Robust Deep Learning Features for Face Recognition under Mismatched Conditions

Omid Abdollahi Aghdam, Hazım Kemal Ekenel
Department of Computer Engineering
Istanbul Technical University
Istanbul, Turkey
{abdollahi15, ekenel}@itu.edu.tr

***Abstract*—In this paper, we addressed the problem of face recognition under mismatched conditions. In the proposed system, for face representation, we leveraged the state-of-the-art deep learning models trained on the VGGFace2 dataset. More specifically, we used pretrained convolutional neural network models to extract 2048 dimensional feature vectors from face images of International Challenge on Biometric Recognition in the Wild dataset, shortly, ICB-RW 2016. In this challenge, the gallery images were collected under controlled, indoor studio settings, whereas probe images were acquired from outdoor surveillance cameras. For classification, we trained a nearest neighbor classifier using correlation as the distance metric. Experiments on the ICB-RW 2016 dataset have shown that the employed deep learning models that were trained on the VGGFace2 dataset provides superior performance. Even using a single model, compared to the ICB-RW 2016 winner system, around 15% absolute increase in Rank-1 correct classification rate has been achieved. Combining individual models at feature level has improved the performance further. The ensemble of four models achieved 91.8% Rank-1, 98.0% Rank-5 identification rate, and 0.997 Area Under the Curve of Cumulative Match Score on the probe set. The proposed method significantly outperforms the Rank-1, Rank-5 identification rates, and Area Under the Curve of Cumulative Match Score of the best approach at the ICB-RW 2016 challenge, which were 69.8%, 85.3%, and 0.954, respectively.**

***Keywords*—Biometric Recognition, Face Recognition, Deep Learning, Convolutional Neural Networks.**

I. Introduction

Biometric face identification for surveillance purposes has many applications, e.g, crime investigation, security systems, in which the features extracted from the face image of a target is compared to the features from all face images in the gallery set as shown in Figure 1. For an efficient comparison, a compact face representation is desired. Conventionally, the engineered techniques such as Fisher Vectors (FV) were used to extract these features [1]. However, the advent of deep learning architectures [4]–[6], [11] and large-scale datasets, such as Labeled Faces in the Wild (LFW) [7] and YouTube Faces (YTF) [8], have facilitated the research in the field of face recognition. Although, the recent advances in the field of face recognition have been significant [9], [10], [12], [13], most of the works are evaluated on LFW and YTF datasets, which are collected under matched conditions. In other words, both training and testing sets in LFW [7] are collected from the web,

978-1-5386-1501-0/18/$31.00

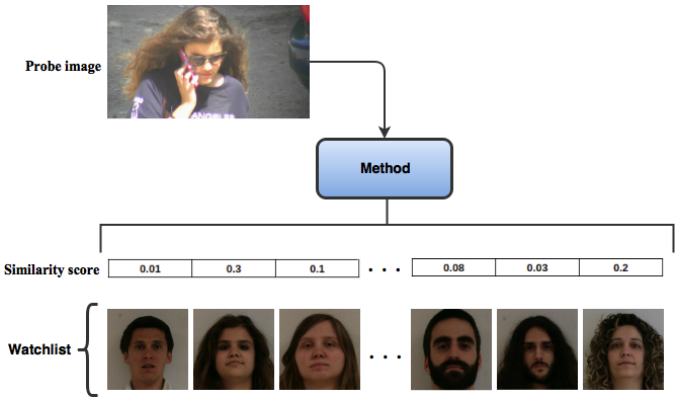

Figure 1: In the ICB-RW 2016 [3], probe/test image is compared to the watchlist/samples in the gallery set. The gallery images are collected under controlled, indoor studio settings. Probe images are acquired from outdoor surveillance cameras.

without motion-blur and in relatively high resolution, or in the case of YTF dataset contains videos recorded under similar conditions. Nonetheless, comparable results for biometric face recognition on real-world visual surveillance scenarios have not been achieved, yet.

Face recognition under matched conditions, where both training and testing images are from the similar domain, as in LFW [7] and YTF [8], is considered as solved, as the recent advances reported on LFW and YTF demonstrate. FaceNet [10] has been proposed as an end to end deep learning architecture based on Inception model [11] followed by L2 normalization and Triplet Loss. The model were trained on a very large-scale private dataset of 260M images. The proposed model achieved the record accuracy of 99.63% on LFW and 95.12% on YTF. Sun, et al. [12] used two VGG [6] architectures including Inception modules [11], and extracted features from 25 crops of different parts of each face per network. The extracted features were concatenated and dimension was reduced to 300 using Principal Component Analysis. Afterwards, a joint Bayesian model is learned for face recognition. Their proposed method achieved 99.54% accuracy on LFW. The SphereFace [13] approach introduced the Angular-Softmax loss and used ResNet architecture [5] to learn face embeddings in training phase. In the test, they applied nearest neighbor classifier with cosine similarity to identify the faces. The applied method achieved 99.42% accuracy on LFW and 95.0% on YTF.

Figure 2: Sample face images from gallery set (three images on the left) and probe set (five images on the right) of two subjects in the ICB-RW 2016 dataset. The gallery set includes left-side, frontal, right-side face images captured under controlled conditions. Probe set contains five images per subject that contain variations in pose, illumination, expression, motion-blur, and occlusion.

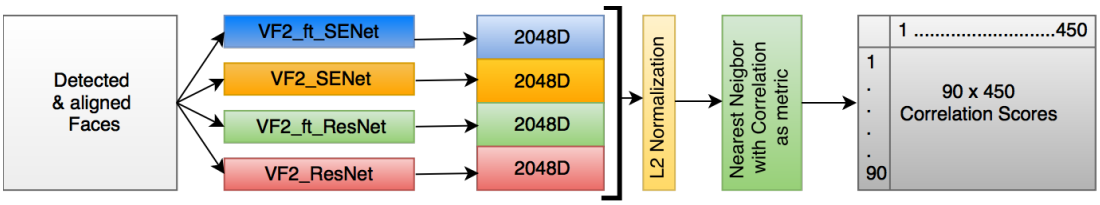

Figure 3: This diagram shows the pipeline for our ensemble model. For each face in the dataset, 2048D feature vectors are extracted per model, concatenated, and L2 normalized. Then, we trained nearest neighbor classifier with correlation as metric.

There are not many publicly available large-scale datasets to address the face recognition problem for surveillance scenarios. To the best of our knowledge, CoxFace [15], ChokePoint [14], and SCFace [16] are the most cited datasets for the surveillance scenarios in the literature. Despite the fact that these datasets are collected under the mismatched conditions, they do not encompass all the characteristic of the real surveillance scenarios, e.g, occlusion, strong motion-blur, pose, illumination, expressions, and focus. Thus, we examined the robustness of the deep face descriptors, learned with Convolutional Neural Networks (CNN), namely ResNet-50 [5] and SENet-50 [4], on the International Challenge on Biometric Recognition in the Wild dataset (ICB-RW 2016) [3], which includes all of the aforementioned variations present in the real-world surveillance scenarios. In the challenge paper [3], the highest performance was reported as 69.8% Rank-1, and 85.3% Rank-5 accuracy on the probe set, with 0.954 Area Under the Curve (AUC) of Cumulative Match Score curve (CMC). Our experiments, which made use of ResNet-50 [5] and SENet-50 [4] models trained on VGGFace2 dataset [2] as the feature extractor, markedly improved the previous results [3]. We consider that our proposed method owes its achievements to deeper CNN architectures and larger number of images in the VGGFace2 dataset (3.31 million images) [2], in which there are images of each subject in different poses and ages, enabling the models to learn a robust face representation.

In the surveillance scenarios, generally there is a single image per person in the gallery, usually a high resolution frontal or left/right-side face, on the other hand, probe images are captured with low resolution cameras, involving variations in pose, illumination, expression, focus, motion-blur, and occlusion as can be seen in Figure 2. Hence, learning robust face representation, invariant to stated image quality problems, is the most critical part of face identification pipeline. In our proposed method, we used face bounding boxes provided in the dataset and the implementation of the work by Kazemi et al. [20], using dlib library [17], to align faces. For learning face features, we leveraged deep learning models, trained on VGGFace2 dataset [2], to extract 2048 dimensional feature vectors of each image in the gallery and probe sets. Finally, we applied the nearest neighbor classifier with correlation metric to identify the probe faces. We calculated a $G \times P$ correlation score matrix for each model, where $G = 90$ and $P = 450$ are the number of images in the gallery —one image per person— and probe set —five images per person. Our best model, which was trained using 8192 dimensional face descriptors, $4 \times 2048D$ features of each image, extracted from four models, achieved 91.78% Rank-1, 98.0% Rank-5 identification rates (IR) and 0.997 AUC of CMC on the test set, significantly improving the best previously reported results [3] by a margin of 21.98%, 12.7%, and 0.045, respectively.

The rest of the paper is organized as follows. Datasets, employed methods, and deep learning models are explained in Section II. Experimental results are presented and discussed in Section III. Finally, in Section IV conclusions and future research directions are given.

II. Method

A. Dataset

The dataset consists of gallery and probe images of 90 individuals. The gallery images are photographed under controlled conditions, in which there are three images per person, i.e. high resolution frontal, left-side, and right-side face images. The probe images are captured outdoors and represent the characteristics of the surveillance systems, e.g. varying poses, illumination, expressions, motion-blur, and occlusion. In our experiments, we only used frontal images in the gallery (90

images) and five probe images per subject (450 images). In the International Challenge on Biometric Recognition in the Wild[1], competitors were asked to identify probe images between individuals in the watchlist. In Figure 2 sample cropped face images from the gallery and probe sets for two subjects are shown.

B. Models

The VGGFace2 dataset [2] is collected in six steps, i.e. name list selection, image downloading, face detection, automatic filtering by classification, near duplicate removal, and final automatic and manual filtering. This dataset contains 3.31 million images of 9,131 subjects across different poses and ages. Thereafter, ResNet-50 [5] and SENet-50 [4] architectures are used for training a face classification model. The models we have utilized in our experiments are publicly available[2] and were trained under following settings:

1) ResNet-50 is trained from scratch on VGGFace2, shortly VF2-ResNet.

2) ResNet-50 is learned on MS-Celeb-1M (MS1M) [19] and fine-tuned on VGGFace2 (VF2-ft-ResNet for short).

3) SENet-50 is learned from scratch on VGGFace2 (VF2-SENet).

4) SENet-50 is trained on MS1M and then fine-tuned on VGGFace2 (VF2-ft-SENet).

C. Face Detection and Alignment

Face detection and alignment are crucial steps for face identification, which are carried out to register the faces such that the facial landmarks in the aligned faces are fixed in a canonical space. Face alignment is a normalization step in face recognition, which should be done to aid the feature extraction step to produce more reliable representations for similarity calculation. In order to do a fair comparison with the previous methods reported in [3], we employed the ground-truth bounding boxes provided in the dataset for face detection as well as face alignment implementation of dlib [17] based on the paper of Kazemi et al. [20]. In Kazemi et al. paper [20], an ensemble of regression trees are used to detect facial landmarks, in our case, corners of the eyes, and bottom of the nose are detected, and the faces in the dataset are aligned using the detected landmarks. We also experimented with 68 facial landmarks detector based on [20] to detect the positions of landmarks for alignment. In all the experimental setups, the former method resulted in higher identification rates.

D. Feature Extraction

For feature extraction in our experiments, pretrained VF2-ResNet, VF2-ft-ResNet, VF2-SENet, and VF2-ft-SENet models, described in subsection II-B, were used. In particular, we removed the top layer (classification) of those models, resized the aligned gallery and probe faces into 224×224 pixels before feeding them to the models, extracting 2048 dimensional output of the final layer as deep face representation. Afterwards, we applied feature-wise L2 normalization and then fed the features into the classifier.

[1]http://icbrw.di.ubi.pt/

[2]http://www.robots.ox.ac.uk/∼vgg/data/vgg_face2/

Table I: Rank-1, Rank-5, and AUC of Cumulative Match Curve are reported for our experiments. For comparison with the proposed approach by Ekenel et al. reported in [3], we used VGGFace [21] in the feature extraction step. The proposed methods are ranked based on AUC of CMC. The best results for Rank-1, Rank-5, and AUC of CMC are in bold font.

Model	Rank-1 IR (%)	Rank-5 IR (%)	AUC (CMC curve)
Ensemble Model	**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
VGGFace	72.00	86.22	0.962

III. Experimental Results

In the ICB-RW 2016 [3], competitors were asked to find the Rank-K list of the most similar faces in the gallery set for the probe images. More specifically, given the 450 probe images, the task of face identification system is providing the Rank-K list of faces in the watchlist —from the most to the least similar. We exploited four models, mentioned in subsection II-B, to extract 2048 dimensional feature vectors from each cropped and aligned faces in the dataset and classified these features using the nearest neighbor classifier with correlation metric. We retrieved the list of nearest frontal faces in the watchlist for every probe images. The experiments resulted in $G \times P$ correlation score matrix, where $G = 90$ is the number of subjects in the gallery —one frontal face per subject— and $P = 450$ is the number of probe images — five images per subject. We also trained an ensemble model of four specified models, as illustrated in Figure 3, in which we concatenated 2048 dimensional feature vectors extracted using four models. Finally, we learned a nearest neighbor classifier with correlation metric to compare 8192 dimensional feature vectors of each probe and gallery face pairs. The performance of the proposed algorithms are measured by Rank-1, Rank-5 identification rates, and Area Under Curve (AUC) of the Cumulative Match Curve (CMC). For this purpose, Rank-K list of the most similar faces in the watchlist for every probe faces are calculated, and the percentage of correct identification for probe faces are determined with different Rank-K values to obtain the CMC curve.

Our proposed approach, ensemble of four models, achieved 91.78% Rank-1, 98.0% Rank-5, and 0.997 CMC, considerably improving the best reported results in [3]. The highest Rank-1 accuracy achieved using a single model is 87.11% using VF2-ft-ResNet and VF2-ResNet models in feature extraction step. Whereas, the highest Rank-5 and AUC of CMC is obtained using VF2-ft-SENet model as a feature extractor, 98.22% and 0.995 respectively. The results of the experiments are reported in Table I and compared to the results obtained with the model that used the pretrained VGGFace model [21] for feature extraction. The proposed models in our work have higher accuracies and AUC of CMC than reported results in ICB-RW 2016 [3]. The Figure 4 illustrates the percentage of the correctly identified probe faces against Rank-K matches in the watchlist.

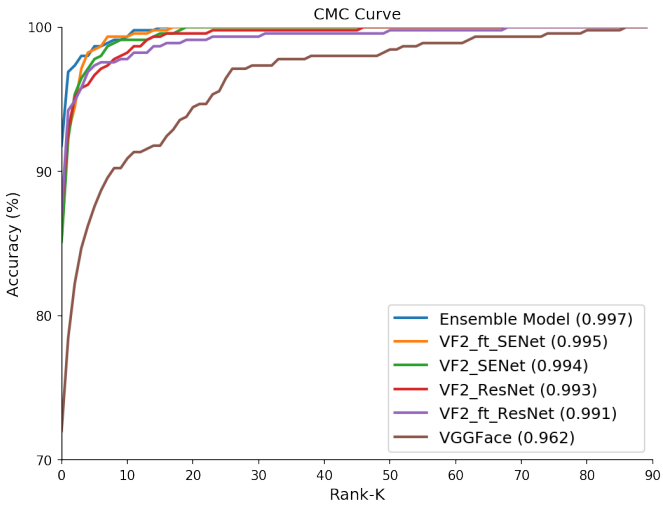

Figure 4: The CMC curve of the experiments in this work compared to the approach using VGGFace in feature extraction step which achieved the highest performance in ICB-RW 2016 challenge [3].

IV. Conclusion

In this paper, we proposed a system for face recognition under mismatched conditions. We utilized ResNet-50 [5] and SENet-50 [4] deep learning models trained on VGGFace2 dataset [2] to extract deep face representation. We presented that rapid development of the deep learning models in the field of face recognition provides us a powerful tool to extract robust face features across different datasets and domains, as in our case, where gallery and probe images are collected under different conditions. Our best model, the ensemble of four models were used in this work to obtain 91.78% Rank-1, 98.00% Rank-5, and 0.997 AUC of the CMC, increasing the best reported results in [3] by 21.98%, 12.7%, and 0.045, respectively. Furthermore, we analyzed the misclassified faces by the proposed system. Figure 5 shows Rank-1 misclassified faces in our best model. We observed that there are still many problems to be addressed to solve the problem of face recognition under mismatched conditions. We noticed that the misclassified faces still suffer from strong pose variations, occlusions, e.g, sunglasses, cellphones, and motion-blur effect of the images captured in real-world surveillance conditions. We plan to address these issues in our future works. We also plan to explore different fusion strategies, besides feature fusion.

Figure 5: Rank-1 misclassified faces by the proposed ensemble model. It shows that sunglasses, strong pose variations, and blur are the most important reasons for misclassification in the ICB-RW 2016 dataset [3].

References

[1] K. Simonyan, O.M. Parkhi, A. Vedaldi, and A. Zisserman, "Fisher Vector Faces in the Wild", BMVC, 2013.

[2] Q. Cao, L. Shen, W. Xie, O.M. Parkhi, and A. Zisserman, "VGGFace2: A dataset for recognising faces across pose and age", arXiv:1710.08092 2017.

[3] J. Neves, and H.& Proença, "ICB-RW 2016: International Challenge on Biometric Recognition in the Wild", Biometrics (ICB), 2016 International Conference on (pp. 1-6). IEEE.

[4] H. Jie, L. Shen, and G. Sun, "Squeeze-and-excitation networks", arXiv:1709.01507 2017.

[5] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition", CVPR, 2016.

[6] K. Simonyan, and A. Zisserman, Very deep convolutional networks for large-scale image recognition", ICLR, 2015.

[7] G.B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments", Technical Report, 2007.

[8] L. Wolf, T. Hassner, and I. Maoz, "Face recognition in unconstrained videos with matched background similarity", CVPR, 2011.

[9] Y. Taigman, M. Yang, and M.A. Ranzato, "Deepface: Closing the gap to human-level performance in face verification", CVPR, 2014.

[10] F. Schroff, D. Kalenichenko, and J. Philbin. "Facenet: A unified embedding for face recognition and clustering", CVPR, 2015.

[11] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions", CVPR, 2015.

[12] Y. Sun, D. Liang, X. Wang, and X. Tang, "DeepID3: Face recognition with very deep neural networks", arXiv:1502.00873 (2015).

[13] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "Sphereface: Deep hypersphere embedding for face recognition", CVPR, 2017.

[14] Y. Wong, S. Chen, and S. Mau, "Patch-based probabilistic image quality assessment for face selection and improved video-based face recognition", CVPRW, 2011.

[15] Z. Huang, S. Shan, R. Wang, H. Zhang, S. Lao, A. Kuerban, and X. Chen, "A benchmark and comparative study of video-based face recognition on COX face database", IEEE Transactions on Image Processing 24.12 (2015): 5967-5981.

[16] M. Grgic, D. Kresimir, and S. Grgic, "SCface–surveillance cameras face database", Multimedia Tools and Applications 51.3 (2011): 863-879.

[17] D.E. King, "Dlib-ml: A Machine Learning Toolkit", Journal of Machine Learning Research 10, pp. 1755-1758, 2009

[18] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. "Caffe: Convolutional architecture for fast feature embedding", In Proceedings of the 22nd ACM International Conference on Multimedia Pages 675-678

[19] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, "Ms-celeb-1m: A dataset and benchmark for large-scale face recognition", ECCV, Springer, Cham, 2016.

[20] V. Kazemi, and J. Sullivan, "One Millisecond Face Alignment with an Ensemble of Regression Trees", CVPR, 2014.

[21] O.M. Parkhi, A. Vedaldi, and A. Zisserman. "Deep Face Recognition", BMVC, 2015.