Exploring Deep Features with Different Distance Measures for Still to Video Face Matching

Yu Zhu and Guodong Guo(✉)

Lane Department of Computer Science and Electrical Engineering,
West Virginia University, Morgantown, WV 26506, USA
yzhu4@mix.wvu.edu, guodong.guo@mail.wvu.edu

Abstract. Still to video (S2V) face recognition attracts many interests for researchers in computer vision and biometrics. In S2V scenarios, the still images are often captured with high quality and cooperative user condition. On the contrary, video clips usually show more variations and of low quality. In this paper, we primarily focus on the S2V face recognition where face gallery is formed by a few still face images, and the query is the video clip. We utilized the deep convolutional neural network to deal with the S2V face recognition. We also studied the choice of different similarity measures for the face matching, and suggest the more appropriate measure for the deep representations. Our results for both S2V face identification and verification yield a significant improvement over the previous results on two databases, i.e., COX-S2V and PaSC.

1 Introduction

Face recognition has been attracting researchers' attentions for several decades in computer vision and biometrics. With the fast development of video surveillance systems and low cost video cameras, video based face recognition becomes an important topic in face recognition [2,3,5,11,12,16]. In practice, face recognition systems usually have still face images in gallery and the probe faces come from videos. In such scenario, it is necessary to study the problem of automatically matching video faces and still faces, a.k.a., Still-to-Video (S2V) face recognition. In this work, we focus on the S2V face recognition where each subject is enrolled with only few still images, while the query is coming from video clips which consist of uncontrolled image frames. One of the major challenges for S2V face recognition lies in the disparity between still images and video clips. Faces captured by video cameras usually show more variations in illumination, head pose, expression, or motion blur, while the still images could be with a high quality.

Recently, several approaches are proposed for S2V face recognition [1–4,9,10,15,20]. One way is to apply still face recognition methods to the S2V scenario [1,9,15]. Another popular methodology for S2V face recognition is by using metric learning, e.g., Neighborhood Components Analysis (NCA) [7], Information Theoretic Metric Learning (ITML) [6], Local Fisher Discriminant Analysis (LFDA) [13] and Large Margin Nearest Neighbor (LMNN) [17]. More recently, point to set metric learning methods, such as Point-to-Set Distance

Z. You et al. (Eds.): CCBR 2016, LNCS 9967, pp. 158–166, 2016.
DOI: 10.1007/978-3-319-46654-5_18

Metric Learning (PSDML) [19] and Learning Euclidean-to-Riemannian Metric (LERM) [10], have been proposed. Moreover, in [20], the authors considered the S2V face recognition as a heterogeneous problem.

However, the above mentioned previous works utilized either gray level features or low level features in their approaches, which to some extent impact the overall face recognition performance. Especially for the still to video face recognition where great variations occurs in video faces, low level features are not sufficient. Therefore, we study the performance of still to video face recognition using deep representations, which has been established as a powerful model for many recognition problems. Our proposed method is the first time that applying deep representations to the Still to Video (S2V) face recognition problem, to the best of our knowledge.

Moreover, classification of faces often involves identifying samples that are close or similar to each other. Distances or similarities are mathematical representations of what are defined by close or similar. Since the choice of distance measures is a critical for S2V face recognition, we also study the choice of distances or similarities when deep features is used.

The rest of this paper is organized as following: The proposed method and distance measures are described in Sect. 2. Experimental results on two databases with comparisons are shown in Sect. 3. Finally, we draw some conclusions in Sect. 4.

2 Approach

2.1 Deep Neural Network

Recently, deep learning methods especially the deep convolutional neural networks (DCNN) have been applied to many computer vision tasks with promising performance, such as object recognition [14], still image based face recognition [18], etc. However, it has not been well studied yet on using deep learning techniques for still to video face matching. In this study, we explore the performance of deeply learned face representations (features) for still to video face recognition. Then, we further studied different distance measures for deep features on S2V face matching.

In exploring the deep features for still to video face matching, we hope that the trained deep network is capable of capturing the face structures with discriminative power, at the same time being robust to the variations, such as head poses, illuminations and blurs. Accordingly, it usually requires a large number of training data for the deep network. However, the size of public S2V databases is usually small, e.g., in PaSC database, there are only 2802 videos collected from 293 subjects. Therefore, we propose to utilize a large number of still face images to train a deep network, e.g., from the WebFace database [18]. Besides, we utilized a much deeper network architecture so that the deep features are expected to be more robust and representative. We applied the GoogLeNet [14], which contains 22-layers with an "inception" structure, to train our deep face model. The main idea of "inception" structure that was included in the GoogLeNet is a combination of network-in-network filters, along with different convolutional

filters in each layer. Within the inception module, a 3×3 filter, a 5×5 filter and a max pooling are performed and then combined together. In this way, there are multiple filter sizes per layer so each layer can have the ability to character different "feature resolutions" that may have in its input. Another property of the GoogLeNet is that auxiliary classifiers are added to connect to intermediate layers. The total loss of the network during training is a weighted sum of the auxiliary classifiers.

2.2 Distance Measures for Face Matching

In our approach, matching of still images and video clips is done by fusing the deep feature distances between still images and video frames in each video clip. Note that, the training data provided with the database are not used in our study. Deep features extracted from the deep model are used directly for face matching without any fine-tuning. There are different measures for the distance between two deep features, however, there is no systematic study or evaluation of different measures. In the following, we study six different distance measures for measuring the similarity between deep features.

Euclidean Distance. Among various distance metrics, the Euclidean distance is one of the most commonly used metric due to its simplicity. The Euclidean distance between 1-D arrays u and v, is defined as:

$$d\,(u, v) = ||u - v||_2, \tag{1}$$

where $||u - v||_2 = \sqrt{(u - v) \cdot (u - v)}$.

Manhattan Distance. The Manhattan distance is the L1-norm of the difference, which is also a special case of the Minkowski distance (with p=1). The Manhattan distance (a.k.a., City Block distance) between 1-D arrays u and v, is defined as:

$$d\,(u, v) = \Sigma_i |u_i - v_i|. \tag{2}$$

In other words, the Manhattan distance computes the sum of absolute differences between two vectors.

Canberra Distance. The Canberra distance between 1-D arrays u and v, is defined as:

$$d\,(u, v) = \Sigma_i \frac{|u_i - v_i|}{|u_i| + |v_i|}. \tag{3}$$

When u_i and v_i are 0 for given i, then the fraction is set to 0. The Canberra is similar to the Manhattan distance, with the distinction that prior to summation, the absolute difference between the variables of the two vectors is divided by the sum of the absolute variable values. It is thus more sensitive to proportional than to absolute differences.

Correlation Distance. Correlation distance is another widely used measures for data analysis, which defined by subtracting the correlation coefficient from 1. Specifically, the Correlation distance between 1-D arrays u and v, is defined as:

$$d\left(u,v\right)=1-\frac{\left(u-\bar{u}\right)\cdot\left(v-\bar{v}\right)}{\left|\left|\left(u-\bar{u}\right)\right|\right|_2\left|\left|\left(v-\bar{v}\right)\right|\right|_2}, \tag{4}$$

where $\bar{u}$ is the mean of the elements of u, and $\cdot$ is the dot product operation.

BrayCurtis Distance. The computation of BrayCurtis distance involves summation of the absolute differences between the variables and dividing this by the sum of the variables in the two vectors. Mathematically, the BrayCurtis distance between 1-D arrays u and v, is defined as:

$$d\left(u,v\right)=\frac{\Sigma|u_i-v_i|}{\Sigma|u_i+v_i|}. \tag{5}$$

Bray-Curtis distance can be considered as a modified version of Manhattan distance, where the summed differences between the vectors are weighed by the sum of corresponding vector components.

Cosine Distance. The cosine distance is a measure that calculates the cosine of the angle between two vectors. The cosine distance between 1-D arrays u and v, is defined as:

$$d\left(u,v\right)=1-\frac{u\cdot v}{||u||_2||v||_2}, \tag{6}$$

where $u\cdot v$ is the dot product of u and v. Cosine distance between two vectors can be seen as a comparison on a normalized space, since the magnitude is not considered but the angle is used to measure how related between the two vectors.

3 Experimental Results

3.1 Databases

COX-S2V Database. COX-S2V database [9] was collected with both still images and video clips. The faces in this database contain large variations, such as illumination, head pose, and motion blurs. Totally there are 1,000 subjects in this database. For each subject, there is one still image with high resolution, and three video clips corresponding to three different camera locations, denoted as Cam1, Cam2 and Cam3, respectively.

Point and Shoot Face Recognition Challenge (PaSC) Database. The PaSC database [1] was collected for the point and shoot face recognition challenge. For both still images and video clips in this database, the faces show different variations such as head pose, background locations, motion blur and poor focus. There are 9,376 still images and 2,802 video clips that have been collected from 293 subjects in this database.

3.2 Experimental Settings

On the COX-S2V database, we follow the same experimental protocol in [9], where face identification is performed between still images and three camera settings, respectively. On the PaSC database, we followed the same protocol in [1], for the face verification task. We use CASIA-WebFace [18] to train the deep network for our S2V face matching. CASIA-WebFace is a large database including about 10,595 subjects and 494,414 images.

On the COX-S2V database, 700 subjects with their still images and video clips formed the test set. The remaining 300 subjects' data are used for training. The experiments ran 10 times with randomly selected gallery/probe combinations. The averaging recognition rates are used as the performance measurement on this database.

On the PaSC database, there are 2872 still images and 280 video clips for training. The 4688 still images formed the target set, while 1401 handheld video clips is the query set in testing. Therefore, the verification is conducted based on the similarity matrix of size 4688×1401. The ROC curve and verification rate (when FAR equals to 0.01) are used as the performance measurements.

Experimental Settings for Deep Network. Firstly, face detection and landmark localization are applied on both COX-S2V and PaSC database. The detected faces are then cropped and aligned to the size of 256×256.

The training of GoogLeNet is implemented using the Caffe toolbox [14] with the WebFace database. We used the mini-batch SDG (Stochasitc gradient descent) with momentum settings (set to 0.9) and the batch size is set to 32. The learning rate is set to 0.01 with decrease in polynomial decay (power of 0.5). The training procedure stops after 2,400,000 iterations.

During the test, the aligned face images from both gallery and probe sets are fed into deep network, we used the layer (*pool5/7x7_s1*) to get the face deep features. Each face image is then represented as the deep feature vector of size 1024. Matching is conducted according to the protocols described above. All experiments are tested on the 64 bit Ubuntu 14.04 platform with 6-Core i7 CPU, 32 G RAM and Titan X GPU.

3.3 Experimental Results on the COX-S2V Database

The experimental results on COX-S2V are shown in Table 1. From the table one can see that, the Euclidean distance gets the worst performance in all three experiments (i.e., Cam 1, 2 and 3), the recognition rates are 30.97 %, 34.53 % and 54.50 %, respectively. One of the reasons might lie in the fact that it suffers from a high sensitivity even to small deformation. For the high dimension (i.e., 1024) deep features, a small deformation may result in a large Euclidean distance change. The Manhanttan and Canberra distances show better performance then the Euclidean distance, but still quite lower than the other three distances. The performance obtained by Cosine distance is slightly higher than the results using correlation distance, but is lower than the BrayCurtis distance. The best

Table 1. The experimental results for S2V face identification on COX-S2V database, and face verification on PaSC database.

Method	COX-S2V			PaSC	
	Still-Video1	Still-Video2	Still-Video3	VR(FAR = 0.01)	VR (FAR = 0.01)
Euclidean	30.97 ± 0.75	34.53 ± 0.86	54.50 ± 1.36	0.04	0.52
Manhattan	36.43 ± 2.38	44.57 ± 4.02	68.93 ± 2.11	0.02	0.48
Canberra	55.31 ± 1.03	65.50 ± 1.38	77.29 ± 0.84	0.08	0.67
Correlation	74.37 ± 1.17	79.61 ± 1.05	94.90 ± 0.64	0.58	**0.82**
Cosine	74.81 ± 1.22	79.76 ± 0.91	95.04 ± 0.59	0.60	0.80
Braycurtis	**77.29±0.84**	**83.10±1.09**	**96.17±0.74**	**0.61**	0.79

Table 2. Comparison with other approaches for S2V face identification on COX-S2V database.

Method	COX-S2V		
	Still-Video1	Still-Video2	Still-Video3
NNC [8]	9.96 ± 0.61	7.14 ± 0.68	17.37 ± 6.16
NCA [7]	39.14 ± 1.33	31.57 ± 1.56	57.57 ± 2.03
LMNN [17]	34.44 ± 1.02	30.03 ± 1.36	58.06 ± 1.35
LERM [10]	45.71 ± 2.05	42.80 ± 1.86	58.37 ± 3.31
GFK [20]	49.86 ± 1.22	42.99 ± 2.17	69.81 ± 1.72
Ours	**77.29±0.84**	**83.10±1.09**	**96.17±0.74**

performance is achieved by the BrayCurtis distance, which gives the recognition rates 77.29 %, 83.10 % and 96.17 % for Cam1, 2 and 3, respectively. These results show that, when the deep feature is used, the choice of distance measures impacts the S2V face recognition performance significantly.

Next we compare the results to other state-of-the-art methods on the COX-S2V database. The results are shown in Table 2. The gray level features were utilized to those methods according to the original papers. From the table one can see that, our proposed approach achieved better results than the other listed methods, where the recognition rates are 77.29 %, 83.10 % and 96.17 %, respectively. The comparisons suggest that deep features are more appropriate for still-to-video face matching.

3.4 Experimental Results on PaSC Database

On the PaSC database, we firstly show the verification results using different distance measures in Table 1. We listed the verification rate (VR) then FAR equals to 0.01 and 0.1, in column 5 and 6, respectively. From the table, one can see that Correlation, Cosine and BrayCurtis distances shows much better performances than the other three distances: Euclidean, Manhattan, and Canberra. The best

Table 3. The experimental results (verification rate) for S2V face verification on PaSC dataset, when FAR equals to 0.01.

Method	Verification rate	Method	Verification rate
NNC [8]	0.05	PLDA-WPCA-LLR [1]	0.26
NCA [7]	0.16	Eigen-PEP [1]	0.24
LMNN [17]	0.17	LPB-SIFT-WPCA-SILD [1]	0.23
LERM [10]	0.17	ISV-GMM [1]	0.11
GFK [20]	0.22	LRPCA [1]	0.10
Ours	**0.61**		

VR (FAR = 0.01) achieves the value 0.61 by BrayCurtis distance. These observations illustrates that, when deep features are utilized as the face representation, different distance measures perform quite differently. The selection of distance measures effects the matching results significantly.

Next we compared our results to the other state-of-the-art approaches on the PaSC dataset in Table 3. The results are shown the verification rate when FAR equals to 0.01, which is the same for the PaSC challenge [1]. In Table 3, it shows that deep feature obtains the best verification rate of 0.61, using the BrayCurtis distance. One can also see that our result is significantly better than the other approaches in the PaSC challenge from [1]. Similar observations can be found in the ROC curves shown in Fig. 1 This further demonstrates the robustness and effectiveness of our method for still to video face recognition.

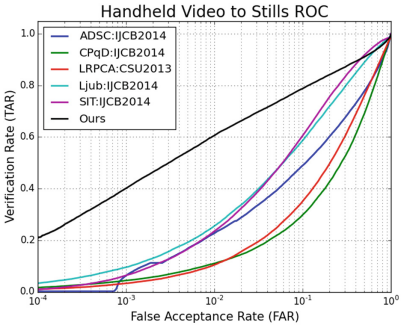

Fig. 1. ROC of the PaSC dataset for S2V face verification.

4 Conclusions

We proposed to utilize the face representations from deep convolutional neural networks dealing with the still to video (S2V) face recognition. We found deeply learned features provides more robustness and discriminative power for face

matching, which is a more appropriate way for the S2V face recognition. We further studied the choice of different similarity measure and show the impact to the final recognition performance. Experimental results on two still to video face databases illustrate that, the deep features achieved significantly better performance in both verification and identification, than the other state-of-the-art approaches. Besides, the recognition performance using deep features varies according to different choices of the similarity measures. Among the six different similarity measures, the BrayCurtis measure is considered more appropriate for the S2V.

References

1. Beveridge, J.R., Phillips, P.J., Bolme, D.S., Draper, B.A., Givens, G.H., Lui, Y.M., Teli, M.N., Zhang, H., Scruggs, W.T., Bowyer, K.W., et al.: The challenge of face recognition from digital point-and-shoot cameras. In: IEEE Biometrics: Theory, Applications and Systems (BTAS), pp. 1–8 (2013)
2. Beveridge, J.R., Zhang, H., Draper, B.A., Flynn, P.J., Feng, Z., Huber, P., Kittler, J., Huang, Z., Li, S., Li, Y., et al.: Report on the fg 2015 video person recognition evaluation. In: 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), vol. 1, pp. 1–8. IEEE (2015)
3. Chen, X., Wang, C., Xiao, B., Cai, X.: Scenario oriented discriminant analysis for still-to-video face recognition. In: IEEE International Conference on Image Processing (ICIP), pp. 738–742 (2014)
4. Chen, X., Wang, C., Xiao, B., Zhang, C.: Still-to-video face recognition via weighted scenario oriented discriminant analysis. In: IEEE International Joint Conference on Biometrics (IJCB), pp. 1–6 (2014)
5. Cui, Z., Chang, H., Shan, S., Ma, B., Chen, X.: Joint sparse representation for video-based face recognition. Neurocomputing **135**, 306–312 (2014)
6. Davis, J.V., Kulis, B., Jain, P., Sra, S., Dhillon, I.S.: Information-theoretic metric learning. In: Proceedings of the 24th International Conference on Machine Learning, pp. 209–216. ACM (2007)
7. Goldberger, J., Hinton, G.E., Roweis, S.T., Salakhutdinov, R.R.: Neighbourhood components analysis. In: Advances in Neural Information Processing Systems, pp. 513–520 (2005)
8. Gong, B., Shi, Y., Sha, F., Grauman, K.: Geodesic flow kernel for unsupervised domain adaptation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2066–2073 (2012)
9. Huang, Z., Shan, S., Zhang, H., Lao, S., Kuerban, A., Chen, X.: Benchmarking still-to-video face recognition via partial and local linear discriminant analysis on COX-S2V dataset. In: Lee, K.M., Matsushita, Y., Rehg, J.M., Hu, Z. (eds.) ACCV 2012. LNCS, vol. 7725, pp. 589–600. Springer, Heidelberg (2013). doi:10.1007/978-3-642-37444-9_46
10. Huang, Z., Wang, R., Shan, S., Chen, X.: Learning euclidean-to-riemannian metric for point-to-set classification. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1677–1684 (2014)
11. Kim, T.-K., Kittler, J., Cipolla, R.: Discriminative learning and recognition of image set classes using canonical correlations. IEEE Trans. Pattern Anal. Mach. Intell. **29**(6), 1005–1018 (2007)

12. Liu, X., Chen, T.: Video-based face recognition using adaptive hidden markov models. In: IEEE Computer Vision and Pattern Recognition, vol. 1, p. I-340 (2003)
13. Sugiyama, M.: Dimensionality reduction of multimodal labeled data by local fisher discriminant analysis. J. Mach. Learn. Res. **8**, 1027–1061 (2007)
14. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–9 (2015)
15. Wang, H., Liu, C., Ding, X.: Still-to-video face recognition in unconstrained environments. In: IS&T/SPIE Electronic Imaging, p. 94050O. International Society for Optics and Photonics (2015)
16. Wang, R., Guo, H., Davis, L.S., Dai, Q., Covariance discriminative learning: a natural and efficient approach to image set classification. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2496–2503 (2012)
17. Weinberger, K.Q., Blitzer, J., Saul, L.K.: Distance metric learning for large margin nearest neighbor classification. In: Advances in Neural Information Processing Systems, pp. 1473–1480 (2005)
18. Yi, D., Lei, Z., Liao, S., Li, S.Z.: Learning face representation from scratch. arXiv preprint arXiv:1411.7923 (2014)
19. Zhu, P., Zhang, L., Zuo, W., Zhang, D.: From point to set: extend the learning of distance metrics. In: IEEE International Conference on Computer Vision, pp. 2664–2671 (2013)
20. Zhu, Y., Zheng, Z., Li, Y., Mu, G., Shan, S., Guo, G.: Still to video face recognition using a heterogeneous matching approach. In: IEEE Biometrics: Theory, Applications and Systems (BTAS) (2015)