Benchmarking Still-to-Video Face Recognition via Partial and Local Linear Discriminant Analysis on COX-S2V Dataset

Zhiwu Huang[1,2], Shiguang Shan[1], Haihong Zhang[3], Shihong Lao[3], Alifu Kuerban[4], Xilin Chen[1]

[1]Key Lab of Intelligent Information Processing, Institute of Computing Technology,
Chinese Academy of Sciences, Beijing 100190, China
[2]University of Chinese Academy of Sciences, Beijing 100049, China
[3]OMRON Social Solutions Co. Ltd, Kyoto, Japan
[4]College of Information Science and Engineering, Xinjiang University
{zhiwu.huang, shiguang.shan}@vipl.ict.ac.cn,
angelazhang@ssb.kusatsu.omron.co.jp, lao@ari.ncl.omron.co.jp,
ghalipk@xju.edu.cn, xilin.chen@vipl.ict.ac.cn

Abstract. In this paper, we explore the real-world Still-to-Video (S2V) face recognition scenario, where only very few (single, in many cases) still images per person are enrolled into the gallery while it is usually possible to capture one or multiple video clips as probe. Typical application of S2V is mug-shot based watch list screening. Generally, in this scenario, the still image(s) were collected under controlled environment, thus of high quality and resolution, in frontal view, with normal lighting and neutral expression. On the contrary, the testing video frames are of low resolution and low quality, possibly with blur, and captured under poor lighting, in non-frontal view. We reveal that the S2V face recognition has been heavily overlooked in the past. Therefore, we provide a benchmarking in terms of both a large scale dataset and a new solution to the problem. Specifically, we collect (and release) a new dataset named COX-S2V, which contains 1,000 subjects, with each subject a high quality photo and four video clips captured simulating video surveillance scenario. Together with the database, a clear evaluation protocol is designed for benchmarking. In addition, in addressing this problem, we further propose a novel method named Partial and Local Linear Discriminant Analysis (PaLo-LDA). We then evaluated the method on COX-S2V and compared with several classic methods including LDA, LPP, ScSR. Evaluation results not only show the grand challenges of the COX-S2V, but also validate the effectiveness of the proposed PaLo-LDA method over the competitive methods.

1 Introduction

Recently, the availability of affordable cameras and low cost storage devices has contributed to a rapid increase in the usage of surveillance systems. Vast amounts of video footage are continuously acquired to monitor government compounds, military installations, commercial sites, and private premises. As a result, various video-based face recognition (VFR) applications, e.g., [1–5], emerged in recent years. Among them, many works assume that both gallery and query set are video sequences, for which this kind of application can be called Video-to-Video (V2V) face recognition.

In contrast, Zhou et al. [6, 7] also defined a Still-to-Video (S2V) face recognition scenario in which the gallery contains still image(s) while the probes are video clips. In their work, however, the faces in still images are all low-resolution as the faces in videos or even directly extracted from the videos in one of their collected databases. But, in a real-world S2V face recognition application such as mug shot based watch list, the still images is often captured by a high quality digital camera under controlled environment while the videos are captured with ordinary video recorder, which implies low quality and low resolution.

To simulate the above application scenario, in this paper, we believe that a real-world Still-to-Video face recognition should be more emphasized on the following scenario: a face recognition system only enrolls one single high resolution still image per person into the gallery, while a sequence of low resolution video frames are used for probing. Evidently, the S2V scenario designed in this paper is much more challenging and better fits real-world application, compared with the scenario in previous work. To advance the research on this S2V scenario, we also collect and release a dataset called COX-S2V database.

Due to intrinsic non-rigid transform and extrinsic uncontrolled environment, human face images captured by video cameras from a distance often contain nonlinear variations caused by variations in pose, illumination, or expression. The difficulties are further increased when the resolution of the face images is low, which is however typical in closed-circuit television systems. Furthermore, misalignment would be more serious in the real-world scenario. Therefore, these factors raise *two key issues to be addressed*: 1) How to match a high resolution still image with low resolution video frames? 2) How to cope with those nonlinear variations due to pose, illumination, expression and misalignment?

To address the first issue, to our best knowledge, there are three ways. The first way (e.g.,[8]) is to extract the invariant discriminant information from low-resolution images directly. The second way is a “two-step” based method (e.g., [9–11]) which first uses super-resolution (SR) techniques to enhance the image before face recognition. The last way (e.g., [12–15]) no longer tries to recover a visually improved high-resolution image, but to directly improve recognition performance by learning a mapping between LR image and HR image.

However, as the resolution decreases more, most above methods may become more vulnerable to environmental and intrinsic variations, such as pose, illumination, expression and even misalignment. For the second issue, Arandjelovic et al. [16] proposed to extract signatures of illumination and pose from genetic training faces to represent a shape-illumination manifold. Jia et al. [17] developed a generalized face SR method for feature-domain reconstruction based on multi-linear analysis, which is able to accommodate multiple factors such as pose, illumination and expression.

In this paper, to address the above two issues, we propose a Partial and Local Linear Discriminant Analysis (PaLo-LDA) method. Briefly speaking, on one hand, to match our proposed S2V scenario, we re-formulate the traditional LDA by partially weighting, i.e., emphasizing cross-resolution image pairs. On the other hand, we also implicitly take the pose and lighting variations into account by locally weighting, i.e., emphasizing image pairs with pose (i.e., near-frontal) and lighting (i.e., normal lighting) similar to the still images in the gallery (i.e., frontal, normal lighting). Extensive experimental results demonstrate that our PaLo-LDA method achieves better performance than most state-of-the-art methods on our COX-S2V dataset.

The remainder of the paper is organized as follows: Section 2 formulates the proposed real-world Still-to-Video (S2V) face recognition problem and briefly reviews the existing work. Section 3 presents our recently collected dataset COX-S2V and the accompanying evaluation protocol in detail. Section 4 details the proposed Partial and Local Linear Discriminant Analysis (PaLo-LDA) for S2V face recognition. Section 5 presents our extensive experimental result on COX-S2V dataset, followed by conclusions in Section 6.

2 Still-to-Video (S2V) Face Recognition: Problem and Previous Work

2.1 Problem Formulation

As mentioned above, in a Still-to-Video (S2V) face recognition scenario, for each person, generally there is only one high resolution still image enrolled for each person while a set of low resolution video frames is available for probing. Formally, the problem is defined as follows: Let $S = \{s_1, s_2, \ldots, s_{n_s}\}$ be the still image gallery set, s_i is the gallery image of the i^{th} person, $x_i \in \Re^d$, $1 \leq i \leq n_s$, where d is the dimensionality or total pixels of each face sample and n_s is the number of persons in the gallery set. Assume $V = \{v_1, v_2, \ldots, v_n\}$ is a query person's video sequence, n is total video frame number for the query person, $v_j \in \Re^{d'}$, d' is the dimensionality (or total pixels) of video frames. The label of V is inferred as follows:

$$c = \arg\min_i d(s_i, V). \tag{1}$$

where $d(s_i, V)$ is a distance between the gallery still image s_i and the probe video frame set V.

2.2 Brief Review of Existing Work

In the case of traditional video-based Face Recognition (VFR), both gallery and query set are video sequences rather than still images. In this case, VFR can be generalized to image-set based classification [2–5], where each target person may be enrolled with one or multiple image sets and a query image set need to be assigned to the identity of its nearest gallery set by calculating its distance from each gallery image set. Since controlled still images and uncontrolled video sequences in our proposed S2V scenario are usually captured in different conditions, most existing V2V face recognition methods are not directly applicable to the S2V scenario.

As discussed in the introduction, there are two key issues raised by our proposed real-world S2V face recognition. Three categories of methods to solve the first issue are also be mentioned. Extracting invariant discriminant feature (e.g., [8]) from low-resolution images is the first way. The second category of methods (e.g., [9–11]) adopt super-resolution (SR) techniques to enhance the image following by traditional face recognition. However, these methods usually have limited performance because the target of SR techniques is for visual enhancement but not for recognition performance improvement. The last categories of methods (e.g., [12–15]) directly improve recognition performance by learning a mapping from LR image to HR image. Nevertheless, they are vulnerable to nonlinear changes of pose, illumination, expression and even misalignment.

To solve the second issue, a shape-illumination manifold was represented by Arandjelovic et al. [16] to extract signatures of illumination and pose from genetic training faces. However, a limitation of this approach is that it requires multiple still high resolution gallery images, making it impractical for our proposed face recognition scenarios. Jia et al. [17] developed a generalized tensor-based face SR method which is able to accommodate multiple factors such as pose, illumination and expression. Nevertheless, the tensor manipulations for reconstruction also demand high computational expenses since no explicit connections between LR and HR pairs are established.

In addition, several existing methods deal with the issue by preserving the local structure of the data. Locality Preserving Projection (LPP) [18] seeks for an embedding transformation that nearby data pairs in the original space close in the embedding space. However, since the distribution of data in our proposed problem has a certain particularity(which will be stated in the section 4), these above methods could not work well directly. In a word, to our best knowledge, little existing methods are designed specially for our proposed S2V face recognition scenario.

3 COX-S2V Dataset and Evaluation Protocol

COX-S2V is a new dataset we constructed for the research on the real-world Still-to-Video face recognition problem. We will release it with the publication of this work. The following subsection will details the construction and protocol of COX-S2V dataset.

3.1 Construction of the COX S2V Dataset

The dataset consists of controlled high resolution still images and four uncontrolled low resolution video sequences of 1000 subjects. As shown in Fig.1, the still images were captured with a high quality digital camera under controlled

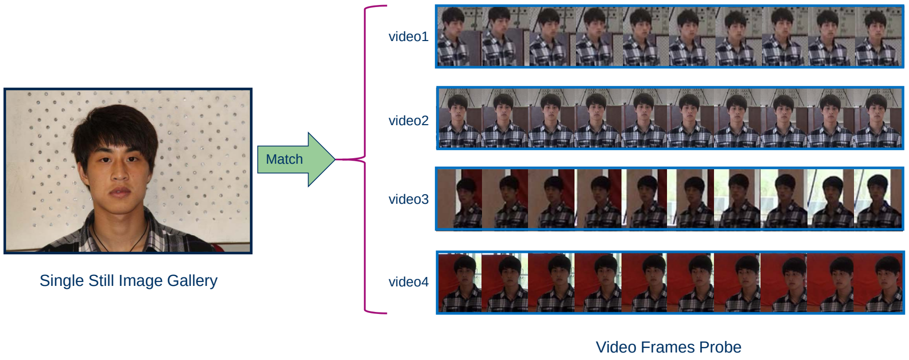

Fig. 1. One example in the COX-S2V dataset. The left is a gallery image, which is a high resolution still image captured with a high quality digital camera under controlled environment; The right shows four video sequence probe sets, which are four low-resolution video sequences of the same subject but captured with off-the-shelf camcorder under uncontrolled environment.

environment and with cooperative subjects, which leads to still images of high resolution and ID photo quality.

The four video clips of people are collected by two different off-the-shelf camcorders. These videos are captured in a gymnasium with high ceilings, enclosed entirely on one side with glass windows. This environment approximates outdoor lighting condition.

Specifically, video1 clips and video2 clips are captured by the first camcorder around 13.5 meters and 6 meters respectively away from the subjects while video3 clips and video4 clips are captured by the second camcorder around 13.5 meters and 6 meters respectively away from the subjects. Both of the camcorder-s are at a height of around 2 meters. In one scenario, the subjects move from the start point forward to the first camcorder along a straight line; In the other scenario, the subjects walk from the start point to the second camcorder along a curve line. In each process of walking, subjects walk naturally without any restriction on expression, head orientation, etc.. Additionally, each video sequence lasts approximately 1 second, contains around 25 frames. More details about the video clips captured by the camcorders are shown in Table 1.

3.2 Evaluation Protocol based on the COX-S2V dataset

Face recognition is often naturally described as part of a Detection-Alignment-Recognition (DAR) pipe-line [19]. To facilitate this process, we have purposefully designed our dataset to represent the output of the detection process. As we emphasize on face recognition, to ensure these faces in the videos are not too hard to detect, all the images are detection verified by several detectors including

Table 1. Details of the video sources captured by different camcorders.

Videos	Face-size	Viewpoint	Illumination	Expression
video1	16×20	controlled	uncontrolled	uncontrolled
video2	48×60	controlled	uncontrolled	uncontrolled
video3	16×20	uncontrolled	uncontrolled	uncontrolled
video4	48×60	uncontrolled	uncontrolled	uncontrolled

a commercial one from OKAO.[1] In COX-S2V dataset, we design a protocol specifically for the proposed S2V face recognition. In the protocol, we use the high resolution still images and video clips of 300 persons for training, and the remaining 700 persons' still images and video frames for testing. For training purpose, all the 300 subjects' data including one high resolution still image and four clips can be used. During testing stage, the high resolution still images from the rest 700 subjects form the gallery set, and the probe set contains four clips from each of the 700 subjects. The rank-1 face recognition rate would be used to test the performance of involved approaches on COX-S2V dataset.

4 Partial and Local Linear Discriminant Analysis (PaLo-LDA) for S2V Face Recognition Scenario

Assume the whole sample set $X = S \cup V = \{x_i | x_i \in S \text{ or } x_i \in V\}$, where S and V are denoted in above section. From here on, we denote C_i and T_i to be the class label and type label ($x_i \in S$ or $x_i \in V$) of the i-th sample x_i, n_k to be the sample number of the k-th class, n to be the number of all samples.

4.1 Related Work

Intuitively and ideally, we would like that pairs of samples from the same class are made close, while the pairs of samples from different classes are separated from each other. One of the most popular dimensionality reduction techniques Linear Discriminant Analysis (LDA)[20, 21] could directly achieve the goal in the pairwise expression way as follows:

$$S^{(w)} = \frac{1}{2} \sum_{i,j=1}^{n} W_{i,j}^{(w)} (x_i - x_j)(x_i - x_j)^T, \tag{2}$$

$$S^{(b)} = \frac{1}{2} \sum_{i,j=1}^{n} W_{i,j}^{(b)} (x_i - x_j)(x_i - x_j)^T, \tag{3}$$

[1] http://www.omron.com/r_d/technavi/vision/okao/detection.html

where

$$W_{ij}^{(w)} = \begin{cases} 1/n_k, & \text{if } C_i = C_j = k, \\ 0, & \text{if } C_i \neq C_j, \end{cases} \tag{4}$$

$$W_{ij}^{(b)} = \begin{cases} 1/n - 1/n_k, & \text{if } C_i = C_j = k, \\ 1/n, & \text{if } C_i \neq C_j. \end{cases} \tag{5}$$

Considering the density of data may be different depending on regions, Zelnik-Manor et al. [22] defined an affinity matrix taking the local scaling of data into account as following:

$$A_{ij} = exp\left(-\frac{d(x_i, x_j)}{\sigma_i \sigma_j}\right), \tag{6}$$

where $d(x_i, x_j)$ represents the Euclidean distance of samples x_i and x_j, σ_i denotes the local scaling of the data samples around x_i, which is determined by

$$\sigma_i = \|x_i - x_i^{(K)}\|. \tag{7}$$

where $x_i^{(K)}$ is the K-th nearest neighbor of x_i.

4.2 Proposed Algorithm

"Partial": As is well known, LDA supposes the samples of each class are generated from single normal distribution. However, since each class contains two types of data in our proposed problem, the data could not be grouped in one single cluster. In fact, one type of our data (*low resolution video frame*) could be grouped in a cluster, while the other type of data (*high resolution still image*) is a single sample which would stay away from the cluster. Additionally, for the proposed S2V problem, it only need to focus on the matching of samples from different types (*Still-to-Video*). Therefore, in each class, we would like to partially put more weight on the pairs of different type samples while giving less weight on the pairs of the same type samples in the above pairwise expression computation. That is, we use partial weighting to emphasize cross-resolution image pairs. Note that, weighting the sample pairs from different classes is not necessary because we want to separate them from each other irrespective of the affinity in the original space. As shown in the following formulations, we denote the partial counterparts of matrices by symbols with tilde.

$$\tilde{S}^{(w)} = \frac{1}{2} \sum_{i,j=1}^{n} \tilde{W}_{i,j}^{(w)} (x_i - x_j)(x_i - x_j)^T, \tag{8}$$

$$\tilde{S}^{(b)} = \frac{1}{2} \sum_{i,j=1}^{n} \tilde{W}_{i,j}^{(b)} (x_i - x_j)(x_i - x_j)^T, \tag{9}$$

where

$$\tilde{W}_{ij}^{(w)} = \begin{cases} A_{ij}/n_k, & \text{if } C_i = C_j = k \wedge T_i \neq T_j, \\ \beta A_{ij}/n_k, & \text{if } C_i = C_j = k \wedge T_i = T_j, \\ 0, & \text{if } C_i \neq C_j, \end{cases} \tag{10}$$

$$\tilde{W}_{ij}^{(b)} = \begin{cases} A_{ij}(1/n - 1/n_k), & \text{if } C_i = C_j = k \wedge T_i \neq T_j, \\ \beta A_{ij}(1/n - 1/n_k), & \text{if } C_i = C_j = k \wedge T_i = T_j, \\ 1/n, & \text{if } C_i \neq C_j. \end{cases} \tag{11}$$

"Local": In the S2V scenario, the data in each class not only belongs to different types (*still image or video frame*), but also has various nonlinear variations of pose, illumination, expression and even misalignment. To deal with this problem, we locally define the weight of sample pairs in the same class. Specifically, we weight the sample pairs in the same class to let far apart sample pairs have less influence on defined scatter matrices than the nearby sample pairs. Consequently, the local weighting implicitly take nonlinear variations of pose, illumination, expression into account, with emphasizing image pairs with similar pose (i.e., near-frontal), lighting (i.e., near-normal) and expression (i.e., near-neutral) video frames to the still images (i.e., frontal, normal lighting and neutral expression).

Owing to the specific data distribution in our proposed S2V scenario, the density of data samples may be different depending on regions. Therefore, it is more reasonable to adopt the local scaling of data. Following this idea and taking into account the above "partial" and "local" constraints, we modify the affinity matrix defined by Zelnik-Manor et al. [22] as following

$$\tilde{A}_{ij} = exp\left(-\frac{\tilde{d}(x_i, x_j)}{\sigma_i \sigma_j}\right), \tag{12}$$

where

$$\tilde{d}(x_i, x_j) = \begin{cases} d(x_i, x_j), & \text{if } T(x_i) \neq T(x_j), \\ m(x_i, x_j), & \text{if } T(x_i) = T(x_j). \end{cases} \tag{13}$$

In Equ. (12-13), $d(x_i, x_j), \sigma_i, \sigma_j$ is the same meaning as those in Eq. (6), $m(x_i, x_j) = max\{max(d(x_i, :)), max(d(x_j, :))\}$, where $max(d(x_i, :))$ computes the maximum Euclidean distance between sample x_i and all the other samples.

After introducing the concepts of "Partial" and "Local", we can turn to calculate our scatter matrices and define our projection matrix. The within-class scatter matrix $\tilde{S}^{(w)}$ and the between-class scatter matrix $\tilde{S}^{(b)}$ can be efficiently calculated following [21]. Now, we can define our projection matrix as following:

$$\tilde{P} = \arg\max \frac{\tilde{P}^T \tilde{S}^{(b)} \tilde{P}}{\tilde{P}^T \tilde{S}^{(w)} \tilde{P}}. \tag{14}$$

The final objective function is a standard generalized eigenvalue problem that can be solved using any eigen-solver. It will produce real eigenvectors and eigenvalues because both $\tilde{S}^{(b)}$ and $\tilde{S}^{(w)}$ are symmetric matrices.

5 Experiments on COX-S2V

5.1 Baseline and Competing Methods

In order to demonstrate the grand challenge of our proposed problem in the COX-S2V datasets, we compare three types of traditional methods in face recognition. The first one utilizes the most popular dimensionality reduction methods after Bicubic interpolation, e.g. Bicubic+LDA[20] or other methods. The second one uses the super-resolution techniques and then applies the traditional algorithms consequently. In the experiment we presents the result with the state-of-the-art super-resolution algorithm ScSR [11] followed by LDA or other methods. The last one is hallucinating feature method. Here, we implement state-of-the-art methods CLPM [14] and CDFE [23]. Note that, above methods SDA [24], LFDA [21], LPP [18], ScSR and CLPM are performed using codes from the original authors.

For the first type of methods, in the experiment, we use Bicubic interpolation to scale both the low resolution and high resolution images to face size of 96×120 pixels. Both of the latter two types perform on the raw face images: the face size from high resolution still image is 96×120 pixels, the face size from *video 1* and *video3* is 16×20 pixels, and the face size from *video 2* and *video4* is 48×60 pixels. Specifically, ScSR super-resolves all the video frames from original resolutions to face size of 96×120 pixels.

In the experiment, the parameters of involved methods are setting as follows: *For SDA*, we fix the number of clusters in each class $K = 5$; *For LPP*, we use *supervised* as the Neighbor Mode, the number of neighborhood $k = 5$, use HeatKernel as the Weight Mode, and let $t = 5$; *For ScSR*, we set the size of dictionary to 1024, parameter $\lambda = 0.15$, the size of each patch to 5×5, the number of training patch to 100000; *For CLPM* , we select the best performance by tuning α from 0.1 to 1 with step of 0.1 and $N(i)$ from 10 to 1000 with step of 50. *For CDFE*, we select the best performance by tuning α from 0.1 to 2 with step of 0.1 and β from 0.1 to 1 with step of 0.1. Note that the above parameter settings are all consistent with original setting of authors. *For our method PaLo-LDA*, the parameter β is selected by tuning it from 0.01 to 0.1 with step of 0.01. Since all the implemented approaches lead to large eigenvalue problems, PCA [25] is applied to reduce the data dimension before feeding it to all methods, with keeping from 60% to 90% of the principal component energy.

In testing stage, we need to calculate the distance between one single still image and a set of video frames. In previous work [26], they defined the distance as Image-to-Set distance. Accordingly, we validate the performance of all involved methods by calculating the following two Image-to-Set distances: Mean Distance ***(MD)*** and Nearest Neighbor Distance ***(NND)*** respectively.

5.2 Experimental Results and Discussion

As shown in Table 2 and Table 3, our method achieves the highest rank-1 recognition rate in all video probe sets compared to state-of-the-art methods. Several

Table 2. Rank-1 Recognition Rate (%) on COX-S2V using ***MD***.

Methods/Probe Set (gray feature)	Video1	Video2	Video3	Video4
Bicubic+PCA+LDA	38.86	60.57	13.57	39.86
Bicubic+PCA+LPP	36.71	60.42	13.71	38.85
Bicubic+PCA+SDA	19.71	25.28	2.86	7.57
Bicubic+PCA+LFDA	17.43	34.00	2.57	10.14
ScSR+PCA+LDA	19.14	37.29	7.57	21.43
ScSR+PCA+LPP	19.00	38.14	7.71	22.29
ScSR+PCA+SDA	14.00	18.43	2.00	5.43
ScSR+PCA+LFDA	14.57	23.86	2.14	8.57
PCA+CLPM	5.14	1.43	1.23	1.57
PCA+CDFE	10.00	5.14	1.71	5.00
Bicubic+PCA+PaLo-LDA	**44.43**	**66.00**	**15.44**	**46.57**

Table 3. Rank-1 Recognition Rate (%) on COX-S2V using ***NND***.

Methods/Probe Set (gray feature)	Video1	Video2	Video3	Video4
Bicubic+PCA+LDA	47.57	68.28	20.00	49.85
Bicubic+PCA+LPP	47.43	68.57	20.12	49.14
Bicubic+PCA+SDA	24.71	33.43	3.29	11.86
Bicubic+PCA+LFDA	21.86	44.00	3.29	16.14
ScSR+PCA+LDA	27.57	50.29	10.43	32.71
ScSR+PCA+LPP	27.29	50.86	10.71	33.00
ScSR+PCA+SDA	20.00	25.00	1.71	8.86
ScSR+PCA+LFDA	18.00	31.57	2.57	12.71
PCA+CLPM	4.43	2.00	1.21	1.86
PCA+CDFE	8.14	12.29	6.57	5.00
Bicubic+PCA+PaLo-LDA	**52.43**	**73.00**	**22.00**	**56.71**

reasons are discussed as follows: LDA puts equal weights on all pairs of sample without considering the cross-resolution and nonlinear data distribution scenario. LPP and LFDA also performs worse for using local affinity to give more weights on the pairs of samples in the same resolution. Although SDA divides each class into several subclasses, it still puts the equal weights on all pairs of samples. Different from above methods, in our method PaLo-LDA, partial weighting addresses the cross-resolution problem, while local weighting implicitly takes other variations (e.g., pose, illumination, lighting etc.) into account. We also did additional validations on the separate effects of partial and local weight schemes, which show average 2.69% and 1.96% gains respectively.

We also conclude several following reasons why state-of-the-art methods ScSR, CLPM and CDFE don't work well: ScSR is visual enhancement oriented but not recognition oriented method, which leads ScSR's performance not well. Since one of the two-view training sets (HR still image set) has only 1 still image per subject and 300 training samples totally, the models of CLPM and CDFE are seriously bias to work worse and even break down on the scenario. Last and may be the most important, these tradition methods are all previously validated in the experimental environment where the low resolution images were artificially obtained by several operation of smoothing and downsampling, and in our dataset, the real-world low resolution images are directly from different cameras with originally lower resolutions, and contain all kinds of nonlinear variations. Therefore, with the grand challenge, our dataset COX-S2V develops a good platform to advance the methods which could work in the real-world low resolution face recognition scenario.

Another deserved discussing point is the Image-to-Set distance. The comparison of Table 1 with Table 2 shows that the performances of most methods using Mean Distance would be worse than those using the Nearest Neighbor Distance. The result illustrates that the Image-to-Set distance plays an important role in advancing the performance of S2V face recognition. Therefore, we will pay more attention on seeking a better Image-to-Set distance in future.

6 Conclusion

In this paper, different from [6], we emphasize a real-world Still-to-Video face recognition problem. Further, we collect a dataset called COX-S2V and designed a reasonable experimental protocol for encouraging more advanced methods to develop on this dataset in the future. As mentioned in this paper, the proposed S2V face recognition scenario raises two key problems. Consequently, we develop a Partial and Local Linear Discriminant Analysis (PaLo-LDA) method to deal with the problems directly. The experimental results on COX-S2V dataset demonstrate that our proposed approach can achieve better performance than existing methods.

Acknowledgements. This work is partially supported by National Basic Research Program of China (973 Program) under contract 2009CB320902; Natural Science Foundation of China (NSFC) under contract Nos. 61173065 and U0835005; and Beijing Natural Science Foundation under contract No. 4111003. Haihong Zhang and Shihong Lao are partially supported by "R&D Program for Implementation of Anti-Crime and Anti-Terrorism Technologies for a Safe and Secure Society", Special Coordination Fund for Promoting Science and Technology of MEXT, the Japanese Government.

References

1. Liu, X., Cheng, T.: Video-based face recognition using adaptive hidden markov models. In: CVPR. (2003) 340–345
2. Arandjelović, O., Shakhnarovich, G., Fisher, J., Cipolla, R., Darrell, T.: Face recognition with image sets using manifold density divergence. In: CVPR 1. (2005) 581–588
3. Kim, T., Kittler, J., Cipolla, R.: Discriminative learning and recognition of image set classes using canonical correlations. IEEE T-PAMI **29** (2007) 1005–1018

4. Cevikalp, H., Triggs, B.: Face recognition based on image sets. In: CVPR. (2010) 2567–2573
5. Wang, R., Guo, H., Davis, L., Dai, Q.: Covariance discriminative learning: A natural and efficient approach to image set classification. In: CVPR. (2012) 2496–2503
6. Zhou, S., Krueger, V., Chellappa, R.: Probabilistic recognition of human faces from video. CVIU **91** (2003) 214–245
7. Zhou, S., Chellappa, R.: Beyond one still image: Face recognition from multiple still images or a video sequence. Face Processing: Advanced Modeling and Methods (2005) 547–567
8. Hwang, W., Huang, X., Noh, K., Kim, J.: Face recognition system using extended curvature gabor classifier bunch for low-resolution face image. In: CVPRW on Biometrics. (2011) 15–22
9. Baker, S., Kanade, T.: Hallucinating faces. In: AFGR. (2000) 83–88
10. Liu, C., Shum, H., Zhang, C.: A two-step approach to hallucinating faces: Global parametric model and local nonparametric model. In: CVPR. (2001)
11. Yang, J., Wang, Z., Lin, Z., Cohen, S., Huang, T.: Coupled dictionary training for image super-resolution. IEEE T-IP **21** (2012) 3467–3478
12. Gunturk, B., Batur, A., Altunbasak, Y., Hayes III, M., Mersereau, R.: Eigenface-domain super-resolution for face recognition. IEEE T-IP **12** (2003) 597–606
13. Hennings-Yeomans, P., Baker, S., Kumar, B.: Simultaneous super-resolution and feature extraction for recognition of low-resolution faces. In: CVPR. (2008)
14. Li, B., Chang, H., Shan, S., Chen, X.: Low-resolution face recognition via coupled locality preserving mappings. Signal Processing Letters **17** (2010) 20–23
15. Huang, H., He, H.: Super-resolution method for face recognition using nonlinear mappings on coherent features. IEEE T-NN **22** (2011) 121–130
16. Arandjelović, O., Cipolla, R.: A manifold approach to face recognition from low quality video across illumination and pose using implicit super-resolution. In: ICCV. (2007)
17. Jia, K., Gong, S.: Generalized face super-resolution. IEEE T-IP **17** (2008) 873–886
18. He, X., Niyogi, P.: Locality preserving projections. In: Advances in neural information processing systems 16. (2004) 153–160
19. Huang, G., Mattar, M., Berg, T., Learned-Miller, E.: Labeled faces in the wild: A database for studying face recognition in unconstrained environments. In: Workshop on Faces in 'Real-Life' Images: Detection, Alignment, and Recognition. (2008)
20. Belhumeur, P., Hespanha, J., Kriegman, D.: Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. IEEE T-PAMI **19** (1997) 711–720
21. Sugiyama, M.: Dimensionality reduction of multimodal labeled data by local fisher discriminant analysis. The Journal of Machine Learning Research **8** (2007) 1027–1061
22. Perona, P., Zelnik-Manor, L.: Self-tuning spectral clustering. In: Advances in neural information processing systems 17. (2004) 1601–1608
23. Lin, D., Tang, X.: Inter-modality face recognition. In: ECCV. (2006) 13–26
24. Zhu, M., Martinez, A.: Subclass discriminant analysis. IEEE T-PAMI **28** (2006) 1274–1286
25. Jolliffe, I., MyiLibrary: Principal component analysis. Volume 2. Wiley Online Library (2002)
26. Lu, J., Tan, Y.: Locality repulsion projections for image-to-set face recognition. In: ICME. (2011) 1–6