Self-portrait to ID Document face matching: CNN-Based face verification in cross-domain scenario

Filipe Costa[1], Marcos Vinícius L. Melo[1], Igor Gadelha[1],
Guilherme Fôlego[1], Larissa Gambaro[2] and André Rodrigues[2]

[1]CPQD - Artificial Intelligence and IoT Solutions, Brazil
[2]CPFL - Strategy and Innovation Directory, Brazil

***Abstract*—Face verification approaches determine whether two given faces are from the same person. Recently, a new demand for face verification application which has become popular in commercial applications is the self-portrait and ID face matching, in which we compare the faces of a selfie shot by a subject and the face in a picture of her identification document. In this work, we proposed a novel approach for face verification in a cross-domain scenario, assuming we have only two images for each subject in the dataset. The method is based on siamese architecture with triplet-loss function. Experiments show the proposed model reaches good effectiveness for cross-domain face verification with low error rates, in comparison to other works of the literature.**

***Index Terms*—Selfie-document, Face Verification, Triplet-loss, Siamese networks**

I. Introduction

One of the most important tasks in computer vision is face recognition [1], which has been receiving increased attention from the scientific community due to the wide range of applications in a number of environments, such as surveillance, access control system, social networks, and others.

The face recognition problem is usually divided into three main categories [2]: *face verification*, in which we aim at determining if a pair of images corresponds to the same individual; *face identification*, in which we assume there exists a set of several cataloged subjects (usually called *gallery*), then we analyse a query image (defined as *probe*), and check if the query subject corresponds to one of those cataloged; and *watch-list*, which is similar to face identification, but dealing with an open-set environment, where we cannot guarantee that all query subjects are registered in the face gallery.

Specifically for the face verification task, it can be described as a $1:1$ matching problem. Several applications can be designed, including access permission to restricted places or bank accounts, for instance. Several works in the literature address face verification [3]–[6].

Recently, a new demand for face verification application has become increasingly popular in commercial applications. It has become common to companies, such as financial institutions, to allow customers to create accounts using the Internet (*i.e.*, without the necessity of going to a physical location). This helps institutions to acquire new clients, reduce bureaucracy and, assuming the recent COVID-19 pandemic, it can avoid lines in banks, agglomeration and mainly the spread of the disease.

Corresponding author: fcosta@cpqd.com.br

Given this reality, face verification has been used for automatically verifying the user's identity information and to avoid identity frauds. A user can capture a picture of her face (a "selfie") and a document photograph using a smartphone. Then, those images are sent to the company, which will compare the faces of the images and, if there is a match between them, it means there is no problem or fraud, and further actions can be performed. Some works in the literature have approached face verification in this cross-domain scenario, considering selfie-portrait and documents photographs [7]–[10].

Considering that there are only a few works in the literature approaching this problem of face verification in the cross-domain scenario, in this paper we present a novel approach for face verification in cross-domain scenario, considering self-portrait to ID document matching. The proposed approach is based on a siamese architecture with triplet-loss function [11]. Comparison with other works of the literature shows great improvement of results.

One of the considerable barriers to the development of new techniques for face verification in selfie-document scenario is the unavailability of suitable datasets, as document pictures have sensitive information. Thus, researchers have no common base to evaluate or compare their methods with other results. For the experiments, we designed a private dataset containing approximately $70{,}000$ images, in which there is only a pair of images for each subject: one self-portrait photograph, and one ID document picture, both on real-world conditions.

In summary, the main contributions of this work are:

- Dealing with cross-domain face verification in a few-shot scenario, in which we have only a pair of image for each individual
- Collection of a new private dataset containing $69{,}998$ images from $34{,}999$ subjects, containing a pair of images

for each individual (a self-portrait and a document photograph)[1]
- The design of a siamese architecture with triplet-loss function for selfie-document face verification
- Results for cross-domain face verification which outperforms other works of the literature.

This paper is organized as follows. Section II presents related works. Details about the used datasets are provided on Section III. The proposed approach is described in Section IV. Section V presents the experimental protocol, the obtained results and the discussion about them. Finally, we conclude the work and point out future directions in Section VI.

II. Related works

In the last decades, facial recognition and verification have gained a lot of attention within the fields of research in computer vision [12], [13]. This paper proposes the comparison between different approaches to facial verification cross-domain problem.

Usually, facial recognition systems consist of three general stages: face detection, feature extraction, and face recognition [14]. At the first step, several techniques are often used, such as Haar algorithms [15], [16], which are efficient and have high detection rates. For feature extraction, the literature presents a wide variety of options, with two main approaches when the features are located in a face: handcrafted features extraction - using descriptors manually created by computer vision specialists, such as Local Binary Patterns (LBP) [17] and Discrete Cosine Transform (DCT) [18], [19] - and automatic feature learning [20]. Finally, we have the recognition stage, in which we use the extracted descriptors to verify that face corresponds to an individual. It can be done in two different manners: *face verification*, in which we verify if two different faces are from the same individual; or *face identification* in which we compare a face with a previously obtained descriptor data in a database. For this purpose, machine learning techniques can be applied, such as Logistic Regression [21] and Support-Vector Machines [22].

The problem of cross-domain face verification involves a series of difficulties that are found in general facial recognition problems, such as illumination, pose and expression changes [23]. Moreover, when the comparison is between a face in an ID Document and a face on a self portrait, several issues are added, such as low quality of the ID picture - this low quality can be due to a range of factors, such as the fact that the images contained in the document were real world photos, later scanned, printed, and finally photographed again by the application; during these steps, there is great potential of quality reduction. Additionally, depending on the storage and age of the document, the image can suffer the action of many abrasives (such as the friction between the paper and the plastic) and have the quality reduced even more. Furthermore, a possible large time difference between the photos can contribute to hinder the recognition process.

[1]Due to privacy concerns and in according to the General Data Protection Rule (GDPR), we will not be able to make this dataset public.

Over the years, the techniques needed to solve the problem of recognition with ID Documents have improved. At first, the problem was addressed using widely-known gradient map recognition algorithms, considering the only difference from conventional approaches (same-domain images) being the cross-domain data entry [24].

Folego et al. [7] proposed an approach for dealing with selfie-document face matching. First, the authors uses the classic Viola-Jones algorithm [15] for face detection and extraction. Then, a histogram equalization is applied over the faces for increase the contrast. The images are L2-normalized and the authors used a pre-trained VGGFaces [25] model for feature extraction. Finally, the authors compare the features with euclidian distance and consider it a match if the distance is below a pre-defined threshold. The authors perform experiments in a very small database, containing 50 individuals. Similar approaches were proposed in [8], [10], [26].

More recently, a method called Dynamic Weight Imprinting (DWI), proposed by Shi et al. [27], achieved promising results for selfie to ID document face verification. This method consists of an update of weights in a Convolutional Neural Network (CNN) classifier which allows faster convergence and more generalized representations. After that, a pair of sibling CNNs are trained to learn a unified face representation with parameters specific by domain. Latest results presented by the same authors in [9] achieved a true acceptance race (TAR) of about 96%.

Albiero et al. [28] proposed an approach for selfie-document face matching considering adolescent people, using a private dataset, where they match live faces images taken at a restricted age interval of 18-19 years old. In this work, the authors fine-tuned an existing approach for face verification for few-shot learning. The work reported true acceptance rate of approximately 96% for the private dataset.

III. Datasets

In this section, we introduce the datasets that are used in our work. We believe the comparison of the approaches investigated in this work is important to the scientific community, showing the difficulties of dealing with the problem of face verification in cross-domain scenario.

A. LFW

Labeled Faces in the Wild (LFW) [29] is a well-known public benchmark for face verification. The dataset contains more than $13{,}000$ images of faces collected from the web of $1{,}680$ people. There are two or more distinct photos for each subject in the dataset. All the faces in this dataset were detected and extracted by the Viola-Jones face detector [16].

This dataset was used for pre-training the backbone network of the proposed method. More details is given on Section IV.

B. Private Dataset

There are some datasets specific for self-portrait to ID-document face matching, such as the proposed in [8]–[10].

Fig. 1. Samples of the private dataset. Each column is a pair of images from the same subject. The pictures were modified on purpose just for depiction, guaranteeing the privacy of the users' identities.

Unfortunately, these datasets are private or unavailable to download. Thus, in order to evaluate the approach proposed in this paper, we designed a private dataset composed of selfie-portrait photographs and ID documents pictures, captured between January and March of 2020. This dataset originally contained $81,294$ images from $40,647$ different individuals: one self-portrait and one correspondent document ID picture with a photograph with proportion of $3cm \times 4cm^2$.

The images of this dataset were captured in a real-world setting, with great variability, such as different illumination and backgrounds, use of flashlight, different angles of captures, individuals with and without glasses, and among others. We also considered the two most common documents with photograph in Brazil, namely, the General Register (*Registro Geral* - RG) and the Brazilian Driver's License (*Carteira Nacional de Habilitação* - CNH). These two documents differ in size, proportion and mainly the position of the user's photograph and identification information. Figure 1 shows some samples of this dataset. We did not separate the dataset according to the kind of document, aiming at maintaining the variability of the dataset in the real world.

C. Face segmentation and pre-processing

To remove unnecessary regions of the images (*e.g.*, document information, background, etc.), we needed to perform face detection and extraction. For this task, we used the approach proposed in [30], which uses a Multi-task Cascaded Convolutional Neural Networks (MTCNN) for face extraction. This step is important to set only the face of subjects in the photographs as input of our model.

Taking into account MTCNN may not work effectively on faces with different orientations, we carry out a series of rotations, with angles from 0 to 135 degrees, in clockwise and counterclockwise directions, with steps of 45 degrees. This was necessary as the private database can have high degree of positioning in several images, which were captured horizontally or in upside-down orientations.

After rotations, 708,307 possible face images are generated, however, most of these images are mistaken detections and must be removed in order not to compromise our dataset. Then, images are filtered by confidence and area, and only faces with confidence score above 99% were used. After this process, we removed duplicated faces and we ended up with a total of $69,998$ images (pair of ID document and selfie portrait faces), totaling $34,999$ subjects.

[2]Real world dimension.

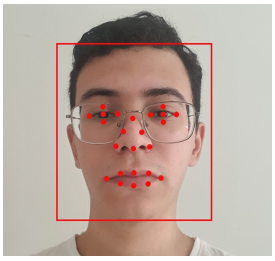

Fig. 2. Face keypoints location example: eyes, nose and mouth. Used to generate face descriptor.

One advantage of the MTCNN-based face detection approach [30] is that it provides the coordinates of eyes, nose and mouth as depicted in Figure 2. Thus, we improve the face alignment based on the eyes, nose and mouth positions, setting up all faces in vertical orientation (*i.e.*, we align the eyes horizontally with nose and mouth below them).

After extracting and aligning the faces, we enhance the contrast using the Contrast Limited Adaptive Histogram Equalization (CLAHE) method [31]. This method equalizes the histogram in image considering small blocks instead of performing a global histogram equalization. This way, it avoids over-brightness noise amplification, typical when capturing printed images (in our case, pictures of a document). We assume the default configurations of this technique.

IV. Cross-domain selfie-document face verification

Dealing with cross-domain face verification has some challenges. As mentioned before, we assume only two images for each individual. This way, it is hard for any architecture to learn specific features for an individual. Furthermore, the nature of both pictures are different, as one is a self-portrait and the other is a re-capture of a picture printed in a document.

In order to deal with these challenges, we designed a siamese architecture for face verification in this cross-domain scenario. Siamese networks were originally proposed by Bromley et al. [32] for signature recognition and have been used in several applications, including object tracking [33], gait recognition [34], and face recognition [35]. It is a type of architecture which contains two or more subnetworks with the same configuration, parameters and weights. In other words, it contains instances of the same base network. This type of architecture is commonly used for comparing inputs and to find similarities between them, which allows it to be used in several applications. To the best of our knowledge, there is no work in the literature using siamese networks for cross-domain face verification specifically for self-portrait to ID Document face matching.

We choose a siamese architecture due to the nature of the selfie-ID face verification problem we approach in this work. Siamese networks are more robust to class imbalances and commonly used in few-shot and one-shot problems [36], since a few images per class is sufficient for learning from semantic similarity between images, putting feature vectors from same classes close together while try to increase the distance from feature vectors from different classes. In fact, as we mentioned before, in our private dataset we have only two images for each subject.

We design the proposed siamese architecture as follows. First, we considered the ResNet50 [37] CNN architecture as the backbone model. This choice was made because this architecture is widely adopted in the literature and in state-of-the-art research, achieving excellent results in previously cited researches, presenting good effectiveness and efficiency. Then, we perform a pre-training of the ResNet50 over the LFW dataset assuming a multi-class problem (face identification). This was performed due to the fact that we have only one pair of images for each individual in the private dataset, which can make the model very sensitive to the lack of data. In this way, our hypothesis is this pre-training will reduce this sensitivity and let the backbone network better differentiate the subjects and improving the results of the experiments.

After the pre-training, we removed the final softmax layer from the ResNet50 and added an L2 normalization layer. As such, we no longer have a classification as output of the backbone network, but a feature vector in a surface of a hypersphere with radius of one. This feature vector is called *embedding*. The L2 normalization was performed due to the fact features extracted in different domains (e.g., selfie vs. ID document) might have significantly different magnitude ranges.

The proposed siamese architecture considers three instances of the backbone model with shared weights. Each one of the instances will receive a different image as input. The first one receives an image called *anchor*, which is the photograph of the individual under investigation. The second input is called *positive*, which is a picture of the same individual of the anchor. Finally, the last input is defined as *negative*, which is a picture of a subject different from anchor and positive inputs. The set of these three images is defined as a *triplet*.

However, considering the private dataset used in this work, in which we have only two images for each individual, we first assume all self-portrait are *anchor* images, the correspondent ID document photographs are *positive* images, and, for each pair (*anchor*, *positive*), we randomly select five ID document faces from different individuals and set each one of them in one triplet as the *negative* input. We empirically opted for five negative images due to time and hardware constraints.Then, we repeat the process with the difference that we set the ID document faces as anchor and the self-portrait faces as positive and negative inputs. This way, for each pair of selfie-document pictures, we end up with 20 inputs for our network, which mitigates the few-shot problem. Figure 3 depicts the triplets generator process.

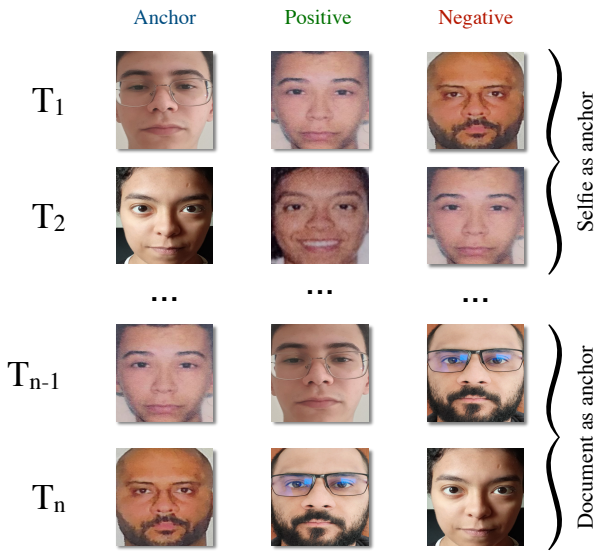

Fig. 3. Triplets generation, with different samples, using in the first scenarios an anchor as selfie image, and in last scenarios an document image as anchor.

We assumed the loss function called *triplet-loss*, proposed in [11] for face recognition. The triplet-loss function is defined by

$$\mathcal{L}(a, p, n) = max(0, D(a, p) - D(a, n) + \alpha), \qquad (1)$$

in which a, p and n are the outputs of base model for anchor, positive and negative images as input, respectively, D is the Euclidian Distance function and α is a margin that is enforced between positive and negative pairs (in our experiments, we empirically set $\alpha = 0.2$). This loss function focuses on approximating anchor and positive embeddings inputs, while separating anchorand negative embeddings.

In the validation step, we aim at reducing the Equal Error Rate (EER), which is an statistic measurement to calculate the performance of the face verification in validation step. It is the location on the Detection Error Tradeoff (DET) curve in which the False Acceptance Rate (FAR) and False Rejection Rate (FRR) are equal. The EER value is only calculated on validation step. The train process using Triplet Loss are depicted in Figure 4.

When the training is finished, a threshold τ_{EER} is defined as the index of the EER. This threshold is used to calculate the performance of face verification on test set (details will be described on Section V).

V. Experiments and Results

In this section, we describe the protocol used for our experiments and the evaluation of the approaches for self-portrait to ID Document face matching.

First, we split the private dataset for the experiments according to the number of individuals, using the same subsets for all the experiments. We assumed 60% of individuals for training, 20% for validation, and 20% for testing. This way, there will not be overlap of individuals between these subsets.

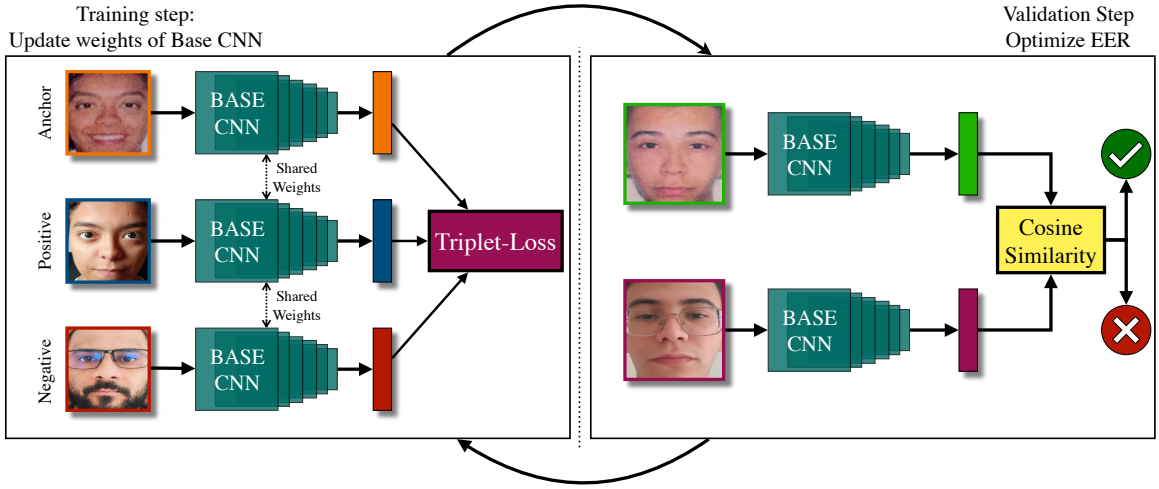

Fig. 4. Example of one epoch of training and validation steps of the Siamese model with Triplet-loss function.

The training step was performed during 100 epochs, with early stopping after 25 epochs without decrease in EER during the validation step. We also assumed batch size equal to 32 and Adam optimizer [38] with learning rate of 10^{-4}, exponential decay rate for the 1st moment 0.9 and default values for other parameters. All training and inference processes were executed in a machine with CPU Intel Xeon Octa-core 2.3GHz with 32GB of RAM and GPU NVidia Tesla T4 with 16 GB of VRAM.

The following metrics were used for evaluating results of our proposed approaches and comparing with other works: Area under ROC Curve (AUC), Precision, Recall and Half-Total Error Rate (HTER). The HTER measurement is done by averaging the False Acceptance Rate (FAR) and the False Rejection Rate(FRR).

We compare the performance of our approach with the methods proposed by Shi et al. [9] and the approach proposed by Folego et al. [7]. Specifically for the work proposed in [7], and aiming at investigating variations which could improve the face verification results, we also implemented two modifications of the feature extractor model. The first one was to perform the face verification with a softmax layer instead of comparing the embeddings with a distance metric. With this, we wanted to see if a classification approach could report good results as performed in [9]. Secondly, we substituted the loss function and used the Focal Loss function, proposed originally by Lin et al. [39] for dense object recognition. This choice was due to focal loss being originally designed to address the issue of class imbalance problem and few-shot problems on object detection problem.

Table I presents the results of the investigated approaches for face verification in selfie-document scenario. We note that the proposed siamese approach with Triplet-loss reported the best results for face verification on selfie-document scenario, considering our private dataset. It can be explained due to the fact that this kind of architecture is robust to low amount of labeled data, since the training was performed by combining images from different subjects in triplets, which mitigate the few-shot problem.

The results also demonstrate that the modifications we designed for the approach proposed by Folego et al [7] do not present improvements in results, discarding our hypothesis that using softmax layer instead of embedding comparisons for this model could improve the effectiveness.

TABLE I
EXPERIMENTS RESULTS

Method	AUC	Precision	Recall	HTER
Folego et al [7]	0.853	0.767	0.786	0.250
Folego et al [7] + Softmax	0.702	0.708	0.472	0.353
Folego et al [7] + Focal Loss	0.621	0.704	0.180	0.416
Shi et al. [9]	0.633	0.731	0.138	0.408
Our approach	**0.971**	**0.918**	**0.905**	**0.111**

We also performed a qualitative analysis of our method. Investigating the error cases, we observed the main problem was images with low quality, blurred images and images with illumination problems, such as dark images and faces with over exposure.

VI. Conclusion and Future Work

In this paper, we address the problem of matching faces in a cross-domain scenario, which involves many complications when comparing images from very different sources. The solution proposed here had a very satisfactory performance, generating a concise facial recognition model with a much smaller number of samples (in the order of thousands) than the large datasets, with a few million images. Also, our private dataset contains only two images per subject, in comparison to others that have hundreds of images from a single individual. Based on our analysis of experimental results, we can conclude that, for the cross-domain verification problem, the Siamese CNN with Triplet-Loss obtained the best

results compared to the already known approaches such as VGGFaces and DIAM-Softmax. With the growing evolution of the techniques used, this study may present better future results with different approaches and dataset improvements, such as size and composition. Furthermore, our proposed system performed well and could be adapted for other cross-domain face verification problems with different kind of samples.

ACKNOWLEDGMENTS

This work was performed as part of the *MOVERS* project, which is conducted by CPQD Foundation in partnership with CPFL Energy Company. The authors would like to thank the CPFL group for technical and financial support, through the Research and Development project PD-00063-3070/2019 with resources from ANEEL's R&D program.

REFERENCES

[1] W. Zhao, R. Chellappa, J. Phillips, and A. Rosenfeld, "Face recognition: A literature survey," *ACM Computing Surveys*, vol. 35, no. 4, pp. 399–458, 2003.

[2] R. Chellappa, P. Sinha, and P. Phillips, "Face recognition by computers and humans," *IEEE Computer*, vol. 43, no. 2, pp. 46–55, 2010.

[3] R. Vareto, S. Silva, F. Costa, and W. R. Schwartz, "Face verification based on relational disparity features and partial least squares models," in *IEEE Conference on Graphics, Patterns and Images*, 2017, pp. 209–215.

[4] Y. Sun, X. Wang, and X. Tang, "Hybrid deep learning for face verification," in *IEEE International Conference on Computer Vision*, 2013, pp. 1489–1496.

[5] ——, "Deep learning face representation by joint identification-verification," *arXiv preprint arXiv:1406.4773*, 2014.

[6] W. Wang, Y. Fu, X. Qian, Y.-G. Jiang, Q. Tian, and X. Xue, "Fm2u-net: Face morphological multi-branch network for makeup-invariant face verification," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5730–5740.

[7] G. Folego, M. A. Angeloni, J. A. Stuchi, A. Godoy, and A. Rocka, "Cross-domain face verification: Matching id document and self-portrait photographs," in *XII Workshop de Visão Computacional*, 2006, pp. 311–316.

[8] R. Paliwal, S. Yadav, and N. Nain, "Faceid: Verification of face in selfie and id document," in *Springer International Conference on Computer Vision and Image Processing*, 2019, pp. 443–454.

[9] Y. Shi and A. K. Jain, "Docface+: Id document to selfie matching," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 1, no. 1, pp. 56–67, 2019.

[10] J. Oliveira, G. Souza, A. Rocha, F. Deus, and A. Marana, "Cross-domain deep face matching for real banking security systems," in *IEEE International Conference on eDemocracy & eGovernment*. IEEE, 2020, pp. 21–28.

[11] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 815–823.

[12] J. Lu, Y.-P. Tan, and G. Wang, "Discriminative multimanifold analysis for face recognition from a single training sample per person," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 39–51, 2013.

[13] D. Yi, Z. Lei, and S. Z. Li, "Towards pose robust face recognition," in *2013 IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3539–3545.

[14] S. Zafeiriou, G. Tzimiropoulos, M. Petrou, and T. Stathaki, "Regularized kernel discriminant analysis with a robust kernel for face recognition and verification," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 23, no. 3, pp. 526–534, 2012.

[15] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, vol. 1, 2001, pp. I–I.

[16] P. Viola and M. J. Jones, "Robust real-time face detection," *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, 2004.

[17] T. Ahonen, A. Hadid, and M. Pietikainen, "Face description with local binary patterns: Application to face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 12, pp. 2037–2041, 2006.

[18] Z. M. Hafed and M. D. Levine, "Face recognition using the discrete cosine transform," *International Journal of Computer Vision*, vol. 43, no. 3, pp. 167–188, 2001.

[19] M. J. Er, W. Chen, and S. Wu, "High-speed face recognition based on discrete cosine transform and rbf neural networks," *IEEE Transactions on Neural Networks*, vol. 16, no. 3, pp. 679–691, 2005.

[20] G. B. Huang, H. Lee, and E. Learned-Miller, "Learning hierarchical representations for face verification with convolutional deep belief networks," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 2518–2525.

[21] P. McCullagh and J. A. Nelder, *Generalized linear models*. Routledge, 2019.

[22] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[23] W. Chen, M. J. Er, and S. Wu, "Illumination compensation and normalization for robust face recognition using discrete cosine transform in logarithm domain," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 36, no. 2, pp. 458–466, 2006.

[24] V. Starovoitov, D. Samal, and B. Sankur, "Matching of faces in camera images and document photographs," in *2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.00CH37100)*, vol. 4, 2000, pp. 2349–2352.

[25] O. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *British Machine Vision Conference*, 2015, pp. 1–12.

[26] T. Bourlai, A. Ross, and A. Jain, "On matching digital face images against scanned passport photos," in *IEEE International Conference on Biometrics, Identity and Security*, 2009, pp. 1–10.

[27] Y. Shi and A. K. Jain, "Docface: Matching id document photos to selfies," in *2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, 2018, pp. 1–8.

[28] V. Albiero, N. Srinivas, E. Villalobos, J. Perez-Facuse, R. Rosenthal, D. Mery, K. Ricanek, and K. W. Bowyer, "Identity document to selfie face matching across adolescence," in *IEEE International Joint Conference on Biometrics*, 2019, pp. 1–9.

[29] G. B. M. Mattar, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database forstudying face recognition in unconstrained environments," in *Workshop on faces in 'Real-Life' Images: detection, alignment, and recognition*, 2008, pp. 1–14.

[30] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016.

[31] S. Pizer, E. Amburn, J. Austin, R. Cromartie, A. Geselowitz, T. Greer, B. Romeny, J. Zimmerman, and K. Zuiderveld, "Adaptive histogram equalization and its variations," *Elsevier Computer Vision, Graphics, and Image Processing*, vol. 39, no. 3, pp. 355–368, 1987.

[32] J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, and R. Shah, "Signature verification using a "siamese" time delay neural network," *Advances in Neural Information Processing Systems*, vol. 6, pp. 737–744, 1993.

[33] Q. Guo, W. Feng, C. Zhou, R. Huang, L. Wan, and S. Wang, "Learning dynamic siamese network for visual object tracking," in *IEEE International Conference on Computer Vision*, 2017, pp. 1763–1771.

[34] C. Zhang, W. Liu, H. Ma, and H. Fu, "Siamese neural network based gait recognition for human identification," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2016, pp. 2832–2836.

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

[36] G. Koch, R. Zemel, and R. Salakhutdinovn, "Siamese neural networks for one-shot image recognition," in *Deep Learning Workshop at International Conference on Machine Learning*, 2015.

[37] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[38] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014.

[39] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *IEEE International Conference on Computer Vision*, 2017, pp. 2980–2988.