SCENARIO ORIENTED DISCRIMINANT ANALYSIS FOR STILL-TO-VIDEO FACE RECOGNITION

Xue Chen, Chunheng Wang, Baihua Xiao, Xinyuan Cai

State Key Laboratory of Management and Control for Complex Systems,
Institute of Automation, Chinese Academy of Sciences

ABSTRACT

In the Still-to-Video (S2V) face recognition, each subject is enrolled with only few high resolution images, while the probe is video clips of complex variations. As faces present distinct characteristics under different scenarios, recognition in the original space is obviously inefficient. Therefore, in this paper, we propose a novel discriminant analysis method to learn separate mappings for different scenario patterns (still, video), and further pursue a common discriminant space for the cross-scenario samples. To maximize the intra-individual correlation of samples in the mapping space, we formulate the learning objective by incorporating the intra-class compactness and the inter-class dispersion. The gradient descend algorithm is used to get the optimal solution. Experimental results on the COX-S2V dataset demonstrate the effectiveness of the proposed method and remarkable superiority over state-of-art methods.

***Index Terms*—** face recognition, Still-to-Video, discriminant analysis

1. INTRODUCTION

In the real-world Still-to-Video (S2V) face recognition scenario, only very few (single, in many cases) still images per person are enrolled into the gallery while multiple video clips are captured as the probe. Generally, the enrolled set are under controlled environment and of high quality. While, the testing videos, captured on arbitrary spots, are under poor lighting, of low resolution and even with considerable blur. S2V face recognition poses a huge challenge due to the great discrepancies of imaging conditions between the still images and the video data. The difference brought by variant conditions could lead to faces of a certain person lying in different subspaces, making the S2V recognition very challenging.

Traditionally, the S2V scenario has been formulated in frameworks of the subspace methods [1, 2]. In the subspace framework, a video is represented as a subspace, and a canonical angle between the subspace and a still image is computed as a matching score. Although classic subspace-based methods could obtain representative face features, performance degenerates severely when wide differences exist in the intra-class samples. A natural way to deal with this problem is to learn a common mapping space for the polymorphous samples. Typically, Hadid et al. [3] applied the classic statistical based methods such as PCA [4] and LDA [5] to obtain a distribution-optimal or discrimination-optimal transform for the image-based representation in video-based face recognition systems. Recently, Huang et al. [6] proposed an improved LDA to learn projections, by using partial weighting to emphasize cross-scenario images in the discriminant analysis. One weakness of these methods is its reliance on a single mapping to build a common discriminant space for samples in different scenarios. As data presents distinctly different characteristics towards specific scenario, it is obviously inefficient to model them in a uniform mode.

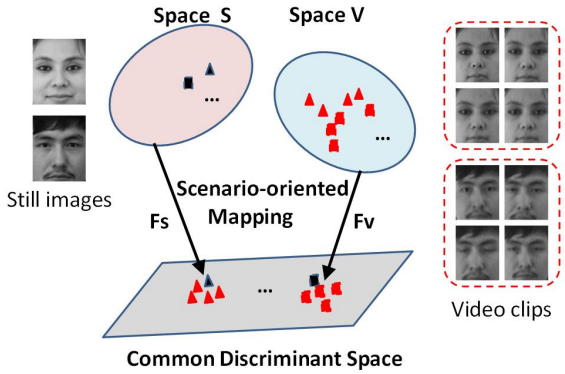

Fig. 1. Overview of the proposed method. $\{F_s, F_v\}$ are mapping functions for space S and V. Shapes represent classes, and colors present scenarios.

In this paper, we propose a novel discriminant analysis method by learning separate transforms for different scenarios, in order to model the underlying data manifold in corresponding modality more effectively. Based on the scenario oriented transforms, this algorithm pursues a common discriminant space where samples from different scenarios are comparable. To obtain good discrimination in the mapping space, we formulate the learning objective by compelling that the intra-class faces from different scenarios cluster together while the inter-class faces separate far away. The gradient de-

scend algorithm is used to get the optimal transforms. Finally, face recognition from still images to videos is performed by matching the closest samples in the learned embedding space. Overview of the proposed method is shown in Fig. 1.

2. SCENARIO ORIENTED DISCRIMINANT ANALYSIS

2.1. Problem description

In S2V face recognition scenario, there is generally only one high resolution still image enrolled for the gallery while a set of low resolution video clips are available for probing. Formally, let $S = \{s_1, s_2, ...s_{N_S}\}$ be the gallery set containing N_S still images, where s_i is the still image of the i^{th} person. Correspondingly, assume $V = \{V_1, V_2, ..., V_{N_V}\}$ be the query set containing N_V video clips, where $V_j = \{v_{j,1}, v_{j,2}, ..., v_{j,N_{V_j}}\}$ denotes the j^{th} query video and N_{V_j} is the number of video frames in V_j. In this way, the identity recognition of video V_j from the gallery S performs the following algorithm $\mathcal{A}$:

$$\mathcal{A} : \hat{i} = \arg \min_{i=1,2,...,N_s} d(s_i, V_j), \tag{1}$$

where $d(,)$ defines the distance between image s_i and video clip V_j.

Typically, the gallery set S and probe set V are captured under different scenarios. One common choice is to seek a learned-based similarity distance in a transformed space [7]:

$$d_f(s_i, V_j) = d(f(V_j), f(s_i)), \tag{2}$$

where the function f indicates a mapping space $f : x \rightarrow f(x)$. Thus, the improved algorithm turns to perform the identity recognition in the mapping space. Motivated by the methods above, the paper proposes a novel learning algrithm to find such a transformed space suitable for the S2V face recognition scenario.

2.2. The proposed method

Classical LDA supposes data of each class are generated from a single normal distribution and seeks a uniform mapping for all the classes [6, 8]. However, in the S2V problem, each class contains two types of data, namely low resolution video frames and high resolution still images, which may not even lie in one single subspace. Instead of using a unique mapping as in LDA, we exploit separate transformation for each scenario and then transform all the data into a common subspace where the similarity computation is meaningful. To obtain good discrimination, we compel that the person-specific faces from different scenarios cluster together while the faces from different person separate far away. Finally, we model the confinement term as the learning objective for subspace optimization.

Assume the training set be $T = \{s_i \in S \cup V_i \in V\}, 1 \leq i \leq N_S$, where $S = \{s_i \in \mathbb{R}^{d_S}\}_{i=1}^{N_S}$ is the still images set as denoted in section 2.1, and $V = \{V_1, V_2, ..., V_{N_S}\}$ holds the corresponding video clip $V_i = \{v_{i,k} \in \mathbb{R}^{d_V}\}_{k=1}^{N_{V_i}}$ for each person i in S. d_S and d_V are the sample dimensions. We denote the transforms for still scene S and video scene V by $F_s(\theta_s) \in \mathbb{R}^{d' \times d_S}$ and $F_v(\theta_v) \in \mathbb{R}^{d' \times d_V}$, where θ_s and θ_v are the transform parameters, and d' is the mapping dimension of transform matrixes. Then, the intra-class compactness term J_w and inter-class repulsion term J_b in the new space $\{\theta_s, \theta_v\}$ are modeled as :

$$\begin{aligned} J_w(\theta_s, \theta_v) &= \frac{1}{N_w} \sum_{i=1}^{N_s} \sum_{k=1}^{N_{V_i}} ||F_v v_{i,k} - F_s s_i||^2, \\ J_b(\theta_s, \theta_v) &= \frac{1}{N_b} \sum_{i=1}^{N_s} \sum_{j=1,2,..,N_s; j \neq i} ||F_v \bar{V_j} - F_s s_i||^2, \end{aligned} \tag{3}$$

where

$$\bar{V_j} = \frac{1}{N_{V_j}} \sum_{k=1}^{N_{V_j}} v_{j,k}, \tag{4}$$

and N_w and N_b are the number of pairs from the same class and different class respectively. For constructing the repulsion term J_b, we use the average of the frames $\bar{V_j}$ in a short video V_j as a representative, and compare it with the still images s_i from other identities. The reason for not using every fames in the video respectively lies in that the number of still images in the training set is much smaller than that of video frames in total, and the discordance would cause serious bias for the following discriminant training. The same problem is also discussed in [6, 9]. In fact, the adjacent frames in a short-time video (eg, the videos in the COX-S2V dataset) change little, so it is reasonable to apply the average frame to represent the whole video frames set in this situation. Particularly, for those complex videos, which contain severe variations in poses, expressions and lights, the special modeling approach should be designed to construct the video representation. This case will be further studied in our future work.

To better discriminate the samples from different classes, we should compel the video frames towards the still images from the same identity and far from those of distinct identities. Based on this principle, we derive the final formulation for the discriminant learning as:

$$J(\theta_s, \theta_v) = J_w(\theta_s, \theta_v) - \alpha * J_b(\theta_s, \theta_v), \tag{5}$$

where α indicates the tradeoff of the compactness and repulsion. From the formation of Eq.(5), we could figure out that minimizing function J means defining a pair of mappings $\{F_s(\theta_s), F_v(\theta_v)\}$ best separating the training samples, even for different scenes.

To solve the problem in Eq.(5) with a simple matrix differentiation, we reform it in the following way. Let $S = [s_1, s_2, ...s_{N_S}] \in \mathbb{R}^{d_S \times N_S}$ collect all the still images for N_S person, and $V_i = [v_{i,1}, v_{i,2}, ..., v_{i,N_{V_i}}] \in \mathbb{R}^{d_V \times N_{V_i}}$ collect the N_{V_i} video frames for the person i in S, $1 \leq i \leq N_S$.

Then $V = [V_1, V_2, ..., V_{N_S}] \in \mathbb{R}^{d_V \times N_v}$ represent the video frames set for all the N_S person, and $\bar{V} = [\bar{V}_1, \bar{V}_2, ..., \bar{V}_{N_S}] \in \mathbb{R}^{d_V \times N_S}$ is the average form of V. In addition, we set:

$$\begin{aligned}
&\bar{S} = [\bar{s}_1, \bar{s}_2, ..., \bar{s}_{N_s}], \quad \bar{s}_i = [s_i, ..., s_i] \in \mathbb{R}^{d_S \times N_{V_i}}, \\
&\bar{V}^\dagger = [\bar{V}_1^\dagger, \bar{V}_2^\dagger, ..., \bar{V}_{N_S}^\dagger], \quad \bar{V}_i^\dagger = \bar{V}/\bar{V}_i \in \mathbb{R}^{d_V \times (N_S - 1)}, \\
&S^\dagger = [s_1^\dagger, s_2^\dagger, ..., s_{N_S}^\dagger], \quad s_i^\dagger = [s_i, ..., s_i] \in \mathbb{R}^{d_S \times (N_S - 1)}.
\end{aligned} \tag{6}$$

where $\bar{V}/\bar{V}_i$ indicates a residue matrix from removing i^{th} column vector $\bar{V}_i$ from matrix $\bar{V}$. Then, we cast the problem in Eq.(5) into the following simplified form:

$$\min_{F_s, F_v} J = \frac{1}{N_w} ||F_v V - F_s \bar{S}||_F^2 - \frac{\alpha}{N_b} ||F_v \bar{V}^\dagger - F_s S^\dagger||_F^2, \tag{7}$$

where $||\square||_F^2$ stands for the Frobenius norm of matrix $\square$.

A straightforward way to optimize Eq.(7) is the gradient descend approach. According the matrix theory, the derivation of $J(\theta_s, \theta_v)$ with respect to $\{\theta_s, \theta_v\}$ can be computed as:

$$\begin{aligned}
\partial J/\partial F_s =& \frac{2}{N_w}(F_s S N^\dagger S_\top - F_v V \overline{S}_\top) \\
&- \frac{2\alpha}{N_b}(F_s S N^\ddagger S_\top - F_v \bar{V}^\dagger S_\top^\dagger), \\
\partial J/\partial F_v =& \frac{2}{N_w}(F_v V - F_s \bar{S}) V_\top - \frac{2\alpha}{N_b}(F_v \bar{V}^\dagger - F_s S^\dagger)\bar{V}_\top^\dagger,
\end{aligned} \tag{8}$$

where $\square_\top$ denote the transposition of matrix $\square$, and $N^\dagger \in \mathbb{R}^{N_S \times N_S}$ and $N^\ddagger \in \mathbb{R}^{N_S \times N_S}$ are diagonal matrixes with the i^{th} diagonal element as N_{V_i} and $(N_S - 1)$ respectively. After obtaining the gradient, the parameter $\{F_s, F_v\}$ can be updated by Eq.(9) until convergence.

$$\begin{aligned}
F_s &= F_s - \gamma \partial J/\partial F_s, \\
F_v &= F_v - \gamma \partial J/\partial F_v,
\end{aligned} \tag{9}$$

where γ is the learning rate.

As for the complexity, the cost of gradient computation results in $O(d' D^2 (N_S^2 + N_v))$ In our application, the sample dimensions d_S and d_V are equal, which are denoted as D here. The mapping dimension d' is rather small ($d' \ll D$). In this way, the complexity is proportional to the category number N_S and the total video frames number N_v. For a normal-sized dataset, the gradient computation in Eqs.(8) is rather efficient. In fact, it just needs dozens of iterations (around ten) before the updating converges in our experiments.

3. EXPERIMENTS

3.1. Dataset and Experimental Settings

COX-S2V is a dataset designed for the real-world Still-to-Video face recognition research [6]. The dataset consists of high resolution still images and four low resolution videos of 1000 subjects. Table 1 gives the shooting condition of the

	Video1	Video2	Video3	Video4
Viewpoint	√	√	×	×
Illumination	×	×	×	×
Expression	×	×	×	×

Table 1. Environment setting of the four videos.

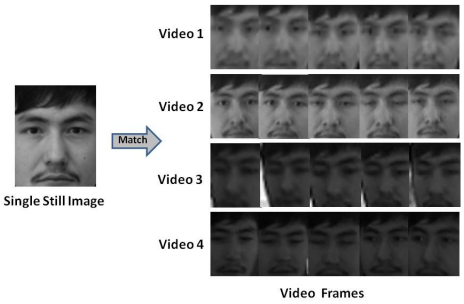

Fig. 2. Examples in the COX-S2V dataset: one still image per person in the gallery set along with four probe videos.

videos. In particular, video 1 and 3 are more blurred than video 2 and 4 for further shooting distance. Some samples are shown in Fig. 2. According to the protocol in [6], we use the still images and video clips of 300 persons for training, and that of the rest 700 persons for testing. During testing stage, the still images serve as the gallery, and videos serve as the probe. The rank-1 recognition rate is used to test the performance.

All images are scaled to 96×120 pixels firstly. The pixel descriptor is used for the baseline assessment. To explore the potentiality of the proposed PMDA, we use local phase quantization (LPQ) [10] and Gabor magnitude [11] to compose the complementary phase-magnitude descriptor for face representation [12]. For LPQ, we set the local window size and the low frequency coefficient as $7 \times 7, 1/7$ respectively. For Gabor, we use 40 Gabor wavelets with 5 scales and 8 orientations. The Gabor kernel's size, the frequency parameters k_{max}, f^v, and the parameter σ are set to $31 \times 31, 1.0, \sqrt{2}$, and 2 respectively. Before training, we apply Principal Components Analysis (PCA) [13] on all the descriptors, and the target dimension is set as 1400 compromising between discrimination preserving and noises compression. Besides, the trade-off parameter α is set to 0.5 for equal weights of compactness and separability constraints. The learning rate γ is 0.001.

3.2. Experiment Results and Analysis

Table 2 shows the performance of the Scenario Oriented Discriminant Analysis (SDA) on the COX S2V dataset. The LPQ and Gabor features promotes the performance significantly comparing with the original pixel feature, especially for video 3 and video 4. The largest increment reaches 18.66%, suggesting that the LPQ and Gabor features are quite effective for

	Video1	Video2	Video3	Video4
Pixel	58.29	75.00	15.71	47.91
LPQ	58.57	80.14	24.29	66.57
Gabor	67.29	83.57	26.86	61.29
LPQ+Gabor	**76.0**	**87.43**	**35.14**	**73.71**

Table 2. Accuracy of different features on COX-S2V (%).

Methods	Video1	Video2	Video3	Video4
PCA+LDA [6]	47.57	68.28	20.00	49.85
PCA+LPP [6]	47.43	68.57	20.12	49.14
PCA+LFDA [6]	21.86	44.00	3.29	16.14
PCA+CDEF [6]	8.14	12.99	6.57	5.00
PCA+PaLo-LDA [6]	52.43	73.00	22.00	56.71
SDA (This work)	**76.0**	**87.43**	**35.14**	**73.71**

Table 3. Comparison with other methods on COX-S2V (%).

capturing information on the uncontrolled environment. The last line of Table 2 gives the accuracy derived from fusing the similarity of LPQ and Gabor with a simple average operation. The fusing results show remarkable improvement to that of single descriptor. Actually, LPQ emphasizes on the facial texture information, while Gabor magnitude emphasizes on the structure information of the face. Accuracy can be obviously improved by combining the two features, for their strong complementation to each other.

Besides, we also compare the SDA method to state-of-art methods under the same configuration in Table 3. As the S2V dataset has been released for a short time, we just compare with the results from the releaser [6], including methods such as LDA [4], LPP [14], LFDA [15]. The best result is reported from the PaLo-LDA method, which has already achieved quite satisfactory result on the difficult environment. While, our approach performs significantly better even than it, with a large increment of $23.57\%, 14.43\%, 13.14\%, 17.00\%$ on the corresponding four videos. The PaLo-LDA is an improved version from classical LDA. It involves computing the weight for each image pair in the training set, which is a time-consuming task as the sample number gets large. Instead of turning to the weighting skills, we learn scenario oriented transforms for the cross-scenario problem. Comparing with the single-mapping PaLo-LDA, our approach could model the underlying data manifold in corresponding modality more effectively, and so recognizes face in this condition much better.

In addition, we also give a detail analysis of the parameter: the mapping dimension of transform matrix. As the recognition is conducted upon features in the mapping space, the mapping dimension imposes significant importance on the proposed method. The original feature has been reduced to 1400 dimension by PCA before training. In this section, we vary the mapping dimension from 200 to 1400 by a step of 200 to test which is more suitable for the S2V facial description. Experimental results on video 2 are drawn in Fig. 3. As

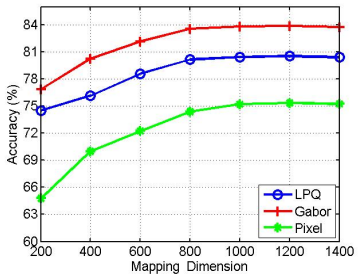

Fig. 3. Performance with different mapping dimensions.

we can see, the accuracy benefits from increasing the mapping dimension. With too low dimension, performance drops for losing much discriminative information in the mapping operation. However, as the dimension exceeds 800, the accuracy increment is less than 0.5%. Moreover, as the dimension continues to increase, performance takes a downward trend for overfitting on the training set. Above all, compromising between the model complexity and accuracy, we set the mapping dimension as 800 for the proposed method.

4. CONCLUSION

In this paper, we develop an approach, named as Scenario oriented Discriminant Analysis, to deal with the Still-to-Video face recognition problem. The algorithm learns specific transforms for data in different scenarios (images, or videos), in order to pursues a common mapping space where compare of the heterogeneous data is effective. Discriminant learning is performed by incorporating the intra-class compactness and the inter-class dispersion. Comparative experiments indicated that the proposed method results in high accuracy and robustness for the S2V face recognition problem.

5. REFERENCES

[1] E. Oja, *Subspace Methods for Pattern Recognition*, Research Study Press, 1983.

[2] Y Ariki and W Ishikawa, "Integration of face and speaker recognition by subspace method," in *ICPR*, 1996, vol. 3, pp. 456–460.

[3] Abdenour Hadid and M Pietikainen, "From still image to video-based face recognition: an experimental analysis," in *Automatic Face and Gesture Recognition, 2004. Proceedings. Sixth IEEE International Conference on*. IEEE, 2004, pp. 813–818.

[4] Peter N. Belhumeur, Joao P. Hespanha, and David J. Kriegman, "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection," *PAMI*, vol. 19, pp. 711–720, 1997.

[5] Hua Yu and Jie Yang, "A direct lda algorithm for high-dimensional data-with application to face recognition," *Pattern recognition*, vol. 34, no. 10, pp. 2067, 2001.

[6] Zhiwu Huang, Shiguang Shan, Haihong Zhang, Shihong Lao, Alifu Kuerban, and Xilin Chen, "Benchmarking still-to-video face recognition via partial and local linear discriminant analysis on cox-s2v dataset," in *ACCV*, pp. 589–600.

[7] Matthieu Guillaumin, Jakob Verbeek, and Cordelia Schmid, "Is that you? metric learning approaches for face identification," in *ICCV*, 2009, pp. 498–505.

[8] Sebastian Mika, Gunnar Ratsch, Jason Weston, Bernhard Scholkopf, and KR Mullers, "Fisher discriminant analysis with kernels," in *Signal Processing Society Workshop*, 1999, pp. 41–48.

[9] Dahua Lin and Xiaoou Tang, "Inter-modality face recognition," in *ECCV*, pp. 13–26. 2006.

[10] Ville Ojansivu and Janne Heikkilä, "Blur insensitive texture classification using local phase quantization," in *Image and Signal Processing*, pp. 236–243. Springer, 2008.

[11] Chengjun Liu and Harry Wechsler, "Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition," *Image processing, IEEE Transactions on*, vol. 11, no. 4, pp. 467–476, 2002.

[12] Yan Li, Shiguang Shan, Haihong Zhang, Shihong Lao, and Xilin Chen, "Fusing magnitude and phase features for robust face recognition," .

[13] Ian T Jolliffe, *Principal component analysis*, Springer verlag, 2002.

[14] X Niyogi, "Locality preserving projections," in *NIPS*, 2004, vol. 16, p. 153.

[15] Masashi Sugiyama, "Dimensionality reduction of multimodal labeled data by local fisher discriminant analysis," *The Journal of Machine Learning Research*, vol. 8, pp. 1027–1061, 2007.