systems 3*. NIPS-3. Denver, Colorado, USA: Morgan

Kaufmann Publishers Inc., 1990, pp. 572–577. ISBN: 1-55860-184-8. URL: http://dl.acm.org/citation.cfm?id=118850.118953.

[29] Mislav Grgic, Kresimir Delac, and Sonja Grgic. "SCface — Surveillance Cameras Face Database." In: *Multimedia Tools Appl.* 51.3 (Feb. 2011), pp. 863–879. ISSN: 1380-7501. DOI: 10.1007/s11042-009-0417-2. URL: http://dx.doi.org/10.1007/s11042-009-0417-2.

[30] Abdenour Hadid, Juha Ylioinas, Messaoud Bengherabi, Mohammad Ghahramani, and Abdelmalik Taleb-Ahmed. "Gender and Texture Classification." In: *Pattern Recogn. Lett.* 68.P2 (Dec. 2015), pp. 231–238. ISSN: 0167-8655. DOI: 10.1016/j.patrec.2015.04.017. URL: http://dx.doi.org/10.1016/j.patrec.2015.04.017.

[31] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. *Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments*. Tech. rep. 07-49. University of Massachusetts, Amherst, 2007.

[32] A. Jain and J. Huang. "Integrating independent components and linear discriminant analysis for gender classification." In: *Sixth IEEE International Conference on Automatic Face and Gesture Recognition, 2004. Proceedings.* 2004, pp. 159–163. DOI: 10.1109/AFGR.2004.1301524.

[33] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "ImageNet Classification with Deep Convolutional Neural Networks." In: *Advances in Neural Information Processing Systems 25*. Ed. by F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger. Curran Associates, Inc., 2012, pp. 1097–1105. URL: http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf.

[34] Yann LeCun and Yoshua Bengio. "The Handbook of Brain Theory and Neural Networks." In: ed. by Michael A. Arbib. Cambridge, MA, USA: MIT Press, 1998. Chap. Convolutional Networks for Images, Speech, and Time Series, pp. 255–258. ISBN: 0-262-51102-9. URL: http://dl.acm.org/citation.cfm?id=303568.303704.

[35] Chen-Yu Lee, Patrick W. Gallagher, and Zhuowen Tu. "Generalizing Pooling Functions in Convolutional Neural Networks: Mixed, Gated, and Tree." In: *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*. Ed. by Arthur Gretton and Christian C. Robert. Vol. 51. Proceedings of Machine Learning Research. Cadiz, Spain: PMLR, 2016, pp. 464–472. URL: http://proceedings.mlr.press/v51/lee16a.html.

[36] Gil Levi and Tal Hassner. "Age and gender classification using convolutional neural networks." In: *CVPR Workshops*. IEEE Computer Society, 2015, pp. 34–42. ISBN: 978-1-4673-6759-2. URL: http://dblp.uni-trier.de/db/conf/cvpr/cvprw2015.html#LeviH15.

[37] Xiao-Chen Lian and Bao-Liang Lu. "Gender Classification by Combining Facial and Hair Information." In: *Advances in Neuro-Information Processing*. Ed. by Mario Köppen, Nikola Kasabov, and George Coghill. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 647–654. ISBN: 978-3-642-03040-6.

[38] R. Lippmann. "An introduction to computing with neural nets." In: *IEEE ASSP Magazine* 4.2 (1987), pp. 4–22. ISSN: 0740-7467. DOI: 10.1109/MASSP.1987.1165576.

[39] Zachary Chase Lipton. "A Critical Review of Recurrent Neural Networks for Sequence Learning." In: *CoRR* abs/1506.00019 (2015). arXiv: 1506.00019. URL: http://arxiv.org/abs/1506.00019.

[40] Jordi Mansanet, Alberto Albiol, and Roberto Paredes. "Local Deep Neural Networks for Gender Recognition." In: *Pattern Recogn. Lett.* 70.C (Jan. 2016), pp. 80–86. ISSN: 0167-8655. DOI: 10.1016/j.patrec.2015.11.015. URL: http://dx.doi.org/10.1016/j.patrec.2015.11.015.

[41] Hossein Moeini and Saeed Mozafari. "Gender Dictionary Learning for Gender Classification." In: 42 (Nov. 2016).

[42] Saeed Mozaffari, Hamid Behravan, and Rohollah Akbari. "Gender Classification Using Single Frontal Image Per Person: Combination of Appearance and Geometric Based Features." In: *20th International Conference on Pattern Recognition, ICPR 2010, Istanbul, Turkey, 23-26 August 2010*. 2010, pp. 1192–1195. DOI: 10.1109/ICPR.2010.297. URL: https://doi.org/10.1109/ICPR.2010.297.

[43] O. M. Parkhi, A. Vedaldi, and A. Zisserman. "Deep Face Recognition." In: *British Machine Vision Conference*. 2015.

[44] P.Jonathon Phillips, Harry Wechsler, Jeffery Huang, and Patrick J. Rauss. "The FERET database and evaluation procedure for face-recognition algorithms." In: *Image and Vision Computing* 16.5 (1998), pp. 295 –306. ISSN: 0262-8856. DOI: https://doi.org/10.1016/S0262-8856(97)00070-X. URL: http://www.sciencedirect.com/science/article/pii/S026288569700070X.

[45] Brunelli Poggio, R. Brunelli, and T. Poggio. *HyberBF Networks for Gender Classification*.

[46] Caifeng Shan. "Learning local binary patterns for gender classification on real-world face images." In: *Pattern Recognition Letters* 33.4 (2012). Intelligent Multimedia Interactivity, pp. 431 – 437. ISSN: 0167-8655. DOI: https://doi.org/10.1016/j.patrec.2011.05.016. URL: http://www.sciencedirect.com/science/article/pii/S0167865511001607.

[47] Karen Simonyan and Andrew Zisserman. "Very Deep Convolutional Networks for Large-Scale Image Recognition." In: *CoRR* abs/1409.1556 (2014). arXiv: 1409.1556. URL: http://arxiv.org/abs/1409.1556.

[48] *Softmax function*. https://en.wikipedia.org/wiki/Softmax_function. Accessed: 2018-05-18.

[49] Nicola Strisciuglio, George Azzopardi, Mario Vento, and Nicolai Petkov. "Supervised vessel delineation in retinal fundus images with the automatic selection of B-COSFIRE filters." In: *Mach. Vis. Appl.* 27.8 (2016), pp. 1137–1149.

[50] Zehang Sun, G. Bebis, Xiaojing Yuan, and S. J. Louis. "Genetic feature subset selection for gender classification: a comparison study." In: *Sixth IEEE Workshop on Applications of Computer Vision, 2002. (WACV 2002). Proceedings.* 2002, pp. 165–170. DOI: 10.1109/ACV.2002.1182176.

[51] Shinichi Tamura, Hideo Kawai, and Hiroshi Mitsumoto. "Male/female identification from 8 × 6 very low resolution face images by neural network." In: 29 (Feb. 1996), pp. 331–335.

[52] J. E. Tapia and C. A. Perez. "Gender Classification Based on Fusion of Different Spatial Scale Features Selected by Mutual Information From Histogram of LBP, Intensity, and Shape." In: *IEEE Transactions on Information Forensics and Security* 8.3 (2013), pp. 488–499. ISSN: 1556-6013. DOI: 10.1109/TIFS.2013.2242063.

[53] M. Uricar, V. Franc, and V. Hlavac. "Facial Landmark Tracking by Tree-Based Deformable Part Model Based Detector." In: *2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*. Vol. 00. 2016, pp. 963–970. DOI: 10.1109/ICCVW.2015.127. URL: doi.ieeecomputersociety.org/10.1109/ICCVW.2015.127.

[54] R. Venkatesan and B. Li. *Convolutional Neural Networks in Visual Computing: A Concise Guide*. Data-Enabled Engineering. Taylor & Francis Group, 2017. ISBN: 9781138747951. URL: https://books.google.nl/books?id=Y2xSAQAACAAJ.

[55] Paul Viola and Michael J. Jones. "Robust Real-Time Face Detection." In: *International Journal of Computer Vision* 57.2 (2004), pp. 137–154. ISSN: 1573-1405. DOI: 10.1023/B:VISI.0000013087.49260.fb. URL: https://doi.org/10.1023/B:VISI.0000013087.49260.fb.