A. Brömme, C. Busch, A. Dantcheva, K. Raja, C. Rathgeb and A. Uhl (Eds.): BIOSIG 2020,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2020 1

Action-Independent Generalized Behavioral Identity Descriptors for Look-alike Recognition in Videos

Ali Khodabakhsh[1], Hugo Loiselle[2]

Abstract: There is a long history of exploitation of the visual similarity of look-alikes for fraud and deception. The visual similarity along with the application of physical and digital cosmetics greatly challenges the recognition ability of average humans. Face recognition systems are not an exception in this regard and are vulnerable to such similarities. In contrast to physiological face recognition, behavioral face recognition is often overlooked due to the outstanding success of the former. However, the behavior of a person can provide an additional source of discriminative information with regards to the identity of individuals when physiological attributes are not reliable. In this study, we propose a novel biometric recognition system based only on facial behavior for the differentiation of look-alikes in unconstrained recording conditions. To this end, we organized a dataset of 85,656 utterances from 1000 look-alike pairs based on videos collected from the wild, large enough for the development of deep learning solutions. Our selection criteria assert that for these collected videos, both state-of-the-art biometric systems and human judgment fail in recognition. Furthermore, to utilize the advantage of large-scale data, we introduce a novel action-independent biometric recognition system that was trained using triplet-loss to create generalized behavioral identity embeddings. We achieve look-alike recognition equal-error-rate of 7.93% with sole reliance on the behavior descriptors extracted from facial landmark movements. The proposed method can have applications in face recognition as well as presentation attack detection and Deepfake detection.

Keywords: Behavioral Biometrics, Face Recognition, Look-alike face, Facial Motion, Triplet Loss.

Fig. 1: Examples of look-alike identity pairs in the proposed 1000 look-alike pairs (1000LP) dataset. Each column shows one pair of look-alikes. The identities in the proposed dataset are a subset of the identities in the VGGFace2 [Ca18] dataset.

1 Introduction

Distinguishing visually similar individuals, be it identical twins or look-alikes with physical make-up or plastic surgery, has been challenging for both humans and face recognition

[1] NTNU, IIK, Norwegian Biometrics Lab, Gjovik, NO, ali.khodabakhsh@ntnu.no
[2] Orange, Région de Caen, Fr, loiselle.hugo@hotmail.fr

algorithms [La11]. In the context of video communication, this vulnerability is further exacerbated as other means of identity verification are often not available. Moreover, the use of look-alikes and make-up for fraud has an advantage over digital manipulation methods as they don't produce any digital footprint in the received signal to be used for detection. Furthermore, despite the rise of advanced digital video manipulation methods such as Deepfakes, subjective tests show higher susceptibility of viewers to fake videos containing look-alikes rather than digitally manipulated videos [KRB19]. Fortunately, a video signal contains additional clues on the identity of the person in the form of facial behavior [Be10, KJ97].

Among existing methods for behavioral face recognition (BFR), the vast majority of studies focus on fixed-phrase authentication or specific emotional responses. Chen et al. [LLJ01] propose use of dense optical flow vector distance for identification in a fixed-phrase scenario. In [Ce06] Cetingul et al. experiment with dense motion features, lip contour motion features, and lip shape features with a hidden-Markov-model (HMM) classifier. Zafeiriou and Pantic [ZP11] use principal component analysis (PCA) followed by linear discriminant analysis (LDA) on dense facial deformation features in spontaneous smile for biometric recognition. Wang and Liew [WL12] show that behavioral lip biometrics based on temporal shape descriptors and motion vector representation outperforms physiological lip biometrics based on texture descriptors. Gavrilescu [Ga16] proposes a multi-state neural network on individual facial expressions extracted in the form of facial action coding system (FACS). More recently, Iengo et al. [Ie19] use neural networks on dynamic facial features to achieve a fixed-phrase recognition rate of 98.2% and Taskirar et al. [Ta19] use statistical properties of facial distances during different phases of smile facial expression for face recognition.

A number of publications have attempted to address unconstrained BFR. Matta and Dugelay [MD06] propose using rigid head displacements along with GMM and Bayesian classifiers for person recognition. Ye and Sim [YS10] use locally similar facial deformation patterns for identification through the calculation of local deformation profile similarity. In [Sh16], Shreve et al. quantify the type and intensity as well as the temporal dynamics of action units (AU) via calculating histogram distances and dynamic time warping (DTW) distance. Yuan et al. [Yu17] propose the usage of active shape models on lip contour along with gaussian mixture models (GMM) for authentication in smartphone applications.

BFR has also been used in multi-modal biometric recognition as well as presentation attack detection (PAD). Notably, Zhao and Pietikanien [ZP07] introduce local binary patterns (LBP) on three orthogonal planes and volume LBPs and thus incorporates immediate neighborhood frames of the video for face recognition. Kim et al. [KKR16] use long short-term memory (LSTM) cells on top of convolutional neural networks (CNN) to capture smile facial dynamics. More recently, Pan and Deravi [PD17] use support vector machine (SVM) on AU histogram features for presentation attack detection. Finally, Agrawal et al. [Ag19] model facial expressions of four individuals using facial landmarks and SVM to detect Deepfakes.

To distinguish look-alikes from each other many image-based methods have been proposed. Klare et al. [KPJ11] provide a taxonomy of facial features and analyze the dis-

criminative power of these features for identical twin identification. The only video-based solution is proposed by Zhang et al. [Zh14], where they extracted six types of face motion from the talking profile of identical twins and use the similarity of aligned motion sequences for classification by an SVM model. To the best of the authors' knowledge, there exists no publicly available video dataset of look-alikes in the literature. The only related video dataset in the literature is the private dataset by Zhang et al. [Zh14] collected from 39 pairs of twins at the Mojiang International Twins Festival. There also exists a couple of related datasets containing solely images. Lamba et al. [La11] collected the only dataset on look-alikes consisting of 500 images from 50 celebrities and their look-alikes. Phillips et al. [Ph11] collected a dataset of 435 twins consisting of 24050 images.

All aforementioned publications rely on small data collected in controlled environments, and few of them address emotion- and utterance-independent detection with limited success, and as such, among all publications regarding this topic, none have addressed the unconstrained BFR in real-world scenarios. In this study, we introduce a general-purpose action-independent identity descriptor extractor based on facial behavior for distinguishing look-alikes. To this end, we also provide the first large-scale look-alike video dataset named "1000 look-alike pairs (1000LP)" which consists of approximately $23,000$ real-world videos collected from a public video-sharing platform[3], for which both humans and state-of-the-art recognition systems fail at differentiation[4]. Among the aforementioned literature, the approach in this article is in the same line of research as is taken by Zhang et al. [Zh14] and Agrawal et al. [Ag19]. The rest of this article is organized as follows: in Section 2 the proposed method is described, while Section 3 includes the details of the collected dataset as well as the experiment setup. The results of the experiments are discussed in Section 4 and the article is concluded in Section 5.

2 Proposed Method

The physiological likeliness of two individuals due to natural similarity or application of physical or digital makeup may lead to false-positives in face recognition. In these cases, the facial behavior can be a source of complementary information for face recognition. Facial behavior contains identifiable information and has a significant role in person identification by humans [Be10, KJ97]. In our proposed method, after face detection and facial landmark extraction in each frame, we train a convolutional deep neural network (CDNN) which maps the sequence of normalized landmark positions in the video to a vector in a generalized behavior space in an end-to-end manner. This approach enables the recognition of persons that are previously unseen by the detector by simply calculating the distance between behavior-vectors extracted from a pair of videos. Furthermore, as the network only sees the landmarks, it is guaranteed to be void of influence by the physiological likeliness of the individuals. Furthermore, landmarks are not as sensitive to disturbances and quality-related issues as other features such as optical and motion fields are and can be extracted with higher confidence.

[3] http://www.youtube.com

[4] The dataset is publicly available for download at `http://ali.khodabakhsh.org/research/1000lp/`

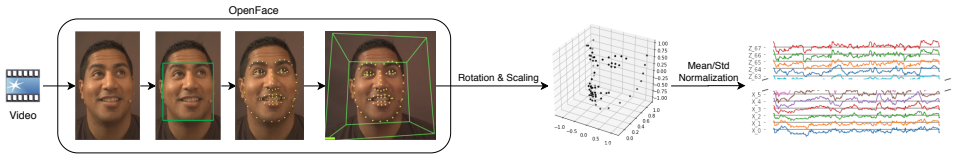

Fig. 2: Feature extraction pipeline.

2.1 Preprocessing

We use the open-source facial behavior analysis toolkit OpenFace [BRM16] to extract the landmark positions from videos. The toolkit provides face detection as well as pose estimation and 3D landmark positions for each frame in the video. For landmark positions to be independent of the camera position and head rotation angle, we use the pose estimation information to rotate the 3D landmarks in 3D space to achieve a frontal pose of zero degrees roll, yaw, and pitch. Further on, the landmark positions in each video are scaled to match a fixed scale used over the whole dataset. The scaling is done such that the inner eye corner landmarks would be on average 0.5 units apart. Finally, the landmarks are individually normalized using their mean and standard deviation across the whole training dataset. The aim of the aforementioned normalization steps is to convert the landmark positions to rotation-independent displacements from the average position. Even though the pose information can also contain additional behavioral identity information, they were left out due to their dependence of the estimated pose to the camera angle and position. Figure 2 visualizes the preprocessing pipeline.

2.2 The proposed recognition system

To extract identity-sensitive yet action-independent information from the time series of landmark movements, it is fruitful to rely on the distribution statistics of the landmark deviations. However, due to the noisy nature of the estimated 3D landmark positions extracted from 2D videos in the pre-processing step, a refinement step proves necessary. However, the refinement criteria are ambiguous as the correct landmark position is not available. Furthermore, the movements are correlated to a large extent and contain redundancies. Motivated by the recent success of x-vectors [Sn18] in the field of speaker recognition, we propose the network architecture shown in Figure 3 for end-to-end learning of the appropriate refinement for the best identification performance before statistical pooling. In this architecture, four 1D-convolutional layers are applied to the input time series. By using max-pooling layers across time, the receptive field of the final layer of the stack can be increased. Following the convolutional layers, a linear mapping is learned to map the output of the last convolutional layer to the feature-embedding space. After calculation of the mean and standard deviation of the feature-embeddings across time, the resulting fixed-length vector is then used for generating identity embeddings by two fully-connected layers. Instead of using class labels for training the network, we use triplet loss [SKP15] to enable better generalization capacity for unseen identities. Furthermore, batch normalization is used after the input layer, the statistical pooling layer, and between the output of

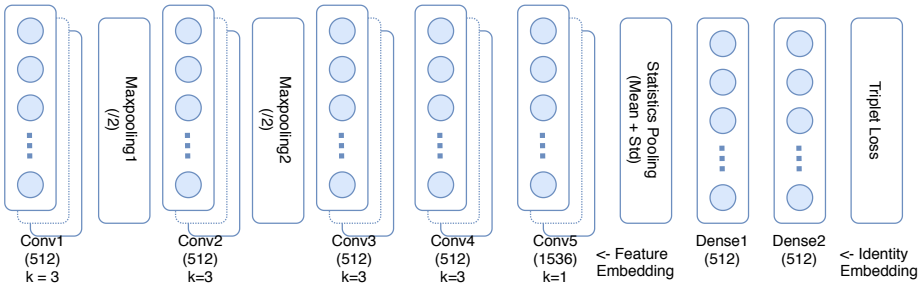

Fig. 3: Proposed network architecture.

neurons and activation functions to reduce the learning time of the network. No activation is used on the output of the feature-embedding mapping layer and the final layer to enable the network to utilize the full embedding space.

The Euclidean distance between identity embeddings can directly be used as a biometric dissimilarity metric. In the case of multiple enrollment samples from multiple identities, it is also possible to use the proposed system as a preprocessing step, and train a softmax layer for classification directly on extracted identity embeddings.

2.3 Look-alike mining

The VoxCeleb2 dataset [CNZ18] contains over 1 million utterances from more than $6,000$ celebrities collected from YouTube. The identities in this dataset are a subset of identities in the VGGFace2 [Ca18] dataset. To mine for Look-alike identities, we used the ArcFace [De19] face recognition system to compare the average embeddings for each identity in the VGGFace2 dataset that appears in VoxCeleb2 dataset as well. After sorting the scores of the resulting 36M comparison pairs, the top $2,000$ pairs with the highest similarity score are selected for a subjective face recognition test. Among the top pairs, there exist pairs of identical twins as well.

In the subjective face recognition test, for each look-alike pair of identities, four images are selected from each identity from the VGGFace2 dataset and shown to participants. The task for the participants was to check whether the two sets of images correspond to the same identity or two different people. The user interface is shown in Figure 4. Due to the large number of comparisons, the test was done by 20 participants, each labeling 200 pairs such that each pair is labeled by two people. From the resulting comparisons, the pairs that were labeled as the same people by at least one participant were selected as look-alikes and formed the 1000 look-alike pairs (1000LP) dataset. Figure 1 shows examples of the resultant look-alike pairs. To assure the reliability of the selected look-alike pairs, the equal-error-rate (EER) is calculated for the resulting look-alike pairs using the ArcFace network, resulting in an unacceptably high EER of 30.32%.

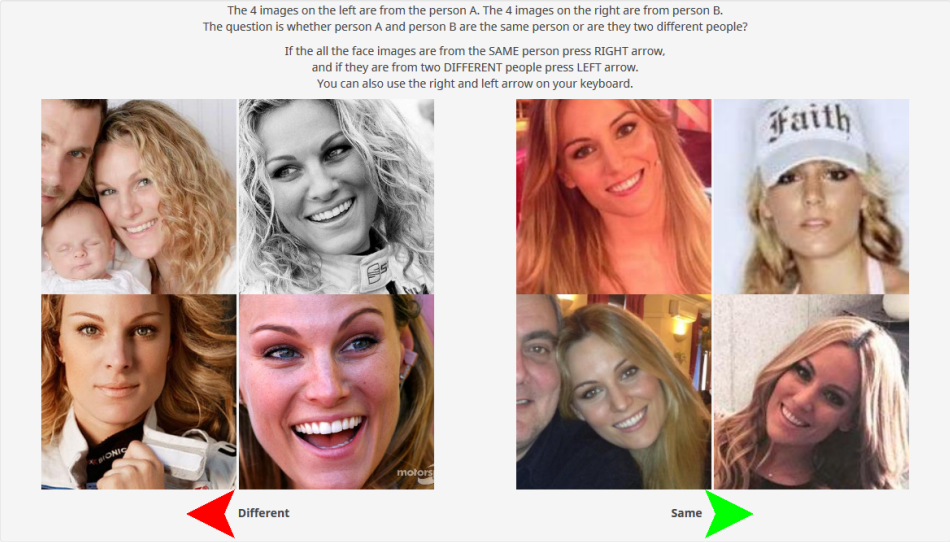

Fig. 4: Subjective face recognition test user interface.

3 Experiment Setup

The selected 1000 pairs of look-alikes consist of 1634 unique identities. The remaining 4500 identities in VoxCeleb2 are available for training the network. The rest of this section describes the details of the organized test dataset and the parameters used for training.

3.1 1000LP Dataset

The utterances available in the VoxCeleb2 dataset are in the format of cropped faces sized 224×224 pixels at 25 frames per second in AVC1 format. There is a total of $1,128,246$ utterances which originate from $150,480$ YouTube videos. After filtering out all utterances with a length of less than 8 seconds and discarding all utterances for which face landmark detection failed, a total of $253,361$ utterances remained for training and $85,656$ utterances for testing. The median length of the remaining utterances is 10.7 seconds. From the 4500 train identities, 15% of them were held for validation purposes, and the remaining were used for training. For the test identities, one-third of videos ($28,368$ utterances) were separated for enrollment, and the remaining videos were used for testing. Resulting from this, $127,332$ test trials were created, out of which $57,288$ are client trials and $70,044$ are impostor trials[5]. Special care is taken in the selection of the enrollment and test utterances such that if an utterance from a YouTube video is used in the enrollment, no utterances from the same video remains in the test trials. Thus, the performance is assured to correspond to the cross-video performance in real-life use.

[5] The dataset is publicly available for download at `http://ali.khodabakhsh.org/research/1000lp/`

		Verification	Identification	
		EER (%)	Top-1 (%)	Top-5 (%)
Euclidean	Segment (∼10 sec)	15.42	60.57	77.83
Distance	Video (∼4 seg)	13.04	**79.84**	**92.61**
Softmax	Segment (∼10 sec)	10.08	65.47	81.00
Classifier	Video (∼4 seg)	**7.93**	73.87	86.33

Tab. 1: The performance of the proposed methods.

3.2 Detector

The network parameters are shown in Figure 3. The breadth of the network along with the dimension of the final embedding is set to 512, with only the exception of expanded feature embedding dimensions of three times the breadth. The total number of trainable parameters in the network was 5.3M. A kernel size of 3 is used in the convolutional stack while max-pooling is done with a stride of two, resulting in a receptive field of 23 frames (roughly one second) before statistical pooling. The normalized input had a dimension of 204 corresponding to 3D coordinates for the 68 landmarks. The model was trained using TensorFlow[6] with a batch size of 256 and the learning rate was manually adjusted towards minimizing validation loss. Semi-hard triplet loss on L2 distance of L2 normalized network outputs was used and the model was trained for 10 epochs. The hyper-parameters are selected according to the highest network performance on validation data.

4 Results and Discussion

The verification and identification performance of the proposed method for Euclidean similarity as well as softmax probabilities are reported in Table 1. The Euclidean similarity scoring performs better in identification mode than softmax probabilities and achieves an identification accuracy of 79.84% on video level. This is remarkable considering the large number of identities enrolled in the system (1634). Despite the high identification accuracy, the EER of the Euclidean similarity measure is 13.04%. Softmax probabilities, however, achieve a much better EER of 7.93% in verification mode. This discrepancy shows that softmax probabilities perform better in separating score distributions of client and impostor trials, but fails to preserve the ranking order of similarities. The detection error tradeoff (DET) curve is shown in Figure 5 visualizing the fact.

In order to be able to interpret the performance of the proposed method, it is compared to the reported results for existing BFR methods in the literature in Table 2. It is important to emphasize that all previous methods have only been tested on videos with controlled and semi-controlled recording environments. Among the methods that operate on non-predetermined motion, the proposed method has the lowest EER and a comparable recognition rate despite the number of enrolled identities being orders of magnitude larger.

[6] https://www.tensorflow.org/

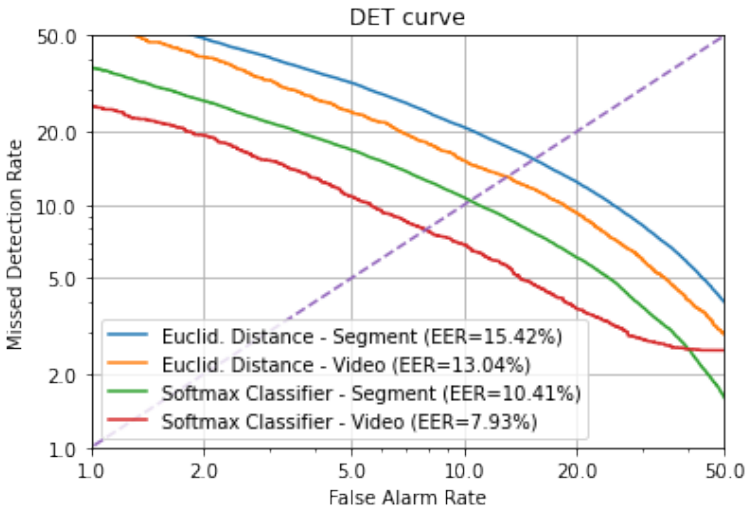

Fig. 5: Detection error tradeoff (DET) curve for the proposed methods.

Ref.	Subj. #	Environment	Motion	Feature	Classifier	Perf.	Metric
[LLJ01]	28	Controlled	Fixed-Phrase Speech	Motion Flow Fields	PCA + LDA	∼87%	Recog. Rate
[Ce06]	50	Controlled	Fixed-Phrase Speech	Grid-based Motion Contour-based Motion Lip Shape	LDA + Bayes	5.2% 12.0% 10.4%	EER
[ZP11]	22	Controlled	Spontaneous Smile	Motion Fields	PCA + LDA	2.5%	EER
[WL12]	40	Controlled	Fixed-Phrase Speech	Lip Shape Deformation Lip Texture Deformation	HMM-UBM	1.92% 8.53%	EER
[Ga16]	64	Controlled	Induced Emotion	Facial Action Units	MLP	91.7%	Percision
[Ie19]	20	Controlled	Fixed-Phrase Speech	Facial Landmarks	DNN	0.64%	EER
[Ta19]	400	Controlled	Spontaneous Smile	Facial Landmark Distances	Euclid. Dist.	31.20%	EER
[MD06]	9	Controlled	Unconstrained Speech	Facial Feature Displacement	GMM	19.1%	EER
[YS10]	97	Controlled	Unconstrained Emotion	Local Deformation Patterns	Similarity	18.86%	EER
[Sh16]	96	Ambiguous	Unconstrained Speech	Facial Action Units	Hist. Sim. DTW	∼62%	Recog. Rate
[Yu17]	20	Ambiguous	Unconstrained Speech	Lip Contour	GMM	96.2%	Recog. Rate
[Ag19]	Clinton Sanders Trump Warren	Ambiguous	Unconstrained Speech	Facial Action Units	SVM	75% 95% 77% 95%	TPR @ 10% FPR
Proposed	1634	Unconstrained	Unconstrained Speech	Facial Landmarks	CDNN	7.93% 79.84% 93.12%	EER Recog. Rate TPR@ 10% FPR

Tab. 2: The performance of the proposed method in contrast to the reported results for existing methods.

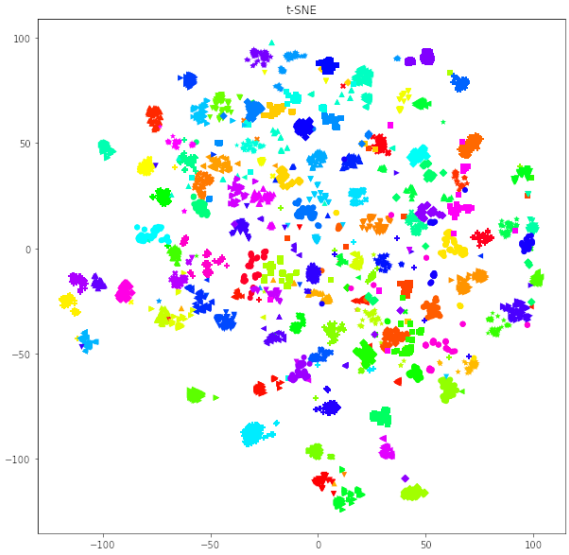

Fig. 6: t-distributed stochastic neighbor embedding for enrollment utterances. For aesthetic reasons, only the identities with more than 50 enrollment utterances are visualized. Different colors and shapes signify different identities.

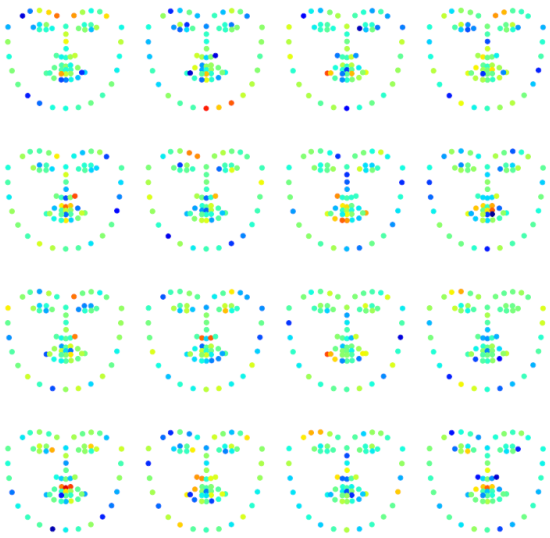

Fig. 7: Facial landmark significance visualization for selected filters in conv1. The significance is measured as the norm of the 3×3 matrix corresponding to x, y, and z coordinates of the landmark in frames $t-1$, t, and $t+1$.

The t-distributed stochastic neighbor embeddings (t-SNE) [MH08] for enrollment utterances for a subset of identities is visualized in Figure 6. It is visible that the enrollment utterances of test set identities form concentrated clusters with few outliers. This signifies that the learned embedding space is able to generalize well across unseen identities, and the failure cases probably correspond to the outliers. Figure 7 shows landmark significance for a selected set of filters in the first convolutional layer of the network. The significance is measured in terms of the norm of the 3×3 matrix corresponding to multiplicative weights in x, y, and z coordinates of each landmark in frames $t-1$, t, and $t+1$. These heatmaps show the reliance of the network on meaningful facial actions such as eyebrow movements, upper lip movements, and movements in the corners of the mouth.

The results of this study show the power of large data in improving the performance and generalizability of BFR systems. Even though this system is trained on 4500 identities, the number of training identities is still much smaller compared to physiological face recognition systems, and there is room for further improvement.

5 Conclusion

In this article, we proposed a novel general-purpose action-independent behavioral identity embedding extraction network with acceptable performance for real-life applications. The network benefits from a large number of training samples and identities and proves capable of extracting descriptive embeddings for unseen identities in unconstrained conditions. We also respond to the lack of publicly available large-scale datasets for look-alike detection, as well as publicly available behavioral face recognition systems by releasing the 1000 look-alike pairs (1000LP) dataset and the code for the proposed method.

The proposed method provides a complementary source of identity information that can be used alongside physiological face recognition systems to make them robust against look-alikes, as well as presentation attacks that try to mimic the physiological likeliness. The proposed method is robust to physical and digital spatial signal manipulations as it relies solely on the temporal behavior of the individual in question. Due to the permanence of behavioral face biometrics [Be10] and its robustness to manipulations and quality degradation, these methods have already found their way into the detection of Deepfakes [Ag19] and can provide a robust alternative to existing narrowly applicable detection methods [Kh18].

References

[Ag19] Agarwal, Shruti; Farid, Hany; Gu, Yuming; He, Mingming; Nagano, Koki; Li, Hao: Protecting World Leaders Against Deep Fakes. In: CVPR Workshops. June 2019.

[Be10] Benedikt, L.; Cosker, D.; Rosin, P. L.; Marshall, D.: Assessing the Uniqueness and Permanence of Facial Actions for Use in Biometric Applications. IEEE SMCS, 2010.

[BRM16] Baltrušaitis, T.; Robinson, P.; Morency, L.: OpenFace: An open source facial behavior analysis toolkit. In: WACV. pp. 1–10, 2016.

[Ca18] Cao, Q.; Shen, L.; Xie, W.; Parkhi, O. M.; Zisserman, A.: VGGFace2: A Dataset for Recognising Faces across Pose and Age. In: FG. pp. 67–74, 2018.

[Ce06] Cetingul, H. E.; Yemez, Y.; Erzin, E.; Tekalp, A. M.: Discriminative Analysis of Lip Motion Features for Speaker Identification and Speech-Reading. IEEE TIPS, 2006.

[CNZ18] Chung, Joon Son; Nagrani, Arsha; Zisserman, Andrew: VoxCeleb2: Deep Speaker Recognition. CoRR, abs/1806.05622, 2018.

[De19] Deng, J.; Guo, J.; Xue, N.; Zafeiriou, S.: ArcFace: Additive Angular Margin Loss for Deep Face Recognition. In: CVPR. pp. 4685–4694, 2019.

[Ga16] Gavrilescu, M.: Study on using individual differences in facial expressions for a face recognition system immune to spoofing attacks. IET Biometrics, 5(3):236–242, 2016.

[Ie19] Iengo, D.; Nappi, M.; Ricciardi, S.; Vanore, D.: Dynamic Facial Features for Inherently Safer Face Recognition. In: ICIP. pp. 2611–2615, 2019.

[Kh18] Khodabakhsh, A.; Ramachandra, R.; Raja, K.; Wasnik, P.; Busch, C.: Fake Face Detection Methods: Can They Be Generalized? In: BIOSIG. pp. 1–6, 2018.

[KJ97] Knight, Barbara; Johnston, Alan: The Role of Movement in Face Recognition. Visual Cognition, 4(3):265–273, 1997.

[KKR16] Kim, S. T.; Kim, D. H.; Ro, Y. M.: Facial dynamic modelling using long short-term memory network: Analysis and application to face authentication. In: BTAS. 2016.

[KPJ11] Klare, B.; Paulino, A. A.; Jain, A. K.: Analysis of facial features in identical twins. In: IJCB. pp. 1–8, 2011.

[KRB19] Khodabakhsh, A.; Ramachandra, R.; Busch, C.: Subjective Evaluation of Media Consumer Vulnerability to Fake Audiovisual Content. In: QoMEX. pp. 1–6, 2019.

[La11] Lamba, H.; Sarkar, A.; Vatsa, M.; Singh, R.; Noore, A.: Face recognition for look-alikes: A preliminary study. In: IJCB. pp. 1–6, 2011.

[LLJ01] Li-Fen Chen; Liao, H. . M.; Ja-Chen Lin: Person identification using facial motion. In: ICIP. volume 2, pp. 677–680 vol.2, 2001.

[MD06] Matta, F.; Dugelay, J.: A Behavioural Approach to Person Recognition. In: 2006 IEEE International Conference on Multimedia and Expo. pp. 1461–1464, 2006.

[MH08] Maaten, Laurens van der; Hinton, Geoffrey: Visualizing data using t-SNE. Journal of machine learning research, 9(Nov):2579–2605, 2008.

[PD17] Pan, S.; Deravi, F.: Facial action units for presentation attack detection. In: EST. pp. 62–67, 2017.

[Ph11] Phillips, P. J.; Flynn, P. J.; Bowyer, K. W.; Bruegge, R. W. V.; Grother, P. J.; Quinn, G. W.; Pruitt, M.: Distinguishing identical twins by face recognition. In: Face and Gesture 2011. pp. 185–192, 2011.

[Sh16] Shreve, M.; Bernal, E. A.; Li, Q.; Kumar, J.; Bala, R.: A study on the discriminability of facs from spontaneous facial expressions. In: ICIP. pp. 1674–1678, 2016.

[SKP15] Schroff, Florian; Kalenichenko, Dmitry; Philbin, James: FaceNet: A Unified Embedding for Face Recognition and Clustering. In: CVPR. June 2015.

[Sn18] Snyder, D.; Garcia-Romero, D.; Sell, G.; Povey, D.; Khudanpur, S.: X-Vectors: Robust DNN Embeddings for Speaker Recognition. In: ICASSP. pp. 5329–5333, 2018.

[Ta19] Taskirar, M.; Killioglu, M.; Kahraman, N.; Erdem, C. E.: Face Recognition Using Dynamic Features Extracted from Smile Videos. In: INISTA. pp. 1–6, 2019.

[WL12] Wang, Shi-Lin; Liew, Alan Wee-Chung: Physiological and behavioral lip biometrics: A comprehensive study of their discriminative power. Pattern Recognition, 2012.

[YS10] Ye, N.; Sim, T.: Towards general motion-based face recognition. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. pp. 2598–2605, 2010.

[Yu17] Yuan, Y.; Zhao, J.; Xi, W.; Qian, C.; Zhang, X.; Wang, Z.: SALM: Smartphone-Based Identity Authentication Using Lip Motion Characteristics. In: SMARTCOMP. 2017.

[Zh14] Zhang, Li; Ma, KengTeck; Nejati, Hossein; Foo, Lewis; Sim, Terence; Guo, Dong: A talking profile to distinguish identical twins. Image and Vision Computing, 2014.

[ZP07] Zhao, G.; Pietikainen, M.: Dynamic Texture Recognition Using Local Binary Patterns with an Application to Facial Expressions. IEEE TPAMI, 29(6):915–928, 2007.

[ZP11] Zafeiriou, S.; Pantic, M.: Facial behaviometrics: The case of facial deformation in spontaneous smile/laughter. In: CVPR WORKSHOPS. pp. 13–19, 2011.