ISTANBUL TECHNICAL UNIVERSITY ★ GRADUATE SCHOOL OF SCIENCE
ENGINEERING AND TECHNOLOGY

UNCONSTRAINED FACE RECOGNITION UNDER MISMATCHED CONDITIONS

M.Sc. THESIS

Omid Abdollahi Aghdam

Department of Computer Engineering

Computer Engineering Programme

SEPTEMBER 2018

ISTANBUL TECHNICAL UNIVERSITY ★ GRADUATE SCHOOL OF SCIENCE
ENGINEERING AND TECHNOLOGY

UNCONSTRAINED FACE RECOGNITION UNDER MISMATCHED CONDITIONS

M.Sc. THESIS

Omid Abdollahi Aghdam
(504151520)

Department of Computer Engineering

Computer Engineering Programme

Thesis Advisor: Assoc. Prof. Dr. Hazım Kemal EKENEL

SEPTEMBER 2018

İSTANBUL TEKNİK ÜNİVERSİTESİ ★ FEN BİLİMLERİ ENSTİTÜSÜ

EŞLEŞMEYEN KOŞULLAR ALTINDA YÜZ TANIMA

YÜKSEK LİSANS TEZİ

Omid Abdollahi Aghdam
(504151520)

Bilgisayar Mühendisliği Anabilim Dalı

Bilgisayar Mühendisliği Programı

Tez Danışmanı: Assoc. Prof. Dr. Hazım Kemal EKENEL

EYLÜL 2018

Omid Abdollahi Aghdam, a M.Sc. student of ITU Graduate School of Science Engineering and Technology 504151520 successfully defended the thesis entitled “UNCONSTRAINED FACE RECOGNITION UNDER MISMATCHED CONDITIONS”, which he/she prepared after fulfilling the requirements specified in the associated legislations, before the jury whose signatures are below.

Thesis Advisor :	**Assoc. Prof. Dr. Hazım Kemal EKENEL** Istanbul Technical University	....
Jury Members :	**Prof. Dr. Muhittin Gökmen** MEF University	....
	Prof. Dr. Mustafa Ersel Kamaşak Istanbul Technical University	....
		....

Date of Submission : **5 September 2018**
Date of Defense : **1 October 2018**

To my family,

FOREWORD

I do appreciate all the supervision I received from my advisor, Assoc. Prof. Dr. Hazım Kemal EKENEL, during the process of my Master's degree in general and thesis in particular.

I also do appreciate all the support I received from my family during my academic career.

Finally, I should also add my appreciation of the support I received from the members of Smart Interaction, Mobile Intelligence, and Multimedia Technologies Lab.

September 2018

Omid Abdollahi Aghdam
Computer Engineer

TABLE OF CONTENTS

Page

FOREWORD .... **ix**
TABLE OF CONTENTS .... **xi**
ABBREVIATIONS .... **xiii**
LIST OF TABLES .... **xv**
LIST OF FIGURES .... **xvii**
SUMMARY .... **xix**
ÖZET .... **xxi**
1. INTRODUCTION **1**
  1.1 Purpose of Thesis 4
  1.2 Literature Review 4
  1.3 Hypothesis 6
2. CONVOLUTIOAL NEURAL NETWORKS **9**
  2.1 Components 9
    2.1.1 Convolution layer 9
    2.1.2 Activation functions 10
    2.1.3 Pooling layer 11
    2.1.4 Batch normalization 12
    2.1.5 Dropout 12
    2.1.6 Fully connected layer 13
    2.1.7 Classification layer 13
  2.2 Loss Functions 13
    2.2.1 Cross entropy loss 14
    2.2.2 Triplet loss 14
    2.2.3 Additive angular margin 15
3. TRANSFER LEARNING **17**
  3.1 Feature Extraction 17
  3.2 Fine Tuning 17
  3.3 Deep Learning Architectures 18
    3.3.1 Residual neural network 18
    3.3.2 Squeeze-and-excitation networks 20
      3.3.2.1 Squeeze: global information embedding 20
      3.3.2.2 Excitation: adaptive re-calibration 20
    3.3.3 Improved residual neural networks 21
4. METHODOLOGY **25**
  4.1 Datasets 25
    4.1.1 ICB-RW 26
    4.1.2 SCFace 26

4.2 Proposed Method 27
4.2.1 Models 28
4.2.1.1 Fine-tuning the models. 29
4.2.2 Face detection and alignment 30
4.2.3 Feature extraction 30
4.2.4 Increasing the amount of information in probe face images 31
4.2.5 Resolution matching between the gallery and probe face images 32
4.2.6 Face identification 33
5. EXPERIMENTAL RESULTS 35
5.1 Experimental Results on ICB-RW 35
5.1.1 Extended bounding boxes and down-sampled gallery face images 37
5.2 Experimental Results on SCFace 38
5.2.1 MTCNN bounding boxes 38
5.2.2 Extended bounding boxes 39
5.2.3 Down-sampled gallery face images 41
6. CONCLUSIONS AND RECOMMENDATIONS 43
REFERENCES 45
APPENDICES 49
APPENDIX A.1 51
CURRICULUM VITAE 53

ABBREVIATIONS

ANN : Artificial Neural Network
AUC : Area Under the Curve
CNN : Convolutional Neural Netrowk
CMC : Cumulative Match Curve
DCNN : Deep Convolutional Neural Netrowk
FR : Face Recognition
HR : High Resolution
HRFR : High Resolution Face Recognition
ICB-RW : International Challenge on Biometric Recognition in the Wild
IR : Identification Rate
LFW : Labled Faces in the Wild
LR : Low Resolution
LRFR : Low Resolution Face Recognition
ResNet : Residual Networks
SCFace : Surveillance Cameras Face database
SENet : Squeeze-and-Excitation Networks
YTF : YouTube Faces

LIST OF TABLES

Page

Table 4.1 : The scale of databases that are used in the experiments. 28
Table 5.1 : Identification rates and CMC with ground-truth bounding boxes....... 37
Table 5.2 : Rank-1, Rank-5 IR and CMC on ICB-RW benchmark [1] using MTCNN bounding boxes. The extension (EXT) and down-sampling (DOWN) factors are 1.2 and 64 × 64, respectively. . 38
Table 5.3 : Rank-1 IR mean and std. for the proposed models (20 runs). MTCNN bounding boxes are used in these experiments................... 39
Table 5.4 : Experimental results on SCFace with and without enlarged bounding boxes. Here, the extension (EXT) factor is 1.35................ 40
Table 5.5 : Experimental results for SCFace with MTCNN bounding boxes, Extended bounding boxes (EXT), and Down-sampled (DOWN) gallery faces. Here, the extension factor is 1.35 and down-sampling factor is 40 × 40... 42
Table A.1 : Experimental results on SCFace [2] with MTCNN bounding boxes, Extended bounding boxes (EXT) with extension factor of 1.35, and Down-sampling (DOWN) factor of 40 × 40. 51

LIST OF FIGURES

Page

Figure 1.1 : Face identification protocol used in surveillance scenarios.............. 2
Figure 2.1 : A 2D convolution operation [3]... 10
Figure 2.2 : A 2×2 pooling operation.. 11
Figure 2.3 : Batch normalization algorithm [4]. 12
Figure 2.4 : A neural network before and after Dropout [5]................................ 13
Figure 2.5 : Triplet loss illustration. [6]. 14
Figure 3.1 : A Residual Block [6]. 18
Figure 3.2 : The number of the parameters of the Residual Neural Networks with different number of layers [7]. $[kernelsize, numberofkernels] \times numberofblocks$ in the architecture. 19
Figure 3.3 : A Squeeze and Excitation block [8]. 21
Figure 3.4 : **Left**: A residual block. **Right**: A SE block added to a residual block, SE-ResNet [8]. 21
Figure 3.5 : SE-ResNet-50 layers parameter settings [8].................................... 22
Figure 3.6 : **Left**: Improved residual block [9]. **Right**: Original residual block [7]... 23
Figure 4.1 : Samples from ICB-RW contains variation in pose, illumination, occlusion, and etc., whereas, SCFace has samples with very low resolution... 25
Figure 4.2 : Samples of two subjects in ICB-RW 2016 database. 26
Figure 4.3 : Sample images of one subject in SCFace [2]. Here, the frontal face images in the gallery (taken with IR night vision and normal vision) and the probe face images captured using five surveillance camera are shown cam1, ..., cam5. Two of the surveillance cameras have IR night vision which are used to capture face images in cam6 and cam7. The number in the end of the probe faces' name represent distance 1, 2, and 3.. 27
Figure 4.4 : Pipeline of the proposed face identification. 28
Figure 4.5 : Detection and alignment step of a subject from ICB-RW 2016 database... 31
Figure 4.6 : Features of each face is extracted using pre-trained models and L2 normalized before inputting to the classifier. 32
Figure 4.7 : Gallery and probe faces of a subject from SCFace and ICB-RW benchmarks. The faces are cropped with bounding boxes detected by MTCNN [10] and different extension factors............................... 32

Figure 4.8 : Gallery faces of two subjects in SCFace cropped with 1.35 extended bounding boxes. **Column 1**: the original resolution. **Columns 2-5**: down-sampled images. **Column 6**: distance 1 probe faces. 33
Figure 5.1 : Pipeline of the ensemble model for face identification. 35
Figure 5.2 : Area Under the Curve of Cumulative Match Curve for Face Identification Results of ICB-RW database [1]. 36
Figure 5.3 : Thirty-seven probe faces from ICB-RW database that are miss-identified by proposed ensemble model. 37
Figure 5.4 : Three probe faces from ICB-RW database that are not identified correctly by any proposed model. 37

UNCONSTRAINED FACE RECOGNITION UNDER MISMATCHED CONDITIONS

SUMMARY

Surveillance cameras are very commonplace, and thus can be utilized to make the world a better, safer place. There are many applications which use massive surveillance data to extract information. The speed cameras are mounted on the roads to detect violations in the traffic rules, as well as the security cameras, which are ubiquitous in buildings to detect anomalies. In addition, there is a great interest to analyze and estimate the people's activities in social events. Extensive usage of surveillance cameras has made the monitoring task laborious, such that it is almost impossible for a person to monitor videos efficiently and act accordingly. Therefore, automatic surveillance techniques have been proposed.

Face recognition, to a great degree, has been addressed in the literature of computer vision. Face recognition is the problem of matching 1:1 face images (one-to-one), face verification, or 1:N face images (one-to-many), face identification. It has use cases in law enforcement, watchlist, security gates and etc. The usual approach in face recognition is to find the matches in the gallery faces with the probe faces by comparing the distance of the face embedding using a similarity measurement. Conventionally, the extracted features that were used for the comparison were hand engineered, however, due to the rapid progress in the deep learning field, e.g. abundance of the large-scale face database and GPUs, nowadays, deep learning based feature learning methods are preferred.

Although, face recognition is a challenging task, as a result of variation in pose, illumination, expression, and occlusion, it is considered to be solved under matched conditions. In matched conditions, face images are literally coming from the same source with relatively high resolution, e.g. faces collected from the Internet. In the mismatched conditions, the face images are coming from the different sources, e.g. surveillance scenarios, where we have the high resolution face images for train set, and the low resolution in the test set.

In this thesis, we focused on the face recognition under mismatched conditions and leveraged deep learning models to learn and extract deep face representations. Afterwards, deep face representations are used to compare the faces using a similarity measurement. For this purpose, we used correlation distance between learned features and nearest neighbor classifier to report Rank-1 Identification Rate.

In order to extract the features, we have employed 50 and 100 layers residual neural network models trained on VGGFace2 and MS-Celeb-1M databases. We extracted face embedding from the last layer of these networks for each faces in the gallery and the probe set. Furthermore, we experimented with the different amount of information included in the face crops in which we extended the detected bounding boxes which resulted in significant performance boost. Additionally, down-sampling the gallery faces before feature extraction increased the Rank-1 identification rate. For evaluating the performance of the deep learning models at learning discriminative features, we examined the proposed method on ICB-RW and SCFace databases.

Our models are trained on VGGFace2 database which is composed of 3.31 million faces of 9131 subjects collected from the celebrity images in different poses and ages and MS-Celeb-1M database which has 10 million images of 100,000 subjects, collected from the celebrity images from the web. The experimental results demonstrate the advantage of using large-scale face database to train deep Convolutional Neural Networks in learning robust face embedding. The experimental results show that quality and variation of the training database is more important than quantity of the database to learn general feature representations. In other words, although, MS-Celeb-1M has 10 times more identity and 3 times more images than VGGFace2, features learned by models trained on VGGFace2 database have better generalization on faces with very low resolution which is a challenging problem in the databases coming from surveillance cameras. Our results on ICB-RW database significantly surpassed the results of previous works, and the experiments on SCFace database achieved state-of-the-art results for distance 3, distance 2, and distance 1 subsets of the probe set. The state-of-the-art results on SCFace benchmark are achieved with an improved version of ResNet-100 trained on MS-Celeb-1M and fine-tuned on VGGFae2 database, which are 76.94% ±1.98, 98.41% ±0.92, and 100% ±0.00 for distance 1, 2, 3, respectively. An ensemble of four model achieved 91.78%, 98.00%, 0.997, Rank-1, Rank-5 IR and area under the curve of cumulative match curve respectively.

EŞLEŞMEYEN KOŞULLAR ALTINDA YÜZ TANIMA

ÖZET

Yaygın olarak kullanılan gözetim kameraları dünyayı daha iyi ve güvenli bir yer haline getirmek için kullanılabilmektedir. Gözetim kamera verileri birçok uygulama tarafından kullanılmaktadır; hız kontrol kameraları trafik kurallarındaki ihlalleri tespit etmek için yollara monte edilmişken, binalarda her yerde bulunan güvenlik kameraları ise anormallikleri tespit etmek için kullanılmaktadır. Ayrıca, insanların sosyal etkinliklerde faaliyetlerini analiz ve tahmin etmek için araştırmacılarda büyük bir ilgi uyandırmaktadır. Gözetim kameralarının yaygın olarak kullanımı, kişilerin kaydedilen veya anlık görüntüleri izleme görevini çok zahmetli hale getirmiştir. Yetkililerin videoları verimli bir şekilde izlemeleri ve buna göre hareket etmeleri neredeyse imkansızdır. Bu nedenle, bu tezde otomatik gözetim teknikleri önerilmiştir.

Yüz tanıma, bilgisayarla görü alanında büyük ölçüde literatürde yer alan konulardan birisidir. Yüz imgelerinin birebir eşlenmesi (1:1) yüz doğrulama (face verification), bir yüz imgesi ile birden çok (1:N) yüz imgesinin karşılaştırılması ise yüz kimlik tanıması (face identification) olarak tanımlanmaktadır. Hukuk alanında, güvenlik soruşturmalarında, gözetleme ve güvenlik kapılarında yüz tanıma kullanılmaktadır. Yüz tanımadaki genel yaklaşım, yüz imgelerinden öznitelikler çıkarmak ve bu öznitelikleri benzerlik ölçümü ile karşılaştırmaktır. Geleneksel olarak elle öznitelikler çıkarılırken, derin öğrenme algoritmalarındaki son gelişmeler ve büyük ölçekli yüz veri kümeleri sayesinde derin öğrenme temelli öznitelik öğrenme yöntemleri tercih edilmektedir.

Poz, aydınlanma, yüz ifadesi ve yüzün başka bir nesne ile kapanması gibi nedenlerlen dolayı yüz tanıma zor bir problem olmasına rağmen, imgeler eşleşen koşullarda toplandığı durumlarda problemin çözüldüğü düşünülmektedir. Eşleşen koşullarda, yüz görüntüleri genel olarak aynı alandan gelmektedir ve göreceli olarak yüksek çözünürlükte olmaktadır; eşleşmeyen koşullarda ise yüz görüntüleri farklı alanlardan gelmektedir. Örneğin, gözetleme senaryolarında, galeri kümesinde yüksek çözünürlüklü yüz görüntüleri varken, prob kümesinde düşük çözünürlüğe sahip yüz imgeleri bulunmaktadır.

Bu tezde, eşleşmeyen koşullar altında yüz tanımaya ve yüz özniteliklerini öğrenmek ve çıkarmak için derin öğrenme modellerinden yararlanmaya odaklanılmıştır. Yüzler arasındaki benzerlik ölçümü derin yüz öznitelikleri kullanılarak yapılmaktadır. Bu amaçla, Rank-1, Rank-5 yüz tanıma doğruluğunu ve Kümülatif Eşleşme Skoru eğrisinin altındaki alanı rapor etmek için öğrenilen özniteliklerin arasındaki uzaklık korelasyon mesafesi ile en yakın komşu sınıflandırıcısı kullanılmıştır.

Öznitellikleri elde etmek için, VGGFace2 veri kümesi üzerinde önceden eğitilmiş 50 katmanlı SENet ve ResNet modelleri, MS-Celeb-1M veri kümesi üzerinde eğitilmiş 50 katmanlı SENet ve ResNet modelleri ve sonradan VGGFace2 üzerinde ince ayar

yapılmış modelleri kullanılmıştır. Tanımlanan dört modelden çıkardığımız öznitelik vektörleri 2048 boyutludur. İlaveten, MS-Celeb-1M veri kümesi üzerinde eğitilmiş 50 katmanlı ve 100 katmanlı geliştirilmiş ResNet modelleri ve aynı modellerin VGGFace2 üzerinde ince ayar yapılmış versiyonları kullanılmıştır. Geliştirilmiş ResNet modellerinden elde edilen öznitelikler 512 boyutludur. Anlatılan 8 derin öğrenme modelleri kullanılarak galeri ve prob kümesinden her yüz için öznitelik vektörleri çıkarılmıştır. Derin öğrenme modelleri kullanılarak öğrenilen özniteliklerin ayırt edici özelliklerini karşılaştırmak amacıyla, derin yüz özniteliklerinin yüz tanımadaki performans ölçümü ICB-RW ve SCFace veri kümeleri üzerinde değerlendirilmiştir.

Yüz özniteliklerini öğrenmek için kullandığımız derin öğrenme modellerinin yüz tanımadaki başarımını ölçmek için sırasıyla her bir model ile üç deney yapılmıştır. Birinci deneyde, MTCNN modelini kullanarak, veri kümelerindeki imgelerde yüz tespiti yapılmış ve modelin verdiği yüz tespit çerçevesi koordinatları kullanılarak kişilerin yüz bölgeleri kesilmiş ve bu kesilmiş yüz imgeleri daha sonra öznitelik çıkartmak için modellere girdi olarak verilmiştir. İkinci deneyde, MTCNN modelinin bulduğu çerçevelerden daha geniş çerçeveler kullanılarak yüzler kesilmiştir ve yüz imgeleri modellere girdi olarak verilmiştir. Kullanılan geniş çerçevelerin referans çerçeveye göre ölçek faktörleri 1.1, 1.2, 1.25, 1.30, 1.35, 1.40 şeklinde belirlenmiştir ve en yüksek başarılar SCFace için referans çerçevenin 1.35, ve ICB-RW için referans çerçevenin 1.2 ölçekte olduğu durumlarda elde edilmiştir. Üçüncü deneyde, yüz çerçevesi olarak her veri kümesi için en yüksek başarımı veren çerçeveler seçilmiştir. Öncelikle, öznitelik çıkarma işleminden önce, galeri kümesindeki yüksek çözünürlüklü yüz imgeleri farklı boyutlara düşürülüp tekrar modellerin girdi boyutuna getirilerek, düşük çözünürlükte imgeler elde edilmeye çalışılmıştır. Çözünürlükleri düşürmek için kullanılan boyutlar 24×24, 32×32, 40×40, 48×48, 64×64 şeklinde seçilmiştir ve daha sonra imgeler, ResNet ve SENet modelleri için 224×224 boyutuna, geliştirilmiş ResNet modelleri için ise 112×112 boyutuna getirilmiştir. Bu deney sonucu elde edilen en yüksek sonuçlar SCFace veri kümesi için 40×40, ve ICB-RW veri kümesi için ise 64×64 boyutları kullanılarak elde edilmiştir.

Bu tezde kullanılan VGGFace2 veri kümesi, farklı yaş gruplarından 9131 deneğe ait çeşitli pozlarda çekilmiş 3.31 milyon yüz imgesinden oluşmaktadır. Kullanılan ikinci veri kümesi olan ICB-RW, 90 denekten oluşmaktadır ve farklı pozlar, aydınlatma durumları, ifadeler, engelleri(gözlük, saç gibi.) barındıran imgeler içermektedir. Bir diğer veri kümesi olan SCFace ise 130 denekten oluşmaktadır. SCFace ve ICB-RW veri kümelerinin galeri kısımları için yüksek kaliteli ön yüz imgeleri kullanılmıştır. ICB-RW veri kümesinin test kısmı farklı poz, aydınlatma ve engel şartlarından ötürü tanıma açısından zorlu bir kümedir. SCFace test setinde ise yüz tanımayı zorlaştıran en büyük problem çok düşük çözünürlüklü imgeler içermesidir. SCFace veri kümesi test seti üç farklı mesafeden çekilmiş imgelerden oluşmaktadır. “Mesafe 1” (4.20 metre), “Mesafe 2” (2.60 metre) ve “Mesafe 3” (1.0 metre) kategorilerinden oluşan test kümesi imgeleri, 5 farklı gözetim kamerasından toplanmıştır. ICB-RW üzerinde elde edilen sonuçlar, önceki çalışmaların başarılımlarını büyük bir farkla geride bırakmıştır. Bu veri kümesi üzerindeki en iyi sonuçlar dört modelden çıkarılan öznitelikleri birleştirerek elde edilmiştir. Bütünleşik modelimizde, VGGFace2 üzerinde eğitilmiş 50 katmanlı ResNet, SENet, ve aynı modellerin önce MS-Celeb-1M veri kümesi üzerinde eğitilip daha sonra VGGFace2 üzerinde ince ayarı yapılmış modelleri bulunmaktadır. Sonuçlar ilk tahminde tespit için 91.78%, ilk 5 tahminde tespit için

98.00% ve CMC için 0.997 şeklindedir. Ayrıca, SCFace veri kümesi üzerinde de tek bir derin evrişimsel sinir ağı modeli kullanılarak, literatürde rapor edilmiş olan en iyi sonuçlar geçilmiştir. Bu veri kümesinde en yüksek yüz tanıma başarımı, “Mesafe 1” için önceden eğitilmiş Geliştirilmiş ResNet-100 modelini VGGFace2 veri kümesi üzerinde ince ayar yapılarak elde edilmiştir. Galeri imgelerinin çözünürlüklerini azaltmak ve daha geniş çerçeve ile yüz imgesi kırpma işlemi yapmak başarımların artmasına önemli katkı sağlamıştır. Elde edilen sonuçlar, önceki çalışmalarla karşılaştırmak için, SCFace veri kümesinde 20 farklı kez rastgele seçilmiş 80 denek üzerinde yüz tanıma yapılmıştır ve elde edilen sonuçların ortalaması ve standard sapması rapor edilmiştir. SCFace veri kümesi için en iyi elde edilen sonuçlar “Mesafe 1” için 76.94 ± 1.98, “Mesafe 2” için 98.41 ± 0.92, ve “Mesafe 3” için 100 ± 0.00’ ilk tahminde tespit yüz tanıma başarımıdır.

1. INTRODUCTION

There are many applications for biometric face identification under surveillance scenarios, e.g. whatchlist, security gates, and law enforcement. In these scenarios, the face descriptors of the probe individuals are matched, using a similarity measurement, to the face descriptors of the face images registered at the watchlist. Figure 1.1 illustrates the face identification protocol. Previously, the hand engineered feature extraction methods were used to extract and compare features to find the matches, e.g. Fisher Vectors (FV) [11]. The recent breakthroughs in deep learning algorithms [7, 8, 12, 13], proliferation of large-scale databases such as VGGFace2 [14], MS-Celeb-1M [15], and availability of high performance GPUs have aided the research in the field of face recognition. Nonetheless, the advancement has been only significant on the database that has high resolution images, where the gallery and the probe faces have similar domain, e.g. Labled Face in the Wild (LFW [16]) contains face images of celebrities of famous individuals which are downloaded from the Internet, whereas, YouTube Face (YTF [17]) database [17] is a collection of videos from YouTube. For the case of real-world biometric surveillance systems, comparable results have not been reported yet.

The recent reports on popular benchmarks, e.g, LFW [16] and YTF [17], for the face recognition under matched conditions exhibit near 100 percent accuracy, thus, the face recognition is considered as solved where the gallery and the probe faces are literally comming from the same source. FaceNet [6], a deep learning approach to the face recognition, demonstrated the robustness of the features extracted using very deep CConvolutional Neural Network (CNN) model. A CNN model with inception [13] modules is used for training, the features are L2 normalized and triplet loss is proposed to learn deep face representations. They used a proprietary database of 260 millions images to train their model. The proposed method achieved accuracies of 95.12%, 99.63% on YTF [17] and LFW [16], respectively. DeepID3 [18] proposed a CNN model using VGG [12] architecture including inception modules. Twenty-five crops of

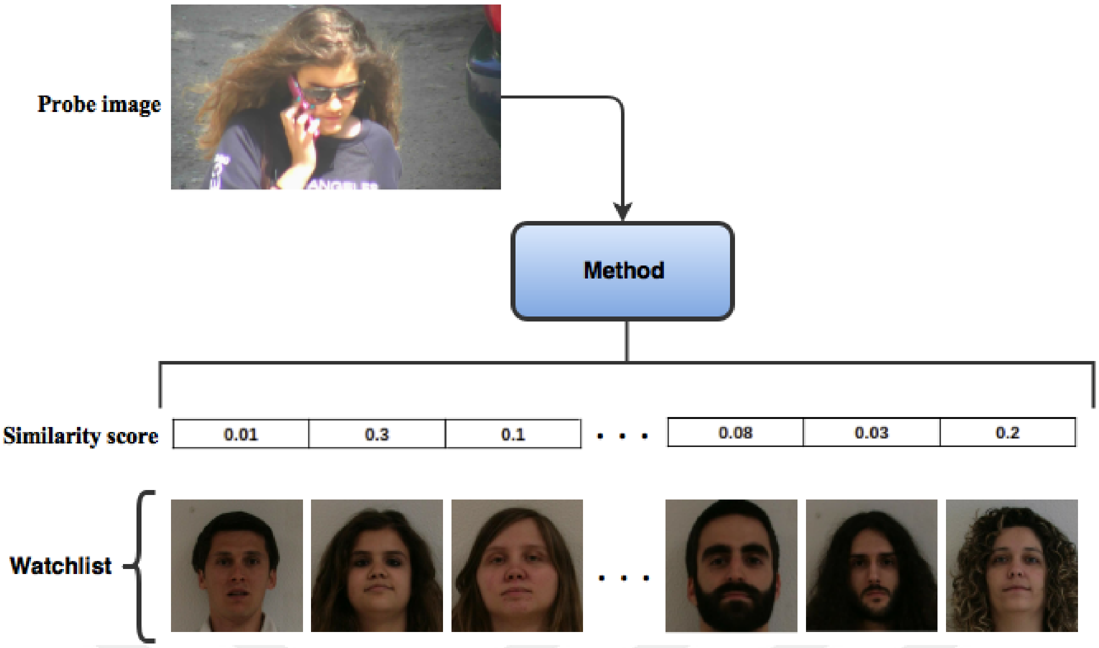


Figure 1.1 : Face identification protocol used in surveillance scenarios.

each face are fed into the model and approximately 30,000 features are extracted which then Principal Component Analysis (PCA) are used to reduce the feature dimension into 300 dimensional vector. Upon that, a joint Bayesian model for the face recognition is trained using 300 features. The reported accuracy of the model on LFW [16] is 99.54%. SphereFace [19] proposed a new loss function, e.g. the Angular-Softmax loss and learned face embedding in training phase with a ResNet [7] model. The nearest neighbor classifier with cosine similarity is applied for the face identification. They reported 99.42%, 95.0% accuracies on LFW [16] and YTF [17], respectively. ArcFace [9] used 50 layers improved ResNet architecture and train the face identification model with additive angular margin loss. The reported best verification accuracy is 99.83% on LFW database [16].

The deep learning advancements have been significant and mostly shift the direction of the research into the deep learning based feature extraction methods in the field of computer vision. In addition, the success of deep learning models on general face recognition inspired researchers to address more challenging problems in the face recognition accordingly, namely, face recognition under surveillance scenarios, where image qualities are very low and there are significant variation in the databases, e.g. pose, illumination, motion-blur, occlusion, expression, focus. SCFace [2] databases includes face images with three different low resolutions, distance 1 (4.20m), distance 2 (2.60m), and distance 3 (1.00m). Also, ICB-RW database is a challenging database which has the most of the aforementioned variations. As a feature extractor, we selected four state-of-the-art Deep Convolutional Nueral Networks (DCNNs) trained or fine-tuned on VGGFace2 [14] which have reported the highest results for face recognition on LFW benchmark [16]. Namely, ResNet-50 [7] and SENet-50 [8] architectures trained on VGGFace2 database [14] and the same architectures trained on MS-Celeb-1M database [15] and fine-tunned on VGGFace2 database [14]. Furthermore, we leveraged LResNet50E-IR [9] and LResNet100E-IR models [9] trained on VGGFace2 [14] and MS-Celeb-1M databases [15] with additive angular margin loss [9]. The models are explained in 4.2.1.

Face recognition under mismatched conditions is a challenging task in which there is usually a single high resolution frontal mugshot per subject in the gallery set, whereas, there are low resolution images captured with surveillance cameras in the probe set, and intrinsically contain variation in illumination, expression, pose, motion-blur, and focus. As a result of stated image quality problems, the model must be capable of learning face representation that are invariant to these changes in the database. In our experiments on ICB-RW database [1], we used provided bounding boxes and aliened the faces using implementation of [20] in dlib library [21]. Also, MTCCN [10] method is employed to detect the faces in SCFace [2] and ICB-RW [1] benchmarks. In our experiments on ICB-RW database, Residual Neural Networks (ResNet-50) [7] and Squeeze and Exitation Networks (SENet-50) [8] models are leveraged to extract 2048 dimensional face representations in the gallery and the probe sets. Afterwards, we normalized 2048D feature vectors and fed them into the classifier. Nearest neighbor with correlation distance as metric is used for classification. We calculated a $G \times P$ matrix of correlation scores for each model, where G is equal to the number identities in the galley images and P equals the number of images in the probe set. The proposed method achieved 91.78%, 98.0% Rank-1 and Rank-5 Identification Rates (IR) respectively on ICB-RW [1] using the ground truth bounding boxes. The Area Under the Curve of Cumulative Match Curve (CMC) on the probe set for ICB-RW dataset is 0.997. Our results significantly outperformed the results reported in [22] by the margins of 21.98%, 12.7%, and 0.045. Moreover, on the SCFace database, where there are probe images captured from 3 distances using 5 surveillance cameras with different qualities, we experimented the proposed method for the subsets of SCFace database, e.g. distance 1, distance 2, and distance 3 which are captured from 4.20m, 2.60m, 1.00m distances from the subjects respectively. We experimented with larger bounding boxes, and down-sampled gallery images to match the resolution of the gallery and the probe sets. We observed that increasing the size of bounding boxes and matching the resolution of the gallery and the probe faces before feature extraction, resulted in significantly higher Rank-1 IR on SCFace database [2], outperforming the state-of-the-art method [23]. Our best results for distance 1, 2, and 3 are 76.94% ±1.98, 98.41% ±0.92, and 100% ±0.00 respectively.

1.1 Purpose of Thesis

Our purpose of this thesis is to investigate the effectiveness and robustness of the deep learning models in extracting face descriptors. Particularity, we are interested in surveillance face recognition, which require features that are robust to strong variation in illumination, pose, occlusion, motion-blur, focus, and have very low resolution. We are motivated by the achievements in computer vision which have leveraged Convolutional Neural Networks on general object recognition tasks to achieve unprecedented results, we will leverage deep learning models to extract deep face representations which then will be employed to compare the faces in the gallery and the probe sets with a similarity measurement. We also explored two factors to increase the performance of deep learning models. We will demonstrate the strong capability of learned features to discriminate faces in very low resolution and show the effect of the proposed factors for improving the face identification results.

1.2 Literature Review

We review face recognition literature in general, and specifically focus on the literature of face recognition under mismatched conditions. Finally, we cover previous works which are most related to the result of this thesis.

In DeepFace [24], the authors have proposed an explicit 3D face modeling following by a nine layers Convolutional Neural Network to learn face features. They exploited 4.4 million face images of 4,030 identities, collected from Facebook and trained a deep network to learn a general face embedding for unconstrained conditions. The extracted

4096 dimensional features of each face are used for face identification having weighted χ^2 as similarity measurement. An ensemble of their proposed approach have 97.35% on LFW [16], and 91.4% on YTF [17] databases.

In FaceNet [6], an 128 dimensional face embedding is learned using an inception modules based CNN model, trained on a very large-scale database. Their private database have 200 millions face images of 8 million identities. Having this enormous database, they leveraged triplet loss to maximize the interclass distance of the learned features and minimize the intraclass distance of the features. The proposed method achieved 99.63%, 95.12% accuracies on LFW [16] and YTF [17] respectively.

Parkhi et al. [25], introduce VGGFace database in which there are 2,622 identities and 2.6 million images in total. They used triplet as the loss function of the model and a VGG architecture [12] is trained on the corresponding database. The face embedding is a 4096 dimensional feature vector. Their method achieved 98.95% accuracy on LFW, and 97.3% accuracy on YTF [17].

In SphereFace [19], the angular softmax loss is introduced. They used a very deep CNN with residual units and the angular softmax loss, trained on CASIA-WebFace database [26] to learn deep face descriptors. Finally, the trained model is leveraged to extract features in the test faces and cosine similarity is used for face identification. The proposed method obtained 99.42%, 95.00% accuracies on LFW [16], and YTF [17] databases respectively.

In ArcFace [9], additive angular margin loss is proposed. The authors leveraged ResNet architecture with 50 and 100 layers to learn face embedding using additive angular margin, which aims to maximize the interclass distance and minimize the intraclass distance simultaneously. The best model achieved state-of-the-art results of 99.83% on LFW [16]database.

The following works is closely related to ours in which the gallery face images are captured in the controlled lighting condition with a professional camera, whereas, the probe face images are captured using surveillance cameras.

In [27], local color vector binary patterns are extracted and nearest neighbor classifier with euclidean distance metric is used to find matches. Gallery face images and distance 3 (1.00m) images are used for learning. Average Rank-1 IR of 67.68% is reported for distance 1, and 2 (4.20m, 2.60m, respectively) of SCFace [2]. In [28], pose and illumination normalization are applied on faces and localized spatial correlation index is used for face matching. They reported 89% Rank-1 IR for distance 3 (1.0m) of SCFace [2]. In [29], Local-Consistency-Preserved Discriminative Multidimensional Scaling (LDMDS) approach is proposed to learn compact intra-class features and maximize inter-class distance. They used 50 subjects, out of 130 subjects available in SCFace [2], for training and reported Rank-1 IR for remaining 80 subjects. They reported 62.7%, 70.7%, and 65.5% Rank-1 IR for distance 1, 2, and 3, respectively.

Following [29], in Deep Coupled ResNet (DCR) [23], the authors proposed a trunk and two branches based deep CNN model. They down-scaled CASIA-WebFace database [26] into three different resolutions, 112×96, 40×40, and 6×6, and then up-scaled two lower resolution images to 112 x 96, which is the input size of their model. They used combination of softmax and center losses in the training step of the trunk network and then freezed the parameters of the trunk network when training the two branches. Combination of softmax, center and square distance losses are used to learn a 512 dimensional face embedding. The proposed method minimizes intraclass distance, maximizes interclass distance, and also minimizes the distance of features extracted from low resolution and high resolution faces of the same identity. Finally, they fine-tuned the two branches using randomly selected 50 subjects of SCFace [2] and reported the result for the remaining 80 subjects. The proposed approach achieved 73.3%, 93.5%, and 98.0% accuracies on distance 1, distance 2, and distance 3.. As it can be seen from the results, performance of the proposed methods deteriorate significantly when the resolution of the probe faces decrease. To the best of our knowledge, the highest results reported on SCFace benchmark [2] is reported in [23]. There are also deep learning based super-resolution methods to deal with low resolution faces, however, these methods are not optimized for Low Resolution Face Recognition (LRFR) [30] or yield modest performance improvement [31].

1.3 Hypothesis

In the machine learning applications we usually ignore the fact that the test and the train data do not have the same distribution. So, we expect that learning a perfect model on training data would resulted in higher accuracy on the test data. However, in the most of the cases this is not true. In other words, the train and the test distributions are not the same. Thus, transfer learning is a technique which have been exploited to bridge the gap between source and target domain distribution [32]. In object recognition problems, one of the main challenges are the lack of the labeled train and test data which are expensive to collect. Thanks to very-large image databases, e.g. ImageNet [33], researchers have demonstrated that embedding learned by Deep Convolutional Neural Networks (DCNNs) are transferable to the similar domains via feature extraction or fine-tuning.

Inspired by the power of deep learning models in learning discriminative features, and access to the models trained on large-scale face databases, e.g.CASIA-WebFace [26], MS-Celeb-1M [15], VGGFace2 [14], we hypothesize that transfer learning can be leveraged to extract robust face features invariant to changes in pose, illumination, resolution, occlusion. We also proposed that increasing the amount of information included in the cropped faces would improve the performance in two ways. Firstly, as the input size of the DCNNs are high (224×224 and 112×112 in our models) enlarging the bounding boxes would decrease the up-sampling factor of the faces which would resulted in less degradation in the faces. Secondly, the low resolation faces has limited information and extending the bounding boxes would allow the models two increase more information about the face, e.g. about the shape o the face, and etc. In this work, we employed ResNet-50 [7] and SENet-50 [8], LResNet50E-IR [9] and LResNet100E-IR [9] architectures, trained or fine-tuned on VGGFace2 and MS-Celeb-1M databases, to extract face features. We used the output of the final layer in each model, prior to the classification layer, as the features of faces. The learned feature vectors are normalized and then are given as input for the nearest neighbor classifier with correlation distance metric as similarity measurement. The experimental results on ICB-RW [1], and SCFace [2] databases are significantly superior to the best reported results in the literature.

2. CONVOLUTIOAL NEURAL NETWORKS

2.1 Components

In this thesis, we leveraged the Convolutional Neural Networks (CNN) to learn robust face embedding. The CNN has different layers in which convolution and fully connected layers have the most of the learn-able parameters in the deep learning models. Here, we briefly review the different parts of the deep learning models.

2.1.1 Convolution layer

The convolution layer is the crucial part of a CNN which contains a set of kernels initialized by random weights at the start of the training phase. During the training, the convolution operation is done between input from previous layers and kernels, and wights of the kernels are updated during back-propagation. Figure 2.1 illustrate a 2D convolution operation between a 2×2 kernel and 3×3 input. In this example, the stride of 1 is used which is the number of the columns or rows that kernel move after each operation. The stride of 2 is used to reduce the dimension of the output by half. The equation 2.1 shows the convolution operation between input (I) and kernel (K). In addition, suppose we have an input size of $W_1 \times H_1 \times D_1$, the output size of the convolution operation with K kernels of size $F \times F$ is calculated by equations 2.2.

$$S(i,j) = (I * K)(i,j) = \sum_m \sum_n I(i-m, j-n)K(m,n) \tag{2.1}$$

$$\begin{aligned} W_2 &= \frac{W_1 - F + 2P}{S+1} \\ H_2 &= \frac{H_1 - F + 2P}{S+1} \\ D_2 &= K \end{aligned} \tag{2.2}$$

Where P is the padding size, and S is the stride.

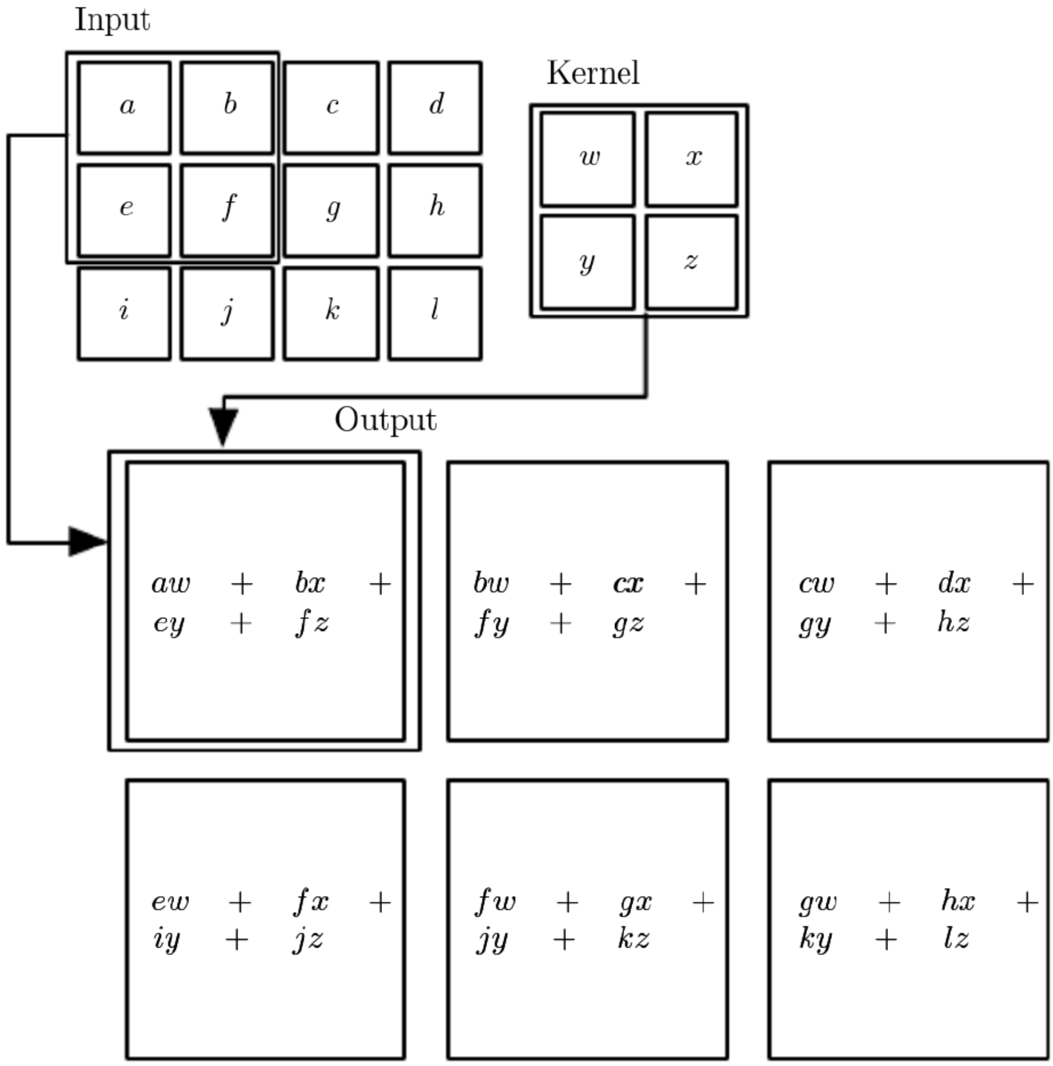


Figure 2.1 : A 2D convolution operation [3].

2.1.2 Activation functions

We need a function that adds non-linearity to the neural networks which are the stack of the linear regression units on top of each other, this is why we have activation functions in neural networks. In other words, a neural network without an activation function is just a sequence of linear functions. There are several activation functions which are used in the deep learning architectures. Sigmoid, Tanh, Rectified Linear Units (ReLU), and Parametric Rectified Linear Unit (PReLU) are the most addressed ones in the literature. The equations of these activation are as follows:

Sigmoid function:

$$f(x) = \frac{1}{1+e^{-x}} \tag{2.3}$$

Tanh function:

$$f(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \tag{2.4}$$

ReLU function:

$$f(x) = max(0,x) \tag{2.5}$$

PReLU function:

$$f(x) = max(0,x) + \alpha min(0,x) \tag{2.6}$$

Of these activation functions ReLU has gained popularity due to efficiency and fast convergence.

2.1.3 Pooling layer

The pooling layer is a sub-sampling layer which is used to reduce the size of the input features. The most addressed pooling functions in the literature of deep learning are max polling, and average pooling. The pooling operation is carried out using a kernel which have a size of usually 2×2, 3×3, 5×5, 7×7, and less frequently 11×11 or larger. In the case of max pooling, the filter is placed in every corresponding position of the input with stride 1 or 2, and the result is the maximum of the input at every receptive field, or it output the mean of input values in average pooling. Figure 2.2 demonstrate a 3×3 input and 2×2 output of each pooling operation with stride 2.

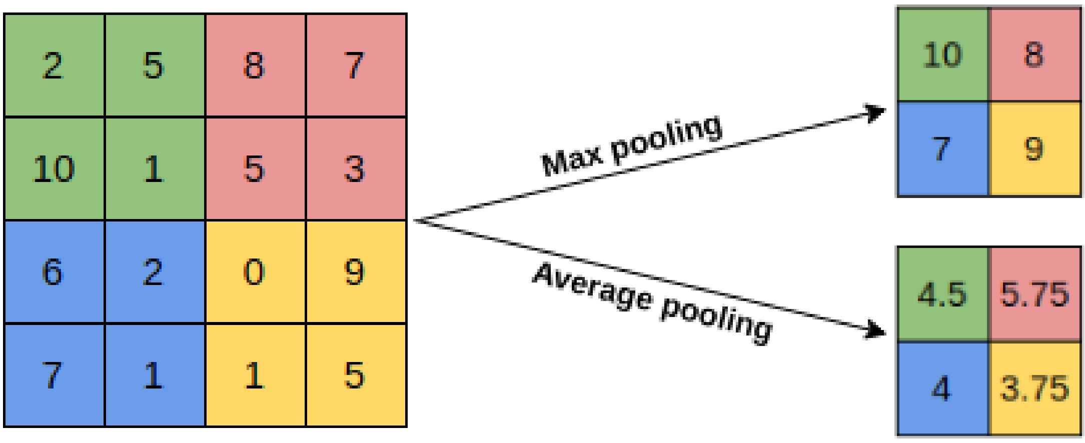
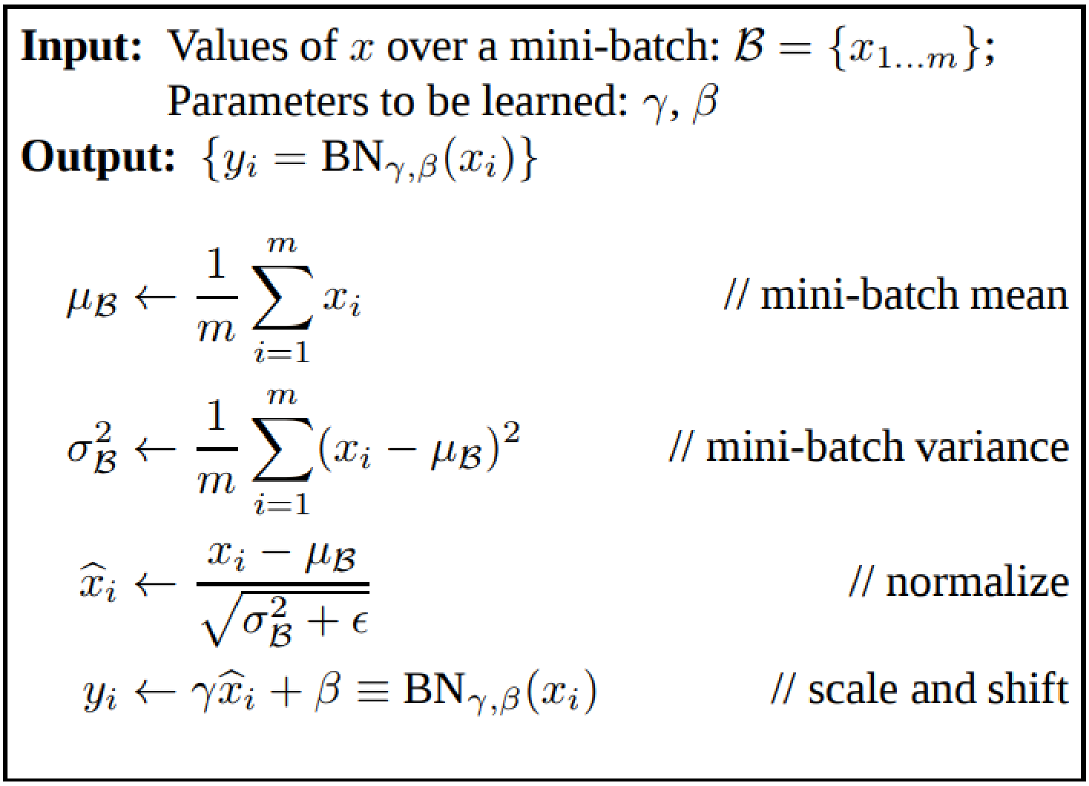


Figure 2.2 : A 2×2 pooling operation.

2.1.4 Batch normalization

In the Batch Normalization (BN) layer input of each layer is normalized for mini-batches during training. By doing so, internal co-variate shift problem is solved [4] and models can be trained using higher learning rate. The batch normalization layer has also the regularization effects which reduce the necessity of using Dropout layer. Figure 2.3 illustrate the algorithm for batch normalization.



Input: Values of x over a mini-batch: $\mathcal{B} = \{x_{1 \ldots m}\}$;
Parameters to be learned: γ, β

Output: $\{y_i = \mathrm{BN}_{\gamma,\beta}(x_i)\}$

$$\mu_{\mathcal{B}} \leftarrow \frac{1}{m}\sum_{i=1}^{m} x_i \qquad \text{// mini-batch mean}$$

$$\sigma_{\mathcal{B}}^2 \leftarrow \frac{1}{m}\sum_{i=1}^{m} (x_i - \mu_{\mathcal{B}})^2 \qquad \text{// mini-batch variance}$$

$$\widehat{x}_i \leftarrow \frac{x_i - \mu_{\mathcal{B}}}{\sqrt{\sigma_{\mathcal{B}}^2 + \epsilon}} \qquad \text{// normalize}$$

$$y_i \leftarrow \gamma \widehat{x}_i + \beta \equiv \mathrm{BN}_{\gamma,\beta}(x_i) \qquad \text{// scale and shift}$$

Figure 2.3 : Batch normalization algorithm [4].

2.1.5 Dropout

As there are many parameters in deep learning architectures, the model might be more complex than necessary which causes high variance and over-fitting on the train set. Dropout is a regularization technique that is used to learn a general model [5]. Dropout is used to remove some neurons and their connections with other neurons by specifying a drop probability, so as to have different set of neurons during each feed-forward. As a result, the learned model is not dependent on some specific features that co-exist. Figure 2.4 shows a neural network before and after applying Dropout.

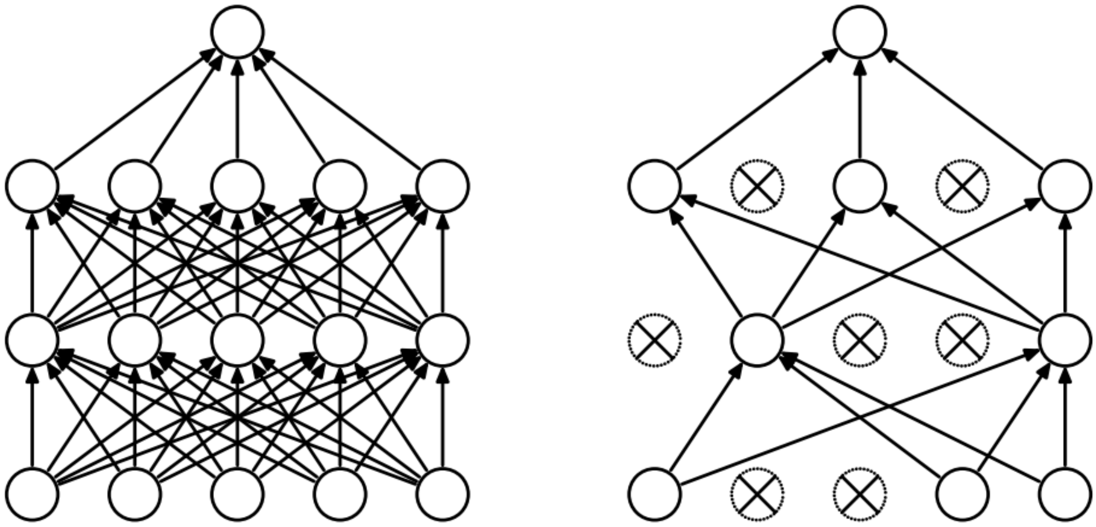


Figure 2.4 : A neural network before and after Dropout [5].

2.1.6 Fully connected layer

In contrast to Convolutional Neural Networks which are connected locally and share the weights, Fully Connected layers (FC) are a type of neural networks in which every input to the next layer is connected to every neuron in that layer. The equation 2.7 shows the operations in fully connected layers, where X is the input, W is the weight matrices, b is bias terms, f is an activation function, and Z is output.

$$Z = f(W^T X + b) \tag{2.7}$$

2.1.7 Classification layer

In every iteration over the mini-batches, the final layer outputs the probability of classes over all the samples in the mini-batch. This probability distribution is used to predict class labels and calculate the loss of the training step.

2.2 Loss Functions

In this section we explain the losses that are used for training the models we leveraged in this thesis and also other losses in the literature.

2.2.1 Cross entropy loss

Cross entropy loss or Softmax loss is used for multi-class classification problems. First, probability of each class is calculated using the softmax function, equation 2.8, afterwards, cross-entropy loss is calculated for each training example, equation 2.9. The total loss is the average of cross-entropy loss on all the training samples.

$$f_j = \frac{e^{z_j}}{\Sigma_k e^{z_k}} \tag{2.8}$$

Where f is the probability of the sample being from class j, and z is the class score.

$$L_i = -log\left(\frac{e^{f_{y_i}}}{\Sigma_j e^{f_j}}\right) \tag{2.9}$$

Where f_{y_i} is the normalized probability of the correct class.

2.2.2 Triplet loss

In the triplet loss, we try to minimize the distance between embedding of an Anchor (a) with the Positive (p) samples, and maximize the distance between embedding of the Anchor with the Negative (n) samples. The desired result on triplet loss function is illustrated in Figure 2.5. The triplet loss function is formulated in equation 2.10. In the triplet loss, selecting the triplets are crucial to learn a powerful model, however, selecting the most useful triplets that would contribute to the training is a cumbersome task. Also, triplets grows exponentially as the size of training databases increase. Thus, recently angular margin [19], [9] are proposed to modify the softmax loss.

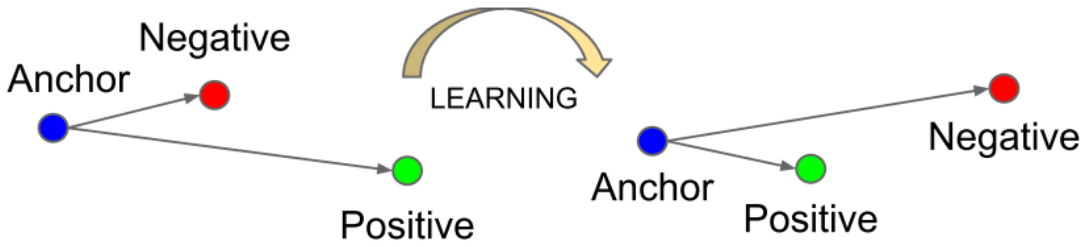


Figure 2.5 : Triplet loss illustration. [6].

$$L = \sum_{i}^{N} \left[\|f(x_i^a) - f(x_i^p)\|_2^2 - \|f(x_i^a) - f(x_i^n)\|_2^2 + \alpha \right]_+ \tag{2.10}$$

2.2.3 Additive angular margin

The proposed ArcFace loss [9] is the modified version of softmax loss in which the objective function is to maximize the intraclass compactness and the interclass distances. The ArcFace loss is calculated with equation 2.11.

$$L = -\frac{1}{m}\sum_{i=1}^{m} log \frac{e^{s(cos(\theta_{y_i}+m))}}{e^{s(cos(\theta_{y_i}+m))} + \sum_{j=1,j\neq y_j}^{n} e^{scos\theta j}} \tag{2.11}$$

subject to

$$W_j = \frac{W_j}{||W_j||}, x_i = \frac{x_i}{||x_i||}, cos\theta_j = W_j^T x_i. \tag{2.12}$$

Where s in 2.11 is the scale factor for features.

3. TRANSFER LEARNING

Transfer learning is a technique which is used when we do not have enough training data to train a deep learning model for our problem. Therefore, we transfer the knowledge from a pre-trained model on a large scale database. For instance, the models trained on very large scale ImageNet database [33] can be leveraged for transfer learning on other classification problems. Transfer learning can be used for feature extraction or fine-tuning. The transfer learning method we use depends on the number of training samples we have for each classes and the similarity of the source and target domains.

3.1 Feature Extraction

In the problems that we do not have enough data, we can use a model trained on a very close domain to target domain for feature extraction. For instance, here, in this thesis we leveraged DCNNs to extract feature for face recognition on ICB-RW [1], and SCFace [2] benchmarks. The DCNNs were trained using a very large scale face databases . The number of samples in ICB-RW and SCFace benchmarks are insufficient for training or fine-tuning the pre-trained models, thus, we leveraged the discriminative power of the learned face embedding for face identification on our relatively small databases.

3.2 Fine Tuning

Based on the magnitude of our target domain we can fine-tune only top or more layers of a pre-trained model. Fine-tuning would help to boost the accuracy on our target domain. By increasing the number of the available samples for each class, we can train more of the top layers. For instance, if there are enough face data, we can fine-tune the classification layer plus some of the top layers to extract more representative features specific to our target domain.

3.3 Deep Learning Architectures

In this section we briefly describe the deep learning architectures we used in our experiments.

3.3.1 Residual neural network

Residual Neural Networks (ResNet) [7] are built using residual blocks. Let us suppose that we try to learn an $H(X)$ function that is the mapping from the input X of a few convolution layers to the output of them. If we add the input X to the output, we let the convolution layers to approximate $F(X) := H(X) + X$. Learning $F(X)$ function is easier than learning $H(X)$ as proposed in [7]. A residual block is shown in Figure 3.1.

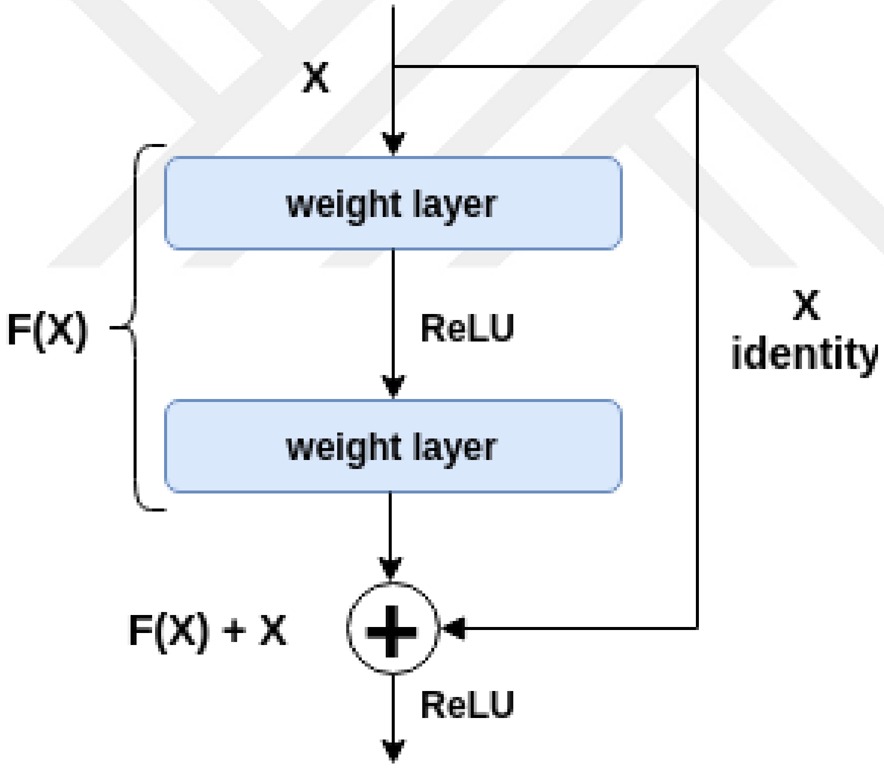


Figure 3.1 : A Residual Block [6].

In [7] the specification for 5 versions of ResNet with different number of layers are reported (as shown in Figure 3.2). We leveraged 50 layers ResNet in our experiments.

layer name	output size	18-layer	34-layer	50-layer	101-layer	152-layer
conv1	112×112	7×7, 64, stride 2				
conv2_x	56×56	3×3 max pool, stride 2				
		$\begin{bmatrix} 3\times3, 64 \\ 3\times3, 64 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 64 \\ 3\times3, 64 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}\times3$
conv3_x	28×28	$\begin{bmatrix} 3\times3, 128 \\ 3\times3, 128 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 128 \\ 3\times3, 128 \end{bmatrix}\times4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}\times4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}\times4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}\times8$
conv4_x	14×14	$\begin{bmatrix} 3\times3, 256 \\ 3\times3, 256 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 256 \\ 3\times3, 256 \end{bmatrix}\times6$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}\times6$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}\times23$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}\times36$
conv5_x	7×7	$\begin{bmatrix} 3\times3, 512 \\ 3\times3, 512 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 512 \\ 3\times3, 512 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix}\times3$
	1×1	average pool, 1000-d fc, softmax				
FLOPs		1.8×10^9	3.6×10^9	3.8×10^9	7.6×10^9	11.3×10^9

Figure 3.2 : The number of the parameters of the Residual Neural Networks with different number of layers [7]. $[kernelsize, numberofkernels] \times numberofblocks$ in the architecture.

3.3.2 Squeeze-and-excitation networks

The Squeeze-and-Excitation (SE) block [8] is proposed to model the inter-dependencies between the channels of a network's convolutional features. The Squeeze-and-Excitation Networks are built upon stacking SE blocks. SE block can be added to the sate-of-the-art networks, e.g. ResNet [7] to increase the capacity of the network in emphasizing descriptive features, meanwhile, eliminating less important features. The inputs to a SE block are passed through squeeze and excitation operations which are described in 3.3.2.1, and 3.3.2.2 respectively.

3.3.2.1 Squeeze: global information embedding

In the squeeze operation, a global average pooling is applied on each channel of the input features. Suppose we have $\mathbf{U} \in \mathbb{R}^{W \times H \times C}$ as an output from a previous convolution layer, the equation 3.1 is the squeeze step of SE block which produce channel statistics $z \in \mathbb{R}^{C}$.

$$z_c = F_{sq}(u_c) = \frac{1}{W \times H} \sum_{i=1}^{W} \sum_{j=1}^{H} u_c(i,j). \tag{3.1}$$

3.3.2.2 Excitation: adaptive re-calibration

To fully utilize the channel statistics in squeeze operation, the results from squeeze operation are passed trough a gate with the sigmoid activation as formulated in equation 3.2.

$$s = F_{ex}(z,W) = \sigma(g(z,W)) = \sigma(W_2\delta(W_1 z)), \tag{3.2}$$

where δ is the ReLU activation function, $W_1 \in \mathbb{R}^{\frac{C}{r} \times C}$, $W_2 \in \mathbb{R}^{C \times \frac{C}{r}}$ and r is the ratio of a dimensionality reduction layer with parameters W_1.

Furthermore, two FC layers sandwiched the dimensionality reduction layer as a bottleneck to decrease the complexity of the model. Finally, U is re-scaled with the activation as shown in equation 3.3.

$$\tilde{x}_c = F_{scale}(u_c, s_c) = s_c.u_c, \tag{3.3}$$

where $\tilde{X} = [\tilde{X}_1, \tilde{X}_2, ..., \tilde{X}_C]$ is the multiplication of the feature map of each channel, $u_c \in \mathbb{R}^{W \times H}$ and the scalar s_c. Figure 3.3 shows a SE block and Figure 3.4 illustrates the SE block added into a residual block.

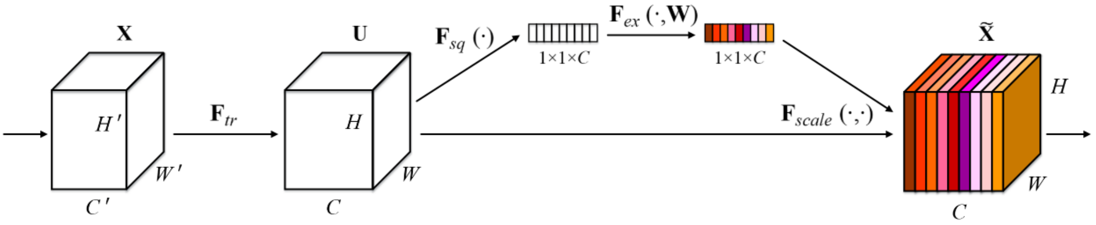


Figure 3.3 : A Squeeze and Excitation block [8].

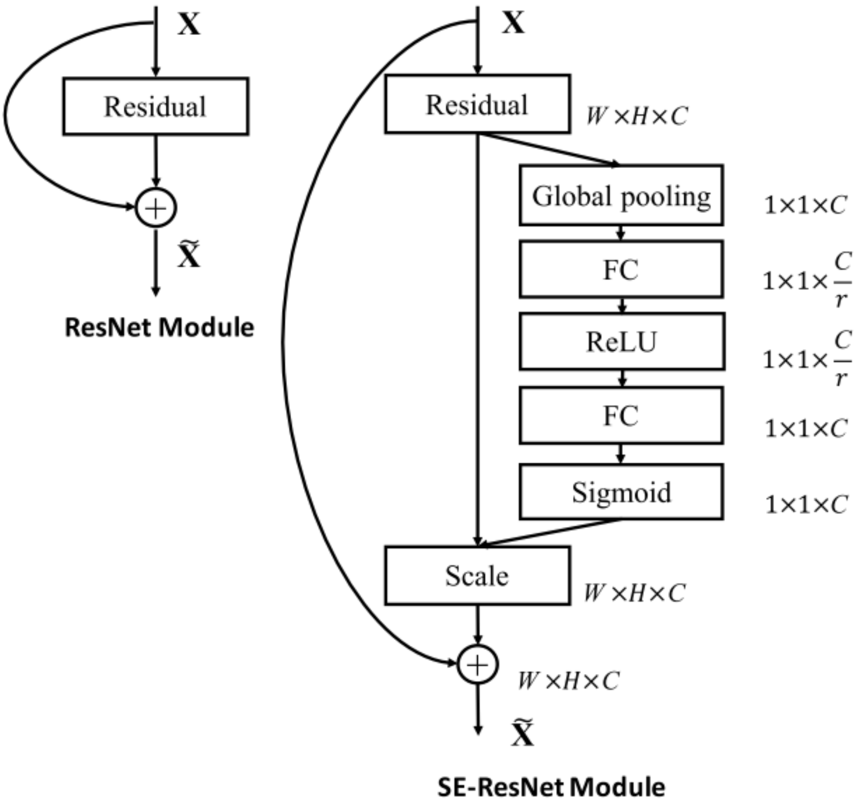
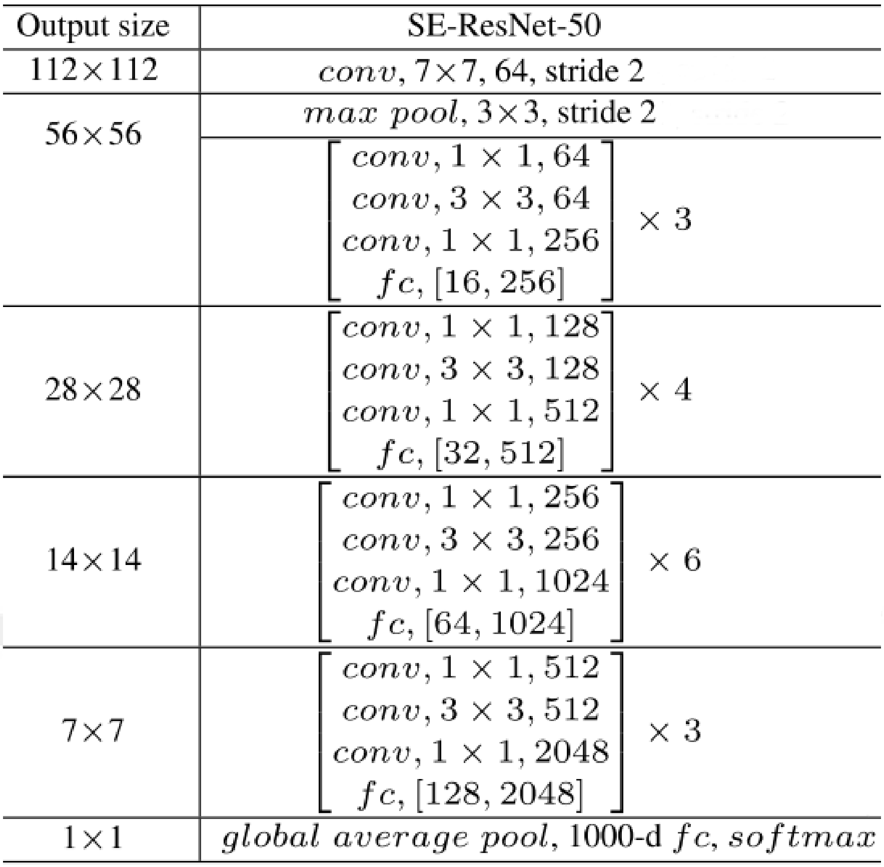


Figure 3.4 : **Left**: A residual block. **Right**: A SE block added to a residual block, SE-ResNet [8].

In our experiments we utilized SE-ResNet with 50 layers and referred to it as SENet-50. The SENet-50 model specification are given in 3.5.

3.3.3 Improved residual neural networks

We also exploited improved Residual Neural Networks [9] for learning face embedding in our experiments. In the improved residual blocks, the first ReLU activation layer is



Output size	SE-ResNet-50
112×112	$conv$, 7×7, 64, stride 2
56×56	$max\ pool$, 3×3, stride 2
	$\begin{bmatrix} conv, 1\times1, 64 \\ conv, 3\times3, 64 \\ conv, 1\times1, 256 \\ fc, [16, 256] \end{bmatrix} \times 3$
28×28	$\begin{bmatrix} conv, 1\times1, 128 \\ conv, 3\times3, 128 \\ conv, 1\times1, 512 \\ fc, [32, 512] \end{bmatrix} \times 4$
14×14	$\begin{bmatrix} conv, 1\times1, 256 \\ conv, 3\times3, 256 \\ conv, 1\times1, 1024 \\ fc, [64, 1024] \end{bmatrix} \times 6$
7×7	$\begin{bmatrix} conv, 1\times1, 512 \\ conv, 3\times3, 512 \\ conv, 1\times1, 2048 \\ fc, [128, 2048] \end{bmatrix} \times 3$
1×1	$global\ average\ pool$, 1000-d fc, $softmax$

Figure 3.5 : SE-ResNet-50 layers parameter settings [8].

removed and the second one is replaced with PReLU. Also, The Batch Normalization (BN) layer is added after the second convolution layer (as shown in Figure 3.6). Moreover, there are some setting in the network input, and output. For the input setting, "**L**" in the network name shows that $conv3\times3$ and $stride = 1$ is used in the first convolution layer of the residual blocks, and in the output setting, "**E**" illustrate that the output of the network is passed trough BN-Dropout-FC-BN layers to learn the final 512 dimensional embedding. In addition, **IR** at the end of the network name specify the usage of improved residual blocks in the architecture.

Two 50 and 100 layers improved residual networks are leveraged in our experiments, namely, LResNet50E-IR [9] and LResNet100E-IR [9].

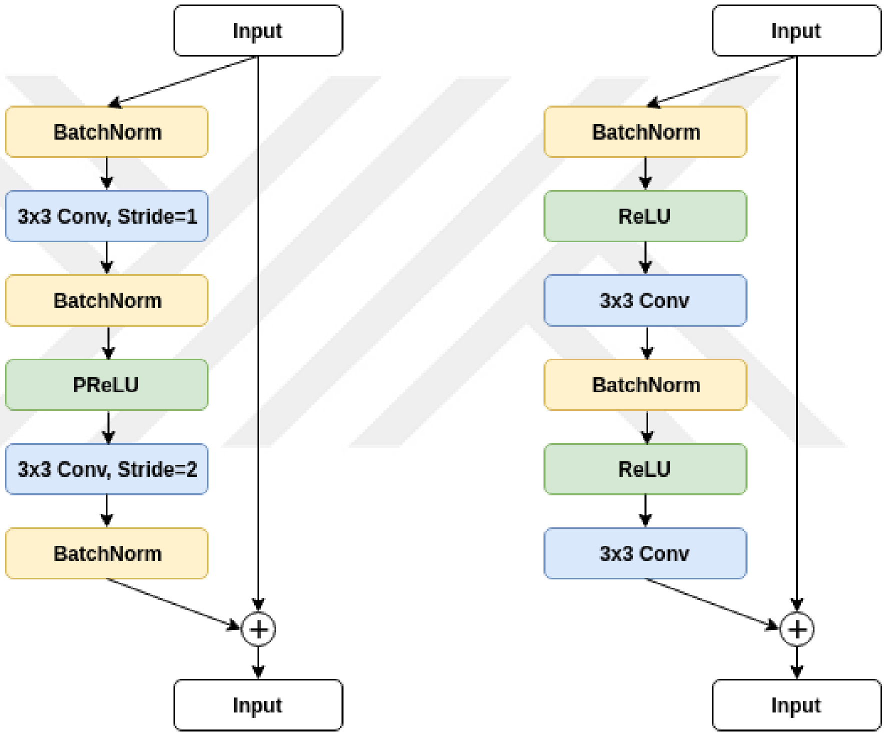


Figure 3.6 : **Left**: Improved residual block [9]. **Right**: Original residual block [7].

4. METHODOLOGY

In this chapter, the methods and databases that are used in the experiments are explored.

4.1 Datasets

For the evaluation of the face embedding learned with deep learning models, the International Challenge on Biometric Recognition in the Wild 2016 database (ICB-RW) [1], and SCFace [2] benchmarks are used. These databases have the characteristic of surveillance data and are collected under mismatched conditions. There are variation in illumination, pose, expression, occlusion, motion-blur, and Focus in ICB-RW [1]. Also, SCFace [2] contains probe face images with three different resolution. Figure 4.1 shows some instance of the aforementioned problems.

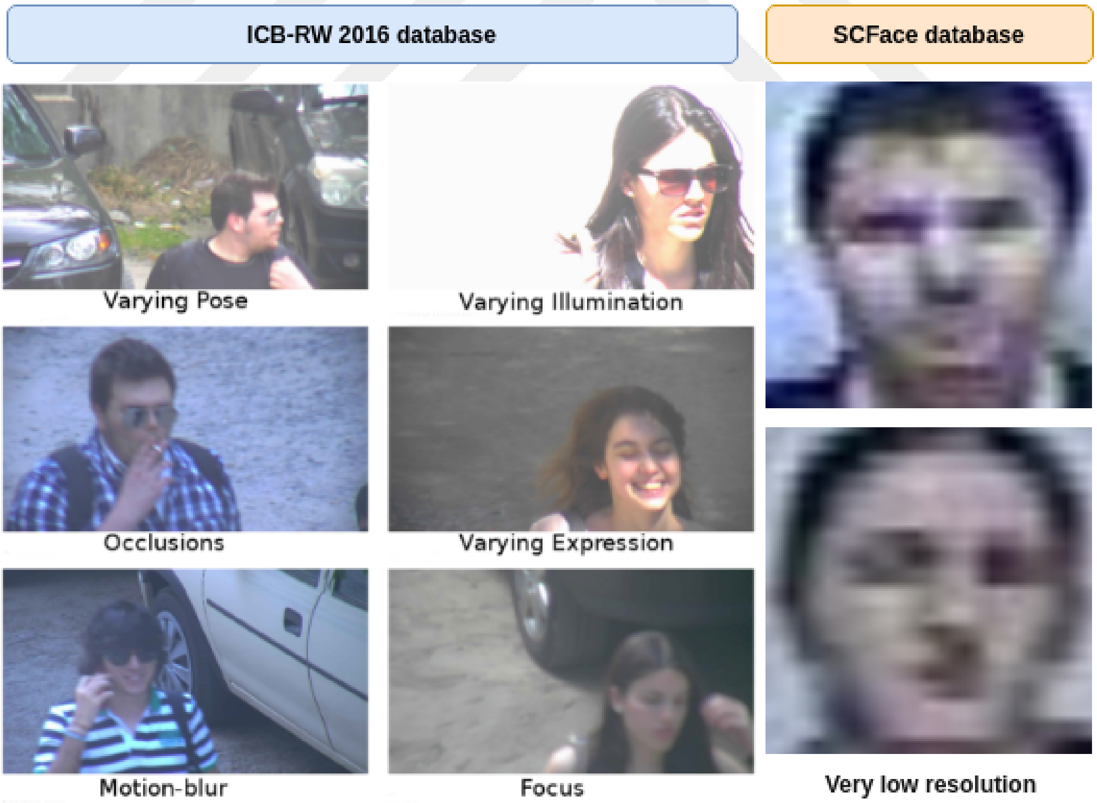


Figure 4.1 : Samples from ICB-RW contains variation in pose, illumination, occlusion, and etc., whereas, SCFace has samples with very low resolution.

4.1.1 ICB-RW

The ICB-RW which is used for evaluation in this thesis, contains gallery and probe images [1]. There are 90 subjects in the database, each of them have 3 gallery and 5 probe images. Figure 4.2 shows train and test set samples for two subjects in the database. The three images on the left are two left/right profiles, and mugshots (gallery images), whereas, 5 images in the right are collected using a surveillance camera and are probe images. The gallery images have high resolution and are captured under controlled conditions, whereas, there is not any control on capturing probe images, so as to including variation in pose, illumination, motion-blur, focus, etc. The probe images are collected from surveillance camera mounted outside the building to represent the characteristics of the surveillance systems. In this work, we only used frontal gallery images and 5 probe images of each subject for identification task. The bounding boxes for the face images are provided in the database.

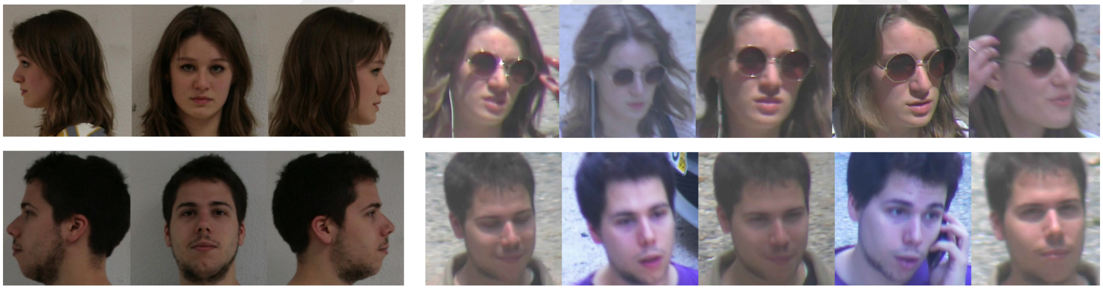


Figure 4.2 : Samples of two subjects in ICB-RW 2016 database.

4.1.2 SCFace

SCFace database [2] are collected from 130 subjects. In order to capture the subject's faces in different poses, 9 high quality images per person are captured in controlled indoor lighting condition. There are 8 different poses for each subject in the gallery, from left profile to the right profile, and 22.5 degree difference between each, so, the images at -90 and +90 degrees are left and right profiles and mugshot image is at 0 degree. In this work, we used only images at 0 degree (frontal mugshot) as the gallery images in face identification. There is also one mugshot, taken with IR night vision camera, for face identification in night vision surveillance scenarios. The probe images are captured with 5 surveillance camera mounted in one room, 2.25m above

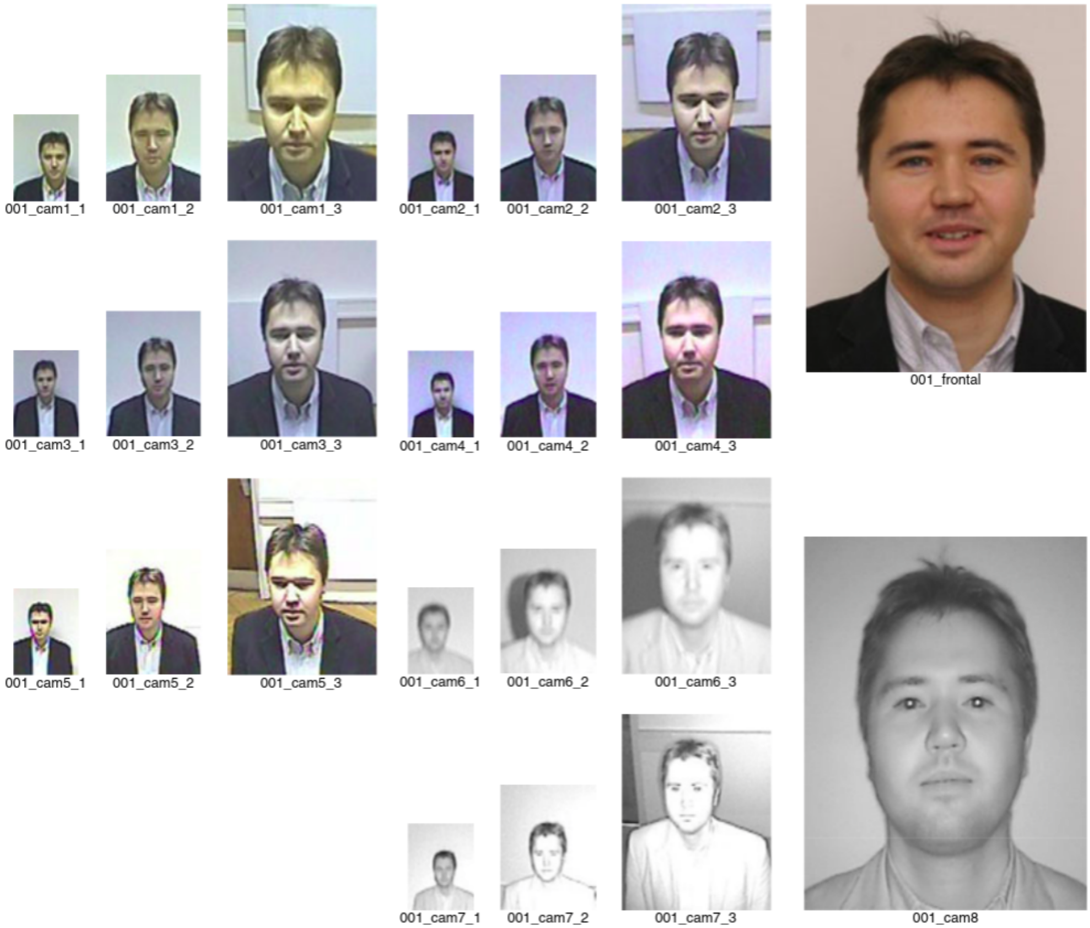


Figure 4.3 : Sample images of one subject in SCFace [2]. Here, the frontal face images in the gallery (taken with IR night vision and normal vision) and the probe face images captured using five surveillance camera are shown cam1, ..., cam5. Two of the surveillance cameras have IR night vision which are used to capture face images in cam6 and cam7. The number in the end of the probe faces' name represent distance 1, 2, and 3.

the ground. Also, two of surveillance cameras which are capable of capturing IR night vision, are used to collect IR probe images. The subjects were ask to stop at three previously marked positions, distance 1 (4.20m), distance 2 (2.60 m), and distance 3 (1.00 m). As a result, 15 probe images are collected per subject for day time test, and 6 probe images are taken for night time test. Figure 4.3 illustrate gallery and probe images for one subject.

4.2 Proposed Method

In this section, we describe the method we proposed for the face identification step-by-step. In the first step, faces are detected and (aligned only for ICB-RW with ground-truth bounding boxes) in all the images of the gallery and the probe set. Later on, eight pre-trained deep learning models are leveraged to extract face embedding as listed in 4.2.1. Finally, the nearest neighbor classifier with correlation distance is used for face identification. We reported, Rank-1, Rank-5 Identification Rate (IR), and Area Under the Curve of Cumulative Match Curve for ICB-RW [1] and Rank-1 IR for SCFace [2]. Pipeline of our method is depicted in Figure 4.4.

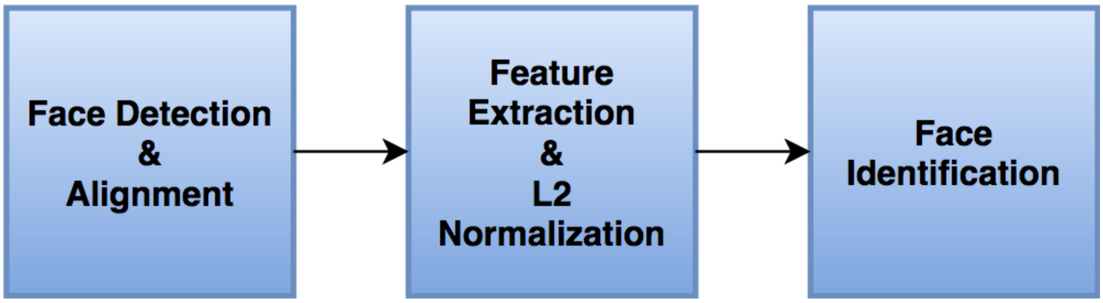

Face Detection & Alignment
Feature Extraction & L2 Normalization
Face Identification

Figure 4.4 : Pipeline of the proposed face identification.

4.2.1 Models

The VGGFace2 [14], and MS-Celeb-1M [15] databases are used to train the models, and the robustness of the learned face embedding are evaluated on ICB-RW [1], and SCFace [2] benchmarks. Table 4.1 illustrates the scale of the databases.

Table 4.1 : The scale of databases that are used in the experiments.

Databases	# of Subjects	# of images	# of images per subject
ICB-RW 2016	90	540	6
SCFace	130	2080	16
VGGFace2	9,131	3.31 M	87/362.6/843
MS-Celeb-1M	100,000	10 M	100

VGGFace2 [14] and MS-Celeb-1M [15] are very large-scale face databases which are collected from the celebrity images on the Internet. In theory, if we train state-of-the-art deep learning networks on these databases, learned face embedding should generalize well on unseen databases. In order to learn face representation, Residual Neural Networks (ResNet) [7], and ResNet with Squeeze and Excitation blocks [8] are trained on aforementioned databases. The eight models that we employed in the experiments are trained as follows:

1) **VF2-ResNet**: A ResNet-50 [7] model trained on VGGFace2 [14]. .

2) **VF2-ft-ResNet**: A ResNet-50 [7]model fine-tuned on VGGFace2 [14] after being trained on MS-Celeb-1M (MS1M) [15].

3) **VF2-SENet**: A SENet-50 [8] model trained on VGGFace2 [14].

4) **VF2-ft-SENet**: A SENet-50 [8] model fine-tuned on VGGFace2 [14] after being trained on MS-Celeb-1M [15].

5) **LResNet50E-IR**: An improved Residual Neural Network [9] with 50 layers trained on MS-Celeb-1M [15] database.

6) **LResNet100E-IR**: An improved Residual Neural Network [9] with 100 layers trained on MS-Celeb-1M [15] database.

7) **VF2-ft-LResNet50E-IR**: An improved Residual Neural Network [9] with 50 layers fine-tuned on VGGFace2 [14] after being trained on MS-Celeb-1M [15].

8) **VF2-ft-LResNet100E-IR**: An improved Residual Neural Network [9] with 100 layers fine-tuned on VGGFace2 [14] after being trained on MS-Celeb-1M [15].

4.2.1.1 Fine-tuning the models.

Off-the-shelf pre-trained models are available for models 1-6. However, VF2-ft-LResNet50E-IR, and VF2-ft-LResNet100E-IR models are fine-tuned as follows:

A) ***Data preprocessing***: The faces in VGGFace2 [14] database are aligned with similarity transform using the center of the eyes, tip of the nose, and the corners of the mouth. The detected faces are cropped and resized to 112×112, and finally pixel values are normalized by subtracting 127.5 and dividing by 128.

B) ***LResNet50E-IR*** model is fine-tuned on VGGFace2 [14] using additive angular margin loss (as explained in 2.2.3) with $m = 0.5$, and $s = 64.0$. Stochastic Gradient Descent with momentum of 0.9 and learning rate of 0.01 are used to fine-tuning the Network with batch size of 64. The training is stopped manually when the accuracy of the test set plateau (99.6% on LFW [16] database). We named the fine-tunned model as *VF2-ft-LResNet50E-IR*.

C) ***LResNet100E-IR*** model is fine-tuned on VGGFace2 [14] with the same setting as in **B**, however the learning rate is set to 0.001 and the training is stopped manually when the accuracy on test set (LFW [16]) achieved 99.7%. We named the fine-tuned model as *VF2-ft-LResNet100E-IR*.

4.2.2 Face detection and alignment

Similar to all machine learning models, we need to normalize the faces before feeding them into the deep learning models for feature extraction. Two step that are required for the purpose of normalization in face recognition are face detection and alignment, the first step is detecting the faces and the second step is alignment using facial landmarks that would help to increase the performance in the cases which include strong variation in poses. Normalizing the faces helps the model to extract discriminant features from the same part of the face images for all the faces in the databases. In our experiments on ICB-RW database, we used both the bounding box meta-data in the database for cropping the faces. Upon that, we used off-the-shelf dlib [21] model, implemented based on [20] to align the faces. The 5 facial landmarks are used which are the eyes corners, and the nose bottom which are detected using the stated model, and the alignment are done using the detected landmarks. We also evaluate the performance of our models on ICB-RW database [1] using face crops using bounding box coordinates detected by MTCNN [10]. For SCFace database, we utilized off-the-shelf MTCNN [10] model to detect the faces. As there is not severe variation in pose we did not align the faces in this database for our experiments. Figure 4.5 shows the face detection and alignment step for a probe face of ICB-RW database.

4.2.3 Feature extraction

In the feature extraction researchers usually use the output of the activation layer of one or many layers of deep learning models as learned features. In this work, we used the 2048 dimensional output of the final activation layer, next to the classification layer as our deep face representations in models 1-4, and 512 dimensional face embedding of the output of the final batch normalization layer in models 5-8. In other words, we removed the classification layer of the pre-trained models described in 4.2.1, and fed the cropped and resized faces in the databases into the models. Afterwards, extracted face representations are L2 normalized (feature-wise) and saved as deep face representation as shown in Figure 4.6.

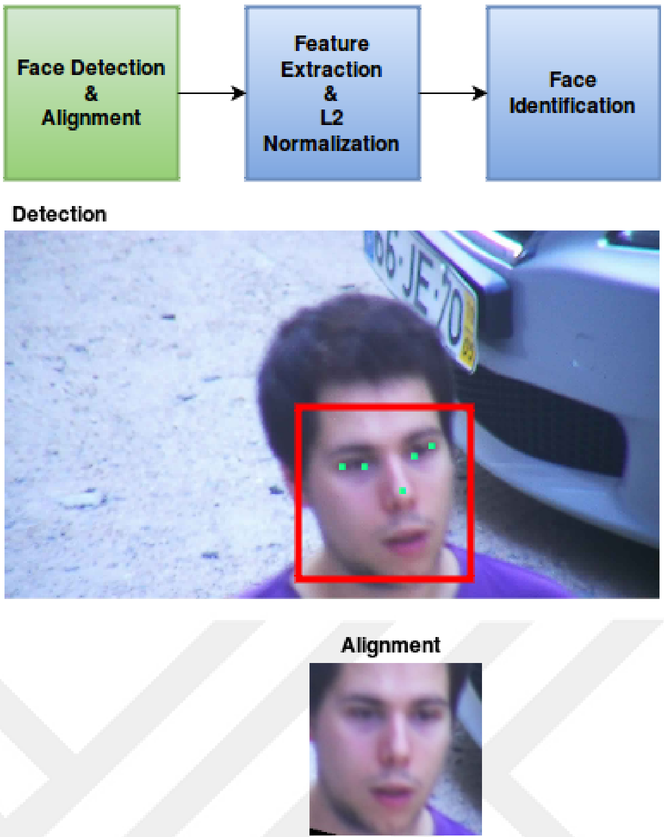

Face Detection & Alignment
Feature Extraction & L2 Normalization
Face Identification
Detection
Alignment

Figure 4.5 : Detection and alignment step of a subject from ICB-RW 2016 database.

4.2.4 Increasing the amount of information in probe face images

To increase the information in the cropped probe face images, we extended the face bonding boxes. In [34], it has been shown that this significantly improve the performance. In this study, we also expect this adjustment to contribute positively to the performance of Low Resolution Face Recognition (LRFR) due to two main reasons. Firstly, the low resolution face images have limited information and enlarging the bounding boxes would allow to include more information, for example about the shape of the face, etc. Secondly, the up-sampling factor would decrease and resulted in less degradation in probe face images. Since the input size of the face images to the deep learning models are relatively high (224×224 or 112×112 pixels in our models), this requires up-sampling of the low resolution face images with a large scaling factor. A larger crop of the face region would decrease the used scaling factor. The cropped faces using MTCNN [10] model and extended boxes are shown in Fig. 4.7. We enlarge the bounding boxes using six different extension factors (1.1, 1.2, 1.25, 1.3, 1.35, 1.40)

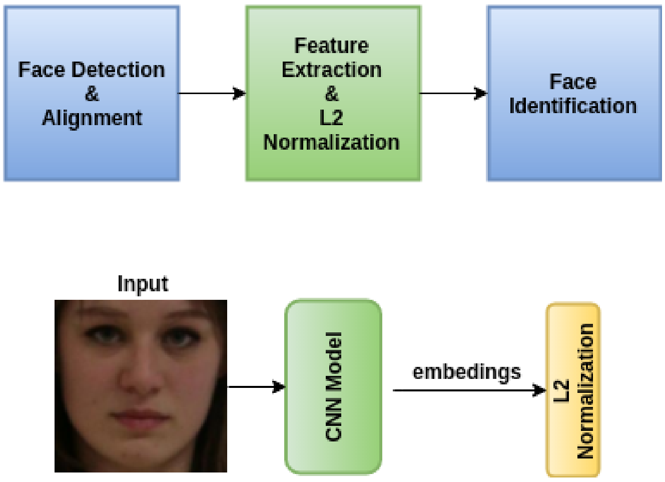

Face Detection & Alignment
Feature Extraction & L2 Normalization
Face Identification
Input
CNN Model
embedings
L2 Normalization

Figure 4.6 : Features of each face is extracted using pre-trained models and L2 normalized before inputting to the classifier.

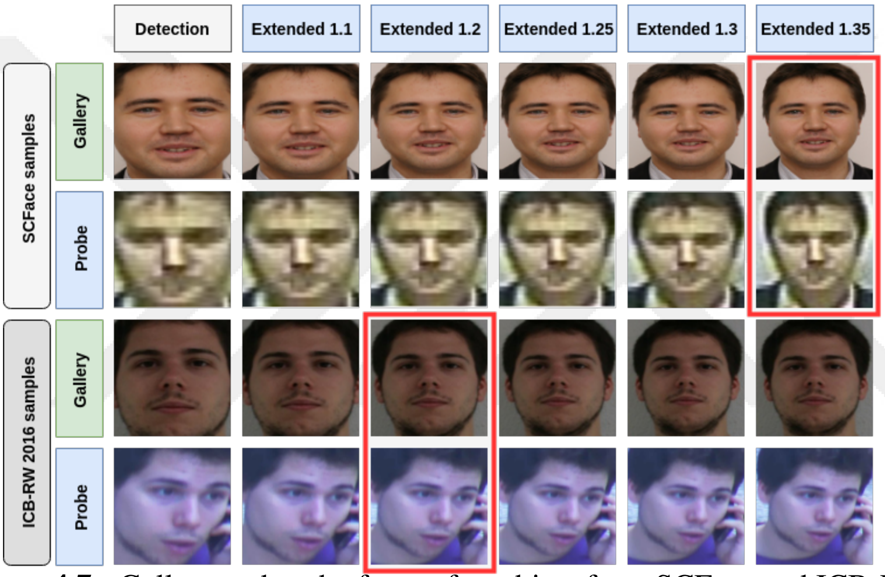


Figure 4.7 : Gallery and probe faces of a subject from SCFace and ICB-RW benchmarks. The faces are cropped with bounding boxes detected by MTCNN [10] and different extension factors.

4.2.5 Resolution matching between the gallery and probe face images

As there are significant resolution variation between the gallery and probe images in LRFR, the features extracted from LR face images in the probe set and HR face images in the gallery set potentially have higher intra-class distance than inter-class distance. We propose that matching the resolution of the gallery face images with the probe face images, intuitively, we would minimize the intra-class distance. This is done by down-sampling and then up-sampling the gallery images to the input size of the DCNNs models before feeding them to the models. In this study, we down-sampled

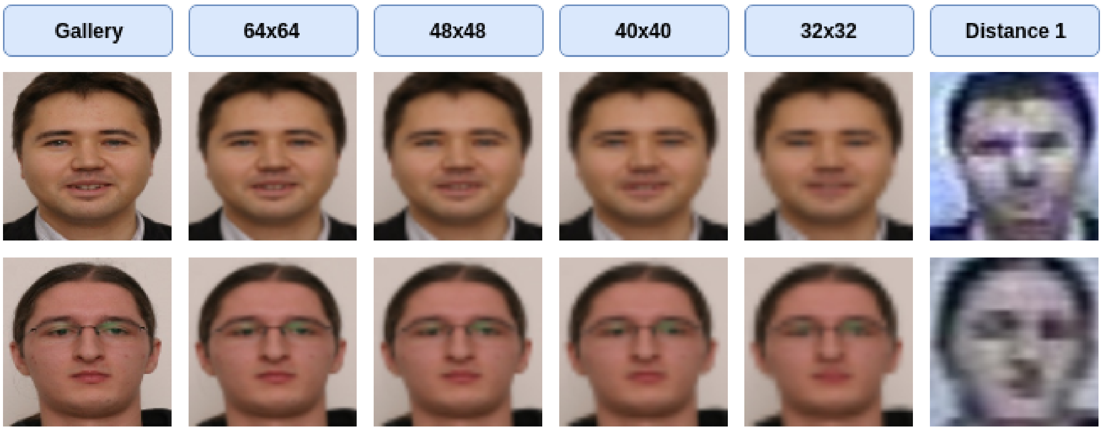


Figure 4.8 : Gallery faces of two subjects in SCFace cropped with 1.35 extended bounding boxes. **Column 1**: the original resolution. **Columns 2-5**: down-sampled images. **Column 6**: distance 1 probe faces.

the cropped gallery faces into five different resolutions (24×24, 32×32, 40×40, 48×48, 64×64) and up-sampled them into the input size of the DCNNs (112×112 or 224×224 depending on the architecture. Fig. 4.8 shows cropped faces of two subjects from SCFace [2] with original resolution (column 1), down-sampled (column 2-5), and distance 1 probe faces (column 6).

4.2.6 Face identification

To do face identification, we used Nearest Neighbor classifier with correlation distance 4.1 as the metric. In particular, extracted features of all the faces in the train (gallery) set are compared to features of all the faces in the test (probe) set to find the matches. A matrix with the size of $G \times P$ for correlation distances are calculated for every experiment, where G equals to the number of subjects in the gallery set and P is the number of the images in the probe set. Rank-1, Rank-5 Identification Rate and Area Under the Curve of Cumulative Match Curve is calculated and reported as the evaluation metric to compare the results with the previous works on ICB-RW database [1]. Also, Rank-1 IR is reported for SCFace database [2] as evaluation metric.

$$Corr.distance(u,v) = 1 - \frac{(u-\bar{u}).(v-\bar{v})}{\|(u-\bar{u})\|_2 \|(v-\bar{v})\|_2} \tag{4.1}$$

Where u, v are face feature vectors and $\bar{u}, \bar{v}$ are mean of face feature vectors.

5. EXPERIMENTAL RESULTS

In this chapter, experimental results on ICB-RW [1] and SCFace [2] databases are reported. ICB-RW database contains probe face images which have variation in pose, expression, illumination, motion-blur, out-of-focus and occlusions. Furthermore, SCFace [2] database encompasses low resolution face images. Thus, these databases characterize real-world surveillance scenarios and can be utilized to study the robustness of the face embeddings, extracted from deep learning models trained on large-scale face databases, e.g. VGGFace2 [14], MS-Celeb-1M [15] (described in 4.2.1), against the aforementioned variation under mismatched conditions. In section 5.1, we used face embedding of four deep learning models (models 1-4) to identify faces in the probe set of ICB-RW database and reported Rank-1, Rank-5, and Area Under the Curve of Cumulative Match Curve. In section 5.2, face features of eight models listed in 4.2.1 are used for face identification and Rank-1 IR is reported on SCFace database.

5.1 Experimental Results on ICB-RW

The ICB-RW database have 90 subjects, one gallery image, and 5 probe images per subject. First, we used bounding-box provided in the database to crop faces and then aligned faces. Afterwards, we leveraged four pre-trained models to extract 2048D feature vectors, namely, VF2-ResNet, VF2-ft-ResNet, VF2-SENet, VF2-ft-SENet. As a final step, we found the list of nearest face embedding of each train face to face embedding of every test images. A 90×450 correlation score matrix is calculated. We also proposed an ensemble of four models (as shown in Figure 5.1), in which face embedding extracted from four models are concatenated and given as input to classifier.

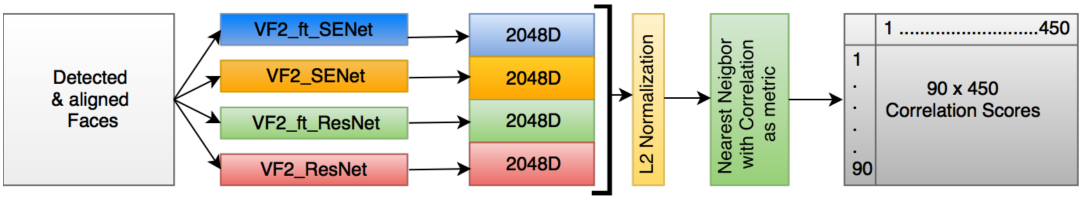


Figure 5.1 : Pipeline of the ensemble model for face identification.

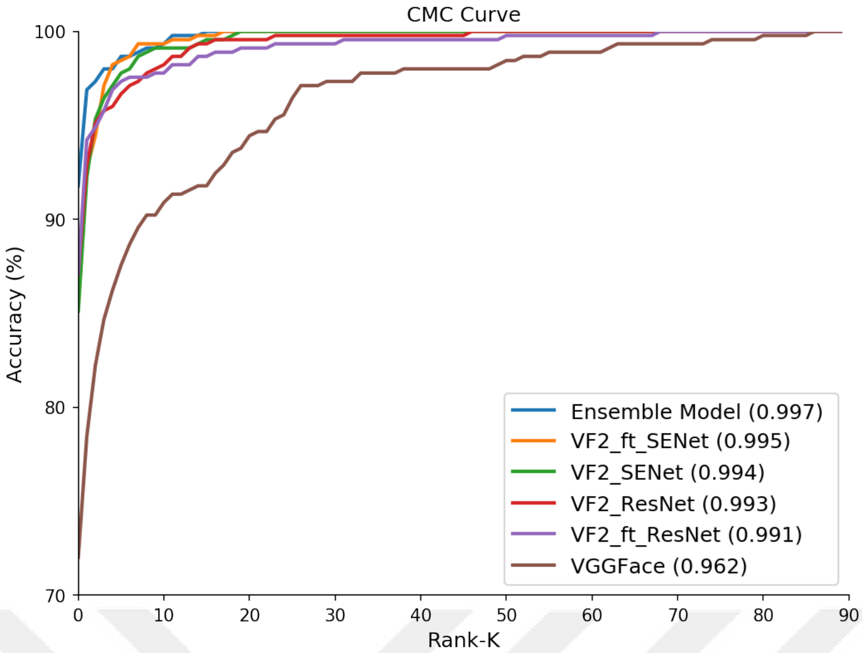


Figure 5.2 : Area Under the Curve of Cumulative Match Curve for Face Identification Results of ICB-RW database [1].

The Rank-1 and Rank-5 IR, and AUC of CMC are selected as evaluation metric. The Rank-G list of identification rates for the nearest face embedding of the gallery (train) to each face embedging of the probe (test) faces are listed, and the percentage of correct identification rate for probe faces are calculated with different Rank-G values to determin the CMC curve. The ensemble model with 8192 dimensional feature representation extracted from 4 models, achieved 91.78% 98.0% Rank-1, and Rank-5 IR respectively. The CMC is 0.997 for the ensemble model, which significantly improved the results in [1]. We achieved 87.11% Rank-1 accuracy, by using the single models of VF2-ft-ResNet and VF2-ResNet in feature extraction step. Also, the method which used VF2-ft-SENet in feature extraction step achieved highest Rank-5 and AUC of CMC, 98.22% and 0.995 respectively. Table 5.1 shows our empirical results in comparison with the results of using VGGFace model [25] in the feature extraction step. Our proposed models significantly increased the results in ICB-RW [1]. The Figure 5.2 depicts the AUC of CMC plot for the proposed models.

Analyzing 37 (out of 450) miss-identified subjects with ensemble models shows that those subjects either wear sunglasses or have strong pose variation (as shown in Figure 5.3). Figure 5.4 depicts three probe faces that have not been identified correctly by any of the models.

Table 5.1 : Identification rates and CMC with ground-truth bounding boxes.

Model	Rank-1 (%)	Rank-5 (%)	CMC
Ensemble	**91.78**	98.00	**0.997**
VF2-ft-SENet	85.33	**98.22**	0.995
VF2-SENet	85.11	97.11	0.994
VF2-ResNet	87.11	96.00	0.993
VF2-ft-ResNet	87.11	96.89	0.991
Ghaleb et al. [22]	71.7	86.5	0.962

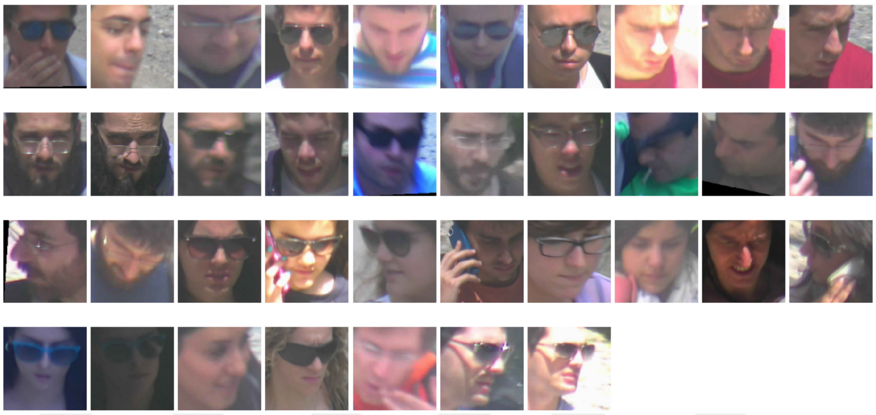


Figure 5.3 : Thirty-seven probe faces from ICB-RW database that are miss-identified by proposed ensemble model.

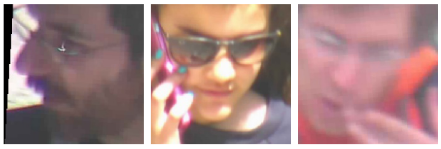


Figure 5.4 : Three probe faces from ICB-RW database that are not identified correctly by any proposed model.

5.1.1 Extended bounding boxes and down-sampled gallery face images

We also evaluated the performance of the proposed DCNNs on ICB-RW with increasing the information in the probe face images (as described in 4.2.4) and matching the resolution between the gallery and the probe face images (as described in 4.2.5). As the resolution of the probe images in ICB-RW [1] benchmark are high (1920 × 1080), enlarging the bounding boxes and decreasing the resolution of the gallery face images result in only modest improvements over the results using the bounding boxes detected by MTCNN [10]. The highest results with face images detected by MTCNN [10] are achieved with 1.2 extension factor and 64×64 down-sampling resolution. However, in overall the highest results are achieved using the ground-truth bounding boxes. The results using MTCNN bounding boxes, 1.2 extension factor and 64 × 64 down-sampling resolution are reported in Table 5.2.

Table 5.2 : Rank-1, Rank-5 IR and CMC on ICB-RW benchmark [1] using MTCNN bounding boxes. The extension (EXT) and down-sampling (DOWN) factors are 1.2 and 64 × 64, respectively.

Model	Rank-1	Rank-5	CMC
Ensemble	87.33	96.00	**0.983**
Ensemble-EXT	69.46	**97.85**	0.982
Ensemble-EXT-DOWN	**91.33**	95.33	0.982
VF2-ResNet-EXT-DOWN	85.11	94.44	0.980
VF2-SENet-EXT-DOWN	85.33	95.11	0.980
VF2-ft-ResNet-EXT-DOWN	85.33	93.56	0.977
VF2-ft-SENet-EXT-DOWN	83.78	94.44	0.979
Ghaleb et al. [22]	71.7	86.5	0.962

5.2 Experimental Results on SCFace

Having observed the significant improvement on ICB-RW database in comparison to previous reports [1], we quantitatively study the face embedding extracted from eight described pre-trained models in 4.2.1 for face identification on SCFace database [2] which encompasses the the probe images captured from 5 surveillance cameras, in three different resolutions, e.g. distance 1 (4.20m), distance 2 (2.60m), and distance 3 (1.00m). In section 5.2.1, we used the bounding boxes detected using off-the-shelf MTCNN [10] model to crop the faces in the gallery and probe sets and reported the Rank-1 face Identification Rates for eight models, as in 4.2.1. In 5.2.2, we further experimented with the extended bounding boxes and reported the results. Finally, we investigate the effect of down-sampling the gallery face images in the face identification results in section 5.2.3.

5.2.1 MTCNN bounding boxes

Following DCR [23], we randomly selected 80 subjects out of 130 subjects in SCFace database [2] and reported the mean and standard deviation of 20 runs for each model as feature extractor. In contrast to DCR [23], we did not used remaining 50 subjects for fine-tuning. Following the same steps as in our experiments on ICB-RW database, we detected the faces using off-the-shelf MTCNN [10] model, and then employed eight deep learning models (explained in 4.2.1) in feature extraction step to extract the face embeddings for every faces in the database, and used nearest neighbor classifier with correlation distance as similarity measurement to classify the probe faces. We reported the Rank-1 Identification Rates of eight single model and an ensemble of four models (models 1-4) in feature extraction step. Table 5.3 demonstrates the Rank-1 Identification Rate results for the prob face images captured in three distinct distances (distance 1,2, and 3, with 4.20, 2.60, 1.00 meters distance from the surveillance cameras). Our ensemble model achieved higher Rank-1 IR than DCR [23] for distance 2 and distance 3. Also, our proposed four single models obtained higher Rank-1 IR than results in DCR [23], however, distance 1 results fell behind the state-of-the-art results [23] by a large margin. We observed that features extracted from our models can not be generalized to probe images with very low resolution. Moreover, we observed that models trained on VGGFace2 [14] have better generalization on low resolution images (distance 1, distance 2) in comparison to the same models trained on MS-Celeb-1M [15].

Table 5.3 : Rank-1 IR mean and std. for the proposed models (20 runs). MTCNN bounding boxes are used in these experiments.

Model	d1 (4.20 m)	d2 (2.60 m)	d3 (1.00 m)
Ensemble	54.03 ±1.72	**94.35** ±1.01	**99.37** ±0.32
VF2-ResNet	47.41 ±1.93	92.69 ±1.03	98.53 ±0.47
VF2-SENet	47.84 ±2.23	91.91 ±1.04	98.47 ±0.54
VF2-ft-SENet	42.95 ±1.88	88.54 ±0.98	98.38 ±0.50
VF2-ft-ResNet	38.70 ±2.12	89.30 ±1.64	97.65 ±0.74
LResNet100E-IR	32.07 ±1.46	85.90 ±1.90	98.24 ±0.56
VF2-ft-LResNet100E-IR	42.36 ±1.68	88.56 ±1.60	96.61 ±0.55
LResNet50E-IR	17.23 ±1.85	63.49 ±2.73	88.21 ±1.14
VF2-ft-LResNet50E-IR	25.44 ±1.56	73.83 ±2.28	87.30 ±1.69
DCR [23]	**73.3**	93.5	98.00
LDMDS [29]	62.7	70.7	65.5

5.2.2 Extended bounding boxes

In these set of experiments, we extend the bonding boxes detected with MTCNN [10] model and used the extended boxes to crop the faces. The rest of the experimental setup is as in 5.2.1. Eight models are used to extract face features and face identification is done with nearest neighbor classifier with correlation distance as metric. Finally, we reported the mean and standard deviation of Rank-1 IR for 20 runs of each experiment. We conducted our experiments with six different extension factors (1.1, 1.2, 1.25, 1.3, 1.35, 1.40) to extend the bounding boxes and reported the highest results which are achieved with 1.35. Table 5.4 listed the Rank-1 IR for the models with and without extending the bounding boxes. As it can be seen from the results, extending the bounding boxes improve the results consistently. The probe face images in SCFace [2] contain limited information due to low resolution. Thus, extending the bounding boxes help the model to learn better face representations for two main reasons. First, the input size of the DCNNs are relatively high (112 × 112 or 224 × 224 in our cases) and cropped faces must be up-sampled which degrade the face information. Enlarging he bounding boxes, decrease the up-sampling factor which results in less deterioration. Second, extending the bounding boxes increase the information to be presented by face embeddings, e.g. shape of the faces and etc. These two reasons improve the performance of the models.

Table 5.4 : Experimental results on SCFace with and without enlarged bounding boxes. Here, the extension (EXT) factor is 1.35.

Model	d1 (4.20 m)	d2 (2.60 m)	d3 (1.00 m)
Ensemble	54.03 ±1.72	94.35 ±1.01	99.37 ±0.32
Ensemble-EXT	69.46 ± 1.83	**97.85** ± 0.50	99.68 ± 0.25
VF2-ResNet	47.41 ±1.93	92.69 ±1.03	98.53 ±0.47
VF2-ResNet-EXT	60.13 ± 2.34	95.65 ± 0.87	99.34 ± 0.34
VF2-SENet	47.83 ±1.52	91.31 ±1.35	98.30 ±0.50
VF2-SENet-EXT	57.91 ± 2.01	95.17 ± 0.82	98.41 ± 0.73
VF2-ft-ResNet	39.08 ±2.06	89.60 ±1.30	97.61 ±0.91
VF2-ft-ResNet-EXT	56.64 ± 2.22	95.46 ± 0.85	99.06 ± 0.46
VF2-ft-SENet	42.21 ±1.93	88.51 ±1.19	97.85 ±0.57
VF2-ft-SENet-EXT	58.46 ± 1.53	94.86 ± 0.80	98.85 ± 0.44
LResNet50E-IR	17.23 ±1.85	63.49 ±2.73	88.21 ±1.14
LResNet50E-IR-EXT	27.49 ± 2.07	80.38 ± 0.96	94.76 ± 1.00
VF2-ft-LResNet50E-IR	25.44 ±1.56	73.83 ±2.28	87.30 ±1.69
VF2-ft-LResNet50E-IR-EXT	51.23 ± 2.08	87.64 ± 2.17	94.44 ± 1.07
LResNet100E-IR	32.07 ±1.46	85.90 ±1.90	98.24 ±0.56
LResNet100E-IR-EXT	60.34 ± 2.25	97.16 ± 0.70	**100 ± 0.00**
VF2-ft-LResNet100E-IR	42.36 ±1.68	88.56 ±1.60	96.61 ±0.55
VF2-ft-LResNet100E-IR-EXT	67.65 ± 2.06	96.71 ± 0.86	99.43 ± 0.31
DCR [23]	**73.3**	93.5	98.00
LDMDS [29]	62.7	70.7	65.5

Although, the results for distance 1 improved significantly, nonetheless, the results are less than the state-of-the-art results. Thus, we used down-scaling to blur the gallery faces before feature extraction and hypothesize that might help to minimize the distance between face descriptors of high quality images and low quality images. In 5.2.3, we reported the results for our experiments with down-scaled gallery faces.

5.2.3 Down-sampled gallery face images

In this work, we experimented with five different down-sampling factors (24×24, 32×32, 40×40, 48×48, 64×64) and reported the highest Rank-1 IR for the eight models which obtained by 64×64 down-sampling factor. For the bounding boxes we used 1.35 extension factor. We conducted our experiments with 24×24, 32×32, 40×40, 48×48, 64×64 down-sampling factors and reported the highest Rank-1 IR which are achieved with down-sampling the cropped gallery faces into 40×40 and up-sampling them into 224 in models 1-4, and 112×112 in models 5-8. Table 5.5 shows the results of our experiments with MTCNN bounding boxes, Extended bounding boxes (EXT), and Down-sampled (DOWN) gallery faces.

The empirical results with 8 different deep learning models and an ensemble of models 1-4 in feature extraction step shows that matching the resolution between the gallery and the probe face images before feature extraction has a significant effect on improving the Rank-1 IR on SCFace [2]. As it can be seen in Table 5.5, face embedding extracted from VF2-ft-LResNet100E-IR model has significantly improved the previously reported state-of-the-art results [23] by a large margin of 3.64%, 3.95%, and 1.36% for distance 1, distance 2, and distance 3 respectively. Rank-1 Identification Rate on 130 subjects of SCFace [2] database are reported in Appendix A to help future research. Rank-1 IR for 130 subjects of SCFace [2] are 72.62%, 96.62%, and 99.08% for distance 1, 2, and 3 using VF2-ft-LResNet100E-IR model in feature extraction step.

Table 5.5 : Experimental results for SCFace with MTCNN bounding boxes, Extended bounding boxes (EXT), and Down-sampled (DOWN) gallery faces. Here, the extension factor is 1.35 and down-sampling factor is 40×40.

Model	d1 (4.20 m)	d2 (2.60 m)	d3 (1.00 m)
Ensemble	54.03 ±1.72	94.35 ±1.01	99.37 ±0.32
Ensemble-EXT	69.46 ± 1.83	97.85 ± 0.50	99.68 ± 0.25
Ensemble-EXT-DOWN	75.47 ± 1.94	97.32 ± 0.69	98.44 ± 0.75
VF2-ResNet	47.41 ±1.93	92.69 ±1.03	98.53 ±0.47
VF2-ResNet-EXT	60.13 ± 2.34	95.65 ± 0.87	99.34 ± 0.34
VF2-ResNet-EXT-DOWN	64.89 ± 2.03	94.81 ± 0.96	95.95 ± 0.80
VF2-SENet	47.83 ±1.52	91.31 ±1.35	98.30 ±0.50
VF2-SENet-EXT	57.91 ± 2.01	95.17 ± 0.82	98.41 ± 0.73
VF2-SENet-EXT-DOWN	64.89 ± 2.03	94.81 ± 0.96	95.95 ± 0.80
VF2-ft-ResNet	39.08 ±2.06	89.60 ±1.30	97.61 ±0.91
VF2-ft-ResNet-EXT	56.64 ± 2.22	95.46 ± 0.85	99.06 ± 0.46
VF2-ft-ResNet-EXT-DOWN	64.89 ± 2.03	94.81 ± 0.96	95.95 ± 0.80
VF2-ft-SENet	42.21 ±1.93	88.51 ±1.19	97.85 ±0.57
VF2-ft-SENet-EXT	58.46 ± 1.53	94.86 ± 0.80	98.85 ± 0.44
VF2-ft-SENet-EXT-DOWN	64.89 ± 2.03	94.81 ± 0.96	95.95 ± 0.80
LResNet50E-IR	17.23 ±1.85	63.49 ±2.73	88.21 ±1.14
LResNet50E-IR-EXT	27.49 ± 2.07	80.38 ± 0.96	94.76 ± 1.00
LResNet50E-IR-EXT-DOWN	36.70 ± 1.93	85.11 ± 1.33	95.40 ± 0.80
VF2-ft-LResNet50E-IR	25.44 ±1.56	73.83 ±2.28	87.30 ±1.69
VF2-ft-LResNet50E-IR-EXT	51.23 ± 2.08	87.64 ± 2.17	94.44 ± 1.07
VF2-ft-LResNet50E-IR-EXT-DOWN	61.40 ± 2.69	89.31 ± 1.41	94.59 ± 1.16
LResNet100E-IR	32.07 ±1.46	85.90 ±1.90	98.24 ±0.56
LResNet100E-IR-EXT	60.34 ± 2.25	97.16 ± 0.70	100 ± 0.00
LResNet100E-IR-EXT-DOWN	70.25 ± 2.22	**98.41** ± 0.92	**100.00** ± 0.00
VF2-ft-LResNet100E-IR	42.36 ±1.68	88.56 ±1.60	96.61 ±0.55
VF2-ft-LResNet100E-IR-EXT	67.65 ± 2.06	96.71 ± 0.86	99.43 ± 0.31
VF2-ft-LResNet100E-IR-EXT-DOWN	**76.94** ± 1.98	97.45 ± 0.76	99.36 ± 0.26
DCR [23]	73.3	93.5	98.00
LDMDS [29]	62.7	70.7	65.5

6. CONCLUSIONS AND RECOMMENDATIONS

In this thesis, we proposed deep learning based face embedding for face identification in low resolution faces. We conducted experiments on challenging surveillance databases, e.g. ICB-RW [1] and [2], using eight state-of-the-art deep learning architectures and an ensemble of four models as described in 4.2.1. We utilized very large scale face database, VGGFace2 [14] which consist of 3.31 M images of 9,131 identities collected across ages and poses.

We explored three factors that would improve low resolution face recognition performance. These factors are, the variation in appearance and resolution of the training database, the probe and the gallery face images resolution matching, and the information to be included in the cropped probe face images. Our experimental results show that these factors are effective an contribute to improve the performance of deep learning models. Our main contributions can be summarized as follows: 1) The robustness of four state-of-the-art DCNN models, namely, ResNet [7], SENet [8], LResNet50E-IR [9], LResNet100E-IR [9] are explored in learning face embeddings, and two large scale face databases, VGGFace2 [14] and MS-Celeb-1M [15] are utilized, to train and fine-tune them.

We found that the variation in appearance and resolution of the training database is an important factor in learning robust face features, 2) Different sizes of the face crops are used for assessing the effect of included information in the cropped face images. We observed that extending the bounding boxes to crop face images have positive effect on the performance, 3) The effect of the resolution matching between the gallery and probe images are investigated. We down-sampled the high resolution gallery face images to match the resolution of the gallery and the probe face images. Our approach has less computational cost than [31] which used super-resolving the low resolution face images. We observed that down-sampling the gallery and the probe face images increases the performance.

Our empirical results are current state-of-the-art on ICB-RW [1] and SCFace [2] benchmarks. Our Ensemble model achieved 91.78%, 98.00%, 0.997, Rank-1, Rank-5 IR and AUC of CMC respectively on ICB-RW [1] using the ground-truth bounding boxes. On SCFace benchmark [2], our single model, VF2-ft-LResNet100E-IR with extension factor of 1.35 and down-sampling factor of 40×40, achieved $76.94\% \pm 1.98$, $97.48\% \pm 0.76$, $99.36\% \pm 0.26$ on distance 1, 2, and 3 of SCFace benchmark [2] respectively. In addition, we achieved 72.62%, 96.62%, and 99.08% for distance 1, 2, and 3 using VF2-ft-LResNet100E-IR model with the same factors.

REFERENCES

[1] **Neves, J. and Proença, H.** (2016). ICB-RW 2016: International challenge on biometric recognition in the wild, *Biometrics (ICB), 2016 International Conference on*, IEEE, pp.1–6.

[2] **Grgic, M.**, **Delac, K. and Grgic, S.** (2011). SCface–surveillance cameras face database, *Multimedia tools and applications*, *51*(3), 863–879.

[3] **Goodfellow, I.**, **Bengio, Y.**, **Courville, A. and Bengio, Y.** (2016). *Deep learning*, volume 1, MIT press Cambridge.

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

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

[6] **Schroff, F.**, **Kalenichenko, D. and Philbin, J.** (2015). Facenet: A unified embedding for face recognition and clustering, *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp.815–823.

[7] **He, K.**, **Zhang, X.**, **Ren, S. and Sun, J.** (2016). Deep residual learning for image recognition, *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp.770–778.

[8] **Hu, J.**, **Shen, L. and Sun, G.** (2017). Squeeze-and-excitation networks, *arXiv preprint arXiv:1709.01507*.

[9] **Deng, J.**, **Guo, J. and Zafeiriou, S.** (2018). ArcFace: Additive Angular Margin Loss for Deep Face Recognition, *arXiv preprint arXiv:1801.07698*.

[10] **Zhang, K.**, **Zhang, Z.**, **Li, Z. and Qiao, Y.** (2016). Joint face detection and alignment using multitask cascaded convolutional networks, *IEEE Signal Processing Letters*, *23*(10), 1499–1503.

[11] **Simonyan, K.**, **Parkhi, O.M.**, **Vedaldi, A. and Zisserman, A.** (2013). Fisher Vector Faces in the Wild., *BMVC*, volume 2, p. 4.

[12] **Simonyan, K. and Zisserman, A.** (2014). Very deep convolutional networks for large-scale image recognition, *arXiv preprint arXiv:1409.1556*.

[13] **Szegedy, C.**, **Liu, W.**, **Jia, Y.**, **Sermanet, P.**, **Reed, S.**, **Anguelov, D.**, **Erhan, D.**, **Vanhoucke, V.**, **Rabinovich, A.** *et al.* (2015). Going deeper with convolutions, Cvpr.

[14] **Cao, Q.**, **Shen, L.**, **Xie, W.**, **Parkhi, O.M. and Zisserman, A.** (2018). Vggface2: A dataset for recognising faces across pose and age, *Automatic Face & Gesture Recognition (FG 2018), 2018 13th IEEE International Conference on*, IEEE, pp.67–74.

[15] **Guo, Y.**, **Zhang, L.**, **Hu, Y.**, **He, X. and Gao, J.** (2016). Ms-celeb-1m: A dataset and benchmark for large-scale face recognition, *European Conference on Computer Vision*, Springer, pp.87–102.

[16] **Huang, G.B.**, **Ramesh, M.**, **Berg, T. and Learned-Miller, E.** (2007). Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments, **Technical Report07-49**, University of Massachusetts, Amherst.

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

[18] **Sun, Y.**, **Liang, D.**, **Wang, X. and Tang, X.** (2015). Deepid3: Face recognition with very deep neural networks, *arXiv preprint arXiv:1502.00873*.

[19] **Liu, W.**, **Wen, Y.**, **Yu, Z.**, **Li, M.**, **Raj, B. and Song, L.** (2017). Sphereface: Deep hypersphere embedding for face recognition, *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1.

[20] **Kazemi, V. and Josephine, S.** (2014). One millisecond face alignment with an ensemble of regression trees, *27th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, United States, 23 June 2014 through 28 June 2014*, IEEE Computer Society, pp.1867–1874.

[21] **King, D.E.** (2009). Dlib-ml: A machine learning toolkit, *Journal of Machine Learning Research*, *10*(Jul), 1755–1758.

[22] **Ghaleb, E.**, **Ozbulak, G.**, **Gao, H. and Ekenel, H.K.**, (2018), Deep representation and score normalization for face recognition under mismatched conditions.

[23] **Lu, Z.**, **Jiang, X. and Kot, A.C.** (2018). Deep Coupled ResNet for Low-Resolution Face Recognition, *IEEE Signal Processing Letters*.

[24] **Taigman, Y.**, **Yang, M.**, **Ranzato, M. and Wolf, L.** (2014). Deepface: Closing the gap to human-level performance in face verification, *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp.1701–1708.

[25] **Parkhi, O.M.**, **Vedaldi, A.**, **Zisserman, A.** *et al.* (2015). Deep Face Recognition., *BMVC*, volume 1, p. 6.

[26] **Yi, D.**, **Lei, Z.**, **Liao, S. and Li, S.Z.** (2014). Learning face representation from scratch, *arXiv preprint arXiv:1411.7923*.

[27] **Lee, S.H.**, **Choi, J.Y.**, **Ro, Y.M. and Plataniotis, K.N.** (2012). Local color vector binary patterns from multichannel face images for face recognition, *IEEE Transactions on Image Processing*, *21*(4), 2347–2353.

[28] **De Marsico, M.**, **Nappi, M.**, **Riccio, D. and Wechsler, H.** (2013). Robust face recognition for uncontrolled pose and illumination changes, *IEEE transactions on systems, man, and cybernetics: systems*, *43*(1), 149–163.

[29] **Yang, F.**, **Yang, W.**, **Gao, R. and Liao, Q.** (2018). Discriminative multidimensional scaling for low-resolution face recognition, *IEEE Signal Processing Letters*, *25*(3), 388–392.

[30] **Yu, X.**, **Fernando, B.**, **Hartley, R. and Porikli, F.** (2018). Super-Resolving Very Low-Resolution Face Images with Supplementary Attributes, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.908–917.

[31] **Wang, Z.**, **Chang, S.**, **Yang, Y.**, **Liu, D. and Huang, T.S.** (2016). Studying very low resolution recognition using deep networks, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.4792–4800.

[32] **Pan, S.J. and Yang, Q.** (2010). A survey on transfer learning, *IEEE Transactions on knowledge and data engineering*, *22*(10), 1345–1359.

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

[34] **Mehdipour Ghazi, M. and Kemal Ekenel, H.** (2016). A comprehensive analysis of deep learning based representation for face recognition, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp.34–41.

APPENDICES

APPENDIX A.1 : Experimental results for 130 subjects in SCFace

APPENDIX A.1

Here, we report the results for 130 subjects of SCFace benchmark [2]. Rank-1 IR results are reported with MTCNN bounding boxes, Extended bounding boxes (EXT) with extension factor of 1.35, and Down-sampling (DOWN) factor of 40 × 40. Table A.1 shows the results for our experiments.

Table A.1 : Experimental results on SCFace [2] with MTCNN bounding boxes, Extended bounding boxes (EXT) with extension factor of 1.35, and Down-sampling (DOWN) factor of 40 × 40.

Model	d1 (4.20 m)	d2 (2.60 m)	d3 (1.00 m)
Ensemble	47.38	92.77	99.23
Ensemble-EXT	64.00	96.92	99.54
Ensemble-EXT-DOWN	70.77	96.62	97.69
VF2-ResNet	40.92	91.69	98.00
VF2-ResNet-EXT	53.54	94.46	99.08
VF2-ResNet-EXT-DOWN	59.38	94.62	98.46
VF2-SENet	41.85	89.54	97.69
VF2-SENet-EXT	51.85	94.15	97.54
VF2-SENe-EXT-DOWN)	57.38	94.77	96.77
VF2-ft-ResNet	33.08	86.92	96.62
VF2-ft-ResNet-EXT	50.77	94.00	98.62
VF2-ft-ResNet-EXT-DOWN	56.00	94.31	97.85
VF2-ft-SENet	35.69	86.00	97.23
VF2-ft-SENet-EXT	53.08	93.23	98.31
VF2-ft-SENe-EXT-DOWN)	58.62	93.85	99.08
LResNet50E-IR	14.31	58.46	86.46
LResNet50E-IR-EXT	23.85	79.23	94.31
LResNet50E-IR-EXT-DOWN	30.77	81.69	94.00
VF2-ft-LResNet50E-IR	20.46	70.00	84.46
VF2-ft-LResNet50E-IR-EXT	45.69	86.00	93.08
VF2-ft-LResNet50E-IR-EXT-DOWN	56.15	87.08	93.23
LResNet100E-IR	28.15	83.54	97.69
LResNet100E-IR-EXT	54.62	96.15	100
LResNet100E-IR-EXT-DOWN	64.77	**97.54**	**100.00**
VF2-ft-LResNet100E-IR	36.77	87.23	95.54
VF2-ft-LResNet100E-IR-EXT	61.54	88.41	94.64
VF2-ft-LResNet100E-IR-EXT-DOWN	**72.62**	96.62	99.08

YÖK

CURRICULUM VITAE

Name Surname: Omid Abdollahi Aghdam

Place and Date of Birth: Marand, Iran - 12/07/1982

E-Mail: abdollahi@itu.edu.tr

EDUCATION:

- **B.Sc.:** 2014, Payame Noor University of Tabriz, Faculty of Information Technology Engineering, Information Technology Engineering Department
- **M.Sc.:** 2018, Istanbul Technical University, Faculty of Computer and Informatics Engineering, Department of Computer Engineering

PUBLICATIONS, PRESENTATIONS AND PATENTS ON THE THESIS:

- Aghdam, O. A., & Ekenel, H. K. (2018, May). Robust deep learning features for face recognition under mismatched conditions. In 2018 26th Signal Processing and Communications Applications Conference (SIU). IEEE.