# A Benchmark and Comparative Study of Video-Based Face Recognition on COX Face Database

Zhiwu Huang, *Student Member, IEEE*, Shiguang Shan, *Senior Member, IEEE*, Ruiping Wang, *Member, IEEE*, Haihong Zhang, *Member, IEEE*, Shihong Lao, *Member, IEEE*, Alifu Kuerban, and Xilin Chen, *Senior Member, IEEE*

***Abstract*— Face recognition with still face images has been widely studied, while the research on video-based face recognition is inadequate relatively, especially in terms of benchmark datasets and comparisons. Real-world video-based face recognition applications require techniques for three distinct scenarios: 1) Video-to-Still (V2S); 2) Still-to-Video (S2V); and 3) Video-to-Video (V2V), respectively, taking video or still image as query or target. To the best of our knowledge, few datasets and evaluation protocols have benchmarked for all the three scenarios. In order to facilitate the study of this specific topic, this paper contributes a benchmarking and comparative study based on a newly collected still/video face database, named COX[1] Face DB. Specifically, we make three contributions. First, we collect and release a large-scale still/video face database to simulate video surveillance with three different video-based face recognition scenarios (i.e., V2S, S2V, and V2V). Second, for benchmarking the three scenarios designed on our database, we review and experimentally compare a number of existing set-based methods. Third, we further propose a novel Point-to-Set Correlation Learning (PSCL) method, and experimentally show that it can be used as a promising baseline method for V2S/S2V face recognition on COX Face DB. Extensive experimental results clearly demonstrate that video-based face recognition needs more efforts, and our COX Face DB is a good benchmark database for evaluation.**

Manuscript received October 28, 2014; revised March 31, 2015; accepted May 14, 2015. Date of publication October 27, 2015; date of current version November 5, 2015. This work was supported in part by the 973 Program under Contract 2015CB351802 and in part by the Natural Science Foundation of China under Contract 61390511, Contract 61173065, Contract 61222211, and Contract 61379083. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Lei Zhang. *(Corresponding author: Shiguang Shan.)*

Z. Huang, R. Wang, and X. Chen are with the Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences (CAS), Beijing 100190, China (e-mail: zhiwu.huang@vipl.ict.ac.cn; ruiping.wang@vipl.ict.ac.cn; xlchen@ict.ac.cn).

S. Shan is with the Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China, and also with the CAS Center for Excellence in Brain Science and Intelligence Technology, Shanghai, 200031 (e-mail: sgshan@ict.ac.cn).

H. Zhang is with OMRON Social Solutions Company, Ltd., Kyoto 619-0225, Japan (e-mail: angelazhang@ssb.kusatsu.omron.co.jp).

S. Lao is with Core Technology Center of OMRON Corporation, Kyoto, 619-0283, Japan (lao@ari.ncl.omron.co.jp).

A. Kuerban is with the College of Information Science and Engineering, Xinjiang University, Xinjiang 830046, China (e-mail: ghalipk@xju.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2015.2493448

[1]COX Face DB was constructed by Institute of Computing Technology, Chinese Academy of Sciences (**C**AS) under the sponsor of OMRON Social Solutions Co. Ltd. (**O**SS), and the support of **X**injiang University.

***Index Terms*— Video-based face recognition, video-to-still, still-to-video, video-to-video, COX Face DB, benchmarking, point-to-set correlation learning.**

## I. Introduction

RECENTLY, the ubiquitous use of video capturing devices is shifting the focus of face recognition research from image-based scenarios to video-based ones. Video-based face recognition identifies a subject with his/her video sequence as the query or target. Specifically, as shown in Table I, there are three distinct video-based face recognition scenarios. Among them, the V2S scenario matches a query video sequence against still face images such as mug shots, ID photos, driver license photos, etc., which are generally taken in controlled setting and thus generally of high quality. This scenario is very common requirement in watch list screening systems. On the contrary, the S2V scenario queries a still face image against a database of video sequences, which can be applied to locate a person of interest by searching his/her ID photo in the stored surveillance videos. The third scenario, i.e., the V2V case, queries a video sequence against a set of target video sequences, which can be exploited, for example, to track a person by matching his/her video sequence taken somewhere against the surveillance videos recorded elsewhere.

Although face recognition has been studied extensively in the literature, recognizing the subject in unconstrained face videos is still a field in its childhood. To our best knowledge, the state-of-the-art methods (e.g., [1]–[5]) still perform poorly on video face databases with real-world setting such as YouTube Celebrities [2], even though it only contains 47 subjects totally. Furthermore, only a few works (e.g., [6]–[9]) explore the V2S/S2V face recognition scenarios, which are however vital to a large number of practical applications such as mug-shot based watch list screening.

To measure the advance of one research problem, appropriate evaluation protocol defined on suitable database is one of the essential factors. In the past decades, a number of video face databases have been collected. Based on the list in the recent survey [10], here we enumerate more public video face databases in Table II, with brief summary of their key characteristics. Although these databases are publicly available to researchers, we find that they are not sufficient to support research on video-based face recognition. For example, quite

TABLE I
THREE DISTINCT SCENARIOS OF VIDEO-BASED FACE RECOGNITION

| Target set \ Query set | Still images | Video sequences |
|---|---|---|
| Still images | / | Video-to-Still (V2S) |
| Video sequences | Still-to-Video (S2V) | Video-to-Video (V2V) |

TABLE II
SELECTED VIDEO FACE DATASETS. "#SUB"/"#VID" RESPECTIVELY DENOTE THE NUMBER OF SUBJECTS/VIDEOS, "VARIATION" INCLUDE: VARYING POSE (p), ILLUMINATION (l), EXPRESSION (e), RESOLUTION (r), MOTION BLUR (b), WALKING (w)

| Databases | #Sub, #Vid | Variation | Scenario |
|---|---|---|---|
| CMU MoBo [15] | 25, 150 | w | V2V |
| First Honda/UCSD [16] | 20, 75 | p | V2V |
| Sec. Honda/UCSD [17] | 15, 30 | p | V2V |
| CMU FIA [18] | 214, 214 | p,l,e | V2V |
| CamFace [19] | 100, 1400 | p,l | V2V |
| Faces96 [20] | 152, 152 | l,r | V2V |
| VidTIMIT [21] | 43, 43 | p,e | V2V |
| YouTube Celebrities [2] | 47, 1910 | p,l,e,r,b,w | V2V |
| MBGC [22] | 821, 3764 | p,l,e,r,b,w | V2S/V2V |
| ND-Flip-QO [23] | 90, 14 | l,e,r,b | V2V |
| YouTube Faces DB [13] | 1595, 3425 | p,l,e,r,b,w | V2V |
| Chokepoint [24] | 29, 48 | p,l,r,b,w | V2V |
| ScFace [11] | 130, 910 | p,l,r | V2S/V2V |
| UT Dallas [12] | 284, 1016 | p,l,e,r,b,w | V2S/V2V |
| UMD Comcast10 [25] | 16, 12 | p,l,r,b,w | V2V |
| PaSC [14] | 265, 2802 | p,l,e,r,b,w | V2S/V2V |
| Celebrity-1000 [26] | 1000, 7021 | p,l,e,r,b,w | V2V |
| **Our COX Face DB** | **1000**, **3000** | **p,l,e,r,b,w** | **V2S/S2V/V2V** |

a few datasets involve a small number of subjects or consist of a small amount of video sequences. Some video databases even contain only a limited number of frames for each subject. For instance, ScFace [11] selects only 3 frames out of each video sequence. Additionally, videos in currently available databases such as UT Dallas [12] are usually taken by the same camera. Moreover, in terms of evaluation protocol, most of them focus on either V2S/S2V or V2V face recognition scenario. For example, YouTube Faces DB [13] is designed only for V2V identification scenario while PaSC [14] dataset is mainly collected for V2S and V2V face verification tasks. We also notice that, in terms of database scale (the number of subjects), only YouTube Face DB [13] and Celebrity-1000 contain videos of more than 1,000 subjects. Furthermore, most of the subjects in almost all the existing databases are not from Asia, which forms heavy bias in skin color.

In this paper, aiming to simulate video surveillance scenario, we collect a large scale face dataset named *COX Face DB*, for the evaluation of three video-based face recognition tasks, i.e. V2S, S2V, and V2V (as shown in Table I). Compared with most existing datasets, COX Face DB contains more subjects (1,000), more video sequences (3,000), and more frames with natural variations in pose, expression, lighting, blur, and face resolution. In addition, all the subjects are Chinese, which makes it complementary to the existing ones. More importantly, the COX videos were captured by 3 camcorders, while the still frontal face images were taken by a digital camera of high quality.

An early version of this dataset, named COX-S2V, was released in [8]. Compared with that early one, the updated COX Face DB in this paper contains additional videos captured from a third camcorder, and the video sequences are made longer to include more video frames. Furthermore, the new COX Face DB adds two different scenarios of video-based face recognition (i.e., S2V and V2V scenarios) with both face identification and face verification protocols for wider range of evaluations.

Our second contribution is a comparative study of a number of set-based methods for video-based face recognition, by evaluating them on the released COX Face DB. Broadly speaking, video-based face recognition methods can be categorized into sequence-based ones and set-based ones [10]. The former methods (e.g., [2], [6], [27]–[29]) exploit the temporal or dynamic information of the face in the video, while the latter (e.g., [1], [3]–[5], [30], [31]) represent video as image set of the separated video frames, without using the temporal information. Since the existing set-based methods make fewer assumptions on face video sequence and commonly achieve more appealing performance, in this paper, we focus on this category of methods and carry out a comparative study on them on our database.

Based on the comparative study, we find only a small number of set-based methods are suggested for the problem of V2S/S2V face recognition, where still and video are usually treated as point and set respectively. Therefore, to accommodate most of statistical models that are usually employed in set-based methods, we further make a third contribution by proposing a new approach named Point-to-Set Correlation Learning (PSCL) for V2S/S2V face recognition task. Since usual set models such as linear subspace typically reside on a specific type of Riemannian manifold [32]–[34], inspired by our previous work [35], we formulate this kind of tasks as a problem of learning correlations between Euclidean points and Riemannian elements. To handle this problem, by deriving kernel functions on certain types of Riemannian manifolds, our proposed method exploits Kernel Canonical Correlation Analysis (KCCA) [36] to learn maximum correlations between the heterogeneous data (i.e. Euclidean points vs. Riemannian elements). The final extensive experimental results on our COX Face DB show that our method has achieved a promising baseline for the V2S/S2V benchmark tests.

The rest of the paper is organized as follows. Section II describes the released COX Face DB. Section III surveys the state-of-the-art set-based methods for video-based face recognition. Section IV presents the new Point-to-Set Correlation Learning method for the problem of V2S/S2V face recognition. Section V reports comparative testing results on our COX Face DB for both V2S/S2V and V2V face recognitions, followed by conclusions in Section VI.

## II. DESCRIPTION OF COX FACE DB

In this section, we describe our COX Face DB in details, including how the still face images and face videos were collected, how the data is further processed, and the evaluation protocols accompanying the released data.

### *A. Data Collection: Image/Video Capturing Environment, Equipment Setup and Recording Procedure*

As mentioned previously, we aim at a new face database for better evaluating three different scenarios of video-based face recognition, especially for applications like video surveillance. For this purpose, we need to take videos as wildly as possible, to include rich variations in face pose, facial expression, face resolution, and environmental lighting, as well as image qualities in terms of noise and blurring. For V2S/S2V scenarios, we also need to take still face images of each subject to simulate ID photo. With the above considerations in mind, we elaborately designed the image/video collecting procedure, which is described in the following.

*1) Data Collecting Venue:* The videos were taken in a large gym with very high ceilings. The gym has one side wall made of transparent glass. Besides good for the volunteers and devices, such a gym forms an imaging environment with complex half indoor and half outdoor lighting in the daytime. Also, a small part of the videos is collected in the night with strong fluorescent on ceilings open. So, overall speaking, the lighting on the faces in our database is natural and very close to many practical applications.

*2) Data Collecting Devices:* Totally, the devices we used include three digital camcorders and one digital camera. To capture high quality still image of each subject, Canon EOS 500D DC was exploited. Please refer to the following part for how the still images are captured. For video acquisition, we used three SONY HDR-CX350E DV camcorders, which were fixed on tripods about 2 meters high. For more details of how they are mounted, readers are referred to the video capturing part.

*3) Still Image Capturing:* For each subject, still face images were taken with the DC. To capture ID photo like images, the DC is mounted on a tripod about 3 meters away from the subjects, who was asked to sit on a chair with face upright and neutral expression. The photographing room was set up with standard indoor lighting, and the flash of the DC was always used to alleviate shadows due to top lighting. One example of still image is shown in Fig. 2 (a).

*4) Video Capturing:* To simulate surveillance video, we took video of every subject when he/she was walking. To include more variations in the facial appearance, we elaborately pre-designed the walking route, as well as the mounting of the cameras. Specifically, as shown in Fig. 1, the subjects were asked to walk from the starting point to the end point, roughly along the S-shape route, as freely as possible at regular speed. Three camcorders, indexed as Cam1, Cam2 and Cam3, were mounted at 3 fixed locations of about 2 meters high. As shown in Fig. 2 (b), (c), (d), they could respectively capture video of the walking subject when he/she was on the route marked in red, green, and blue.

It is easy to understand, such an S-shape route is good for including more face appearance variations in pose, lighting, blur, and face resolution. Especially, as the subject walked along the two semicircles, he/she was naturally changing his/her face orientations continuously, which leads to changing face poses, as well as varying lighting thanks to outdoor lighting through the big glass wall.

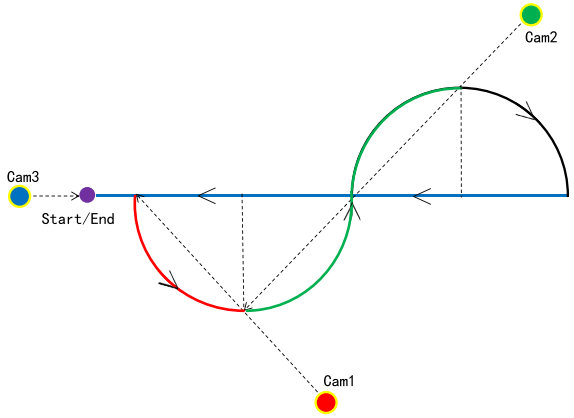



Fig. 1. Walking route of subjects and camcorders setting. Every subject was asked to walk freely from the starting point to the end point, roughly along the S-shape route. Three camcorders, Cam1, Cam2 and Cam3, were placed at 3 fixed locations, respectively capturing video of the subject when walking on the route marked in red, green, and blue. The radius of the two semicircles in the S-shape is 3 meters.


Fig. 2. Example still images collected by our digital camera (a) and example video frames from video clips captured by three camcorders Cam1 (b), Cam2 (c) and Cam3 (d). In (b), (c) and (d), the $t$ value in red in each frame indicates the index of the frame in the video sequence.

### *B. Data Processing and Formatting*

With the still images and videos of all the 1,000 subjects captured, we further process the data in order to facilitate future evaluation. This process is necessary because the camcorders were kept power on during the video recording, which implies many void frames without faces of the subject. In addition, for the purpose of convenient evaluation, it is desirable that all the video frames contain only the head-shoulder or even only the head part of the subject.

To satisfy the above goals, we first manually truncated the long videos recorded by each camcorder into shorter video clips, each of which starts from the appearance of one subject until his/her disappearance (coarsely corresponding to

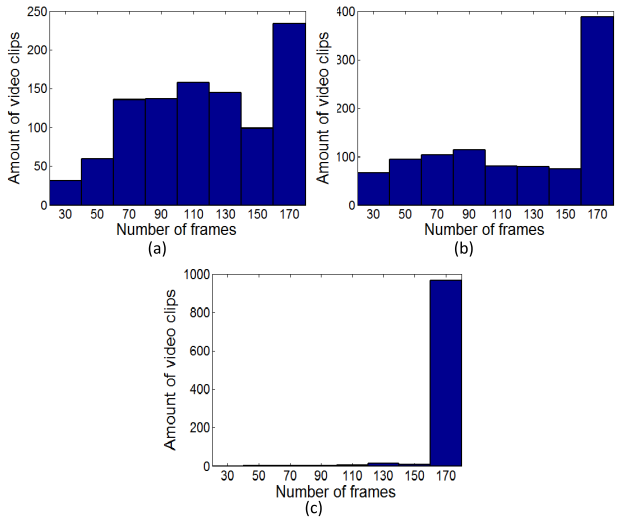



Fig. 3. Statistics of the number of frames in the video clips recorded by 3 different camcorders respectively. (a) Cam1. (b) Cam2. (c) Cam3.

the viewing field as shown in Fig.2). This procedure resulted in 3,000 video clips. Then, we ran a commercialized face detector, OKAO,[2] to detect the faces in the video clips. However, as the face detector is not perfect, it might generate some inaccurate or even incorrect detections. For the convenience of later process, we exploit a simple tracking-like strategy to remove possible outlier detections. Simply speaking, if the center of the detected face in one frame is too far away from those in its previous frames, the face in this frame will be removed as an outlier. This processing unavoidably leads to loss of a small number of video frames, which we think does no hurt to the evaluation. It is also worth pointing out that, as the walking speeds were different from subject to subject, the durations of the video clips of different subjects are different even for the same camera, which also accounts for the varying number of video frames in this dataset. In Fig. 3, we statistically show the number of frames in the videos respectively for three different camcorders, from which we can see that most of the video clips have more than 100 frames per subject and especially video clips from Cam3 mostly have 170 frames.

Another processing relates to the reduction of the memory storing the final database. The video recorded originally is of high resolution, i.e., $1920 \times 1080$, which leads to too large memory requirement preventing wide distribution. However, the video camcorders actually worked in interlaced mode; therefore, we "de-interlaced" the videos by a commercialized tool, Aunsoft Final Mate.[3] Then, to further reduce the size of the database, for each frame with a face detected, an image patch centered at the head of the subject is cropped out. In this process, we carefully avoided any geometry transform which might lead to pixel interpolation. In other words, all the intensities of the pixels were directly taken from the uncompressed video frames. Finally, the cropped head-centered images are saved as JPG files, in one folder per subject per camcorder. Due to the above process, the sizes of the cropped face images vary a lot, from the smallest $66 \times 66$ pixels to the largest $798 \times 798$ pixels.

[2]http://www.omron.com/r_d/technavi/vision/okao/detection.html
[3]http://www.aunsoft.com/final-mate/

TABLE III
AGE DISTRIBUTION OF THE SUBJECTS IN THE DATABASE

| Age | 10-19 | 20-29 | 30-39 | 40-49 | 50-59 | 60-69 | 70-79 |
|---|---|---|---|---|---|---|---|
| Num. | 19 | 933 | 23 | 15 | 6 | 3 | 1 |

For convenient evaluation, the video frames and the still images are all well named. Specifically, the video frames are named in the way of "subjectID-camIndex-frameIndex.jpg", while the still facial images are named as "subjectID-frontal.jpg". In this way, every image/frame in the database was given a unique name with information about the subject's unique ID. In addition, the name of every video frame also includes information about camcorder index, and frame index. For example, the image named "201103180001-1-0133.jpg" is the 133th frame of the video clip captured by Cam1 for the subject with ID "201103180001". All the data is organized in a simple file structure. Specifically, all the 1,000 still images are in one folder named "still", while all the video frames are in another one folder named "video" with three sub-folders, i.e., cam1, cam2, and cam3, each of which has 1,000 sub-folders named after the subject ID and containing all the video frames of that subject.

### C. Basic Demographics of the Subjects in the Database

The data collecting was conducted in Xinjiang University. So, most of the subjects in the database are students, teachers or professors of the university and nearby residents. Among the 1,000 volunteers, 435 are males and 565 are females. And, coarsely, half of them are Mongolian and the other half are Caucasian. As a feature of university volunteers, most of the subjects are young people, as indicated by the age distribution in Table III.

### D. COX Face DB Evaluation Protocol for Video-Based Face Recognition

As our goal is an evaluation protocol for assessing video-based face recognition algorithms, we intend to design protocols respectively for three different video-based face recognition scenarios, i.e., V2S, S2V, and V2V. For most video surveillance applications, both face identification and face verification are required. Therefore, we designed protocols for both face identification and face verification evaluations. Specifically, the identification performance measurement is defined as Rank-1 recognition rate while the verification rate can be reported by Receiver Operating Characteristic (ROC) curve. In keeping with the verification protocol used in previous challenge problem [14], face recognition algorithms must compute a similarity matrix for all pairs of images/videos obtained by matching images/videos in a target set and images/videos in a query set. The resulting similarity matrix becomes the basis for subsequent analysis, enabling performance to be expressed in terms of an ROC curve.

TABLE IV
TRAINING AND TESTING CONFIGURATION FOR **V2S** SCENARIO

| | | #subjects | #images | #videos | Image/Video Sources | Note |
|---|---|---|---|---|---|---|
| Training | | 300 | 300 | 900 | DC, Cam 1, 2, 3 | |
| Testing | Target set | 700 | 700 | - | DC | |
| | Query set 1 | 700 | - | 700 | Cam 1 | V1-S |
| | Query set 2 | 700 | - | 700 | Cam 2 | V2-S |
| | Query set 3 | 700 | - | 700 | Cam 3 | V3-S |

TABLE V
TRAINING AND TESTING CONFIGURATION FOR **S2V** SCENARIO

| | | #subjects | #images | #videos | Image/Video Sources | Note |
|---|---|---|---|---|---|---|
| Training | | 300 | 300 | 900 | DC, Cam 1, 2, 3 | |
| Testing | Target set 1 | 700 | - | 700 | Cam 1 | S-V1 |
| | Target set 2 | 700 | - | 700 | Cam 2 | S-V2 |
| | Target set 3 | 700 | - | 700 | Cam 3 | S-V3 |
| | Query set | 700 | 700 | - | DC | |

We also expect to guarantee fair comparison of different methods. Therefore, in our protocol, for each evaluation scenario, we strictly define the training set and testing set. It is worth noting that, here by "training set" we mean the set of face images is used to train a feature extractor or classification model, while testing set is used to form the target set (or gallery) and query set (or probe). Here, one question is how to determine the ratio of the training and the testing data. While more training data commonly leads to better performance, more testing data implies more accurate performance estimation. Furthermore, for face identification, more subjects in the target set are desirable. Therefore, in our protocols, relatively more data is used for testing. Specifically, in our protocols, we empirically set the ratio as 3:7, i.e., data of 300 subjects selected randomly can be used for model training while that of the remaining 700 subjects for testing. For all the evaluations, 10 groups of random 300/700 partition are predefined accompanying our database. Correspondingly, we hope all the future testing on this database should report the mean and the standard deviation of the face recognition results of the 10 runs. For face verification, following the work [13], to depict the ROC curve, we require the final similarity matrix to be constructed by concating 10 similarity matrices which are respectively computed in 10 runs.

With the common parts described above, we now define the particular part of the distinct protocol for each of the three video-based face recognition scenarios, i.e., V2S, S2V, and V2V face recognition.

*1) Evaluation Protocol for V2S Scenario:* As defined in Table I, in the V2S scenario, the target set contains still images of the persons with known identities, while the query samples are video clips of faces to be recognized, generally by matching against the target still face images. Therefore, for this scenario, we designed the protocol with the training and testing data configured as in Table IV. Specifically, the videos taken by three camcorders form three separated experiments, i.e., V1-S, V2-S, and V3-S. The 10 random partitions of the 300/700 subjects for training and testing are given in the "V2S partitions" folder of the released database.

TABLE VI
TRAINING AND TESTING CONFIGURATION FOR **V2V** SCENARIO

| | | #subjects | #videos | Image/Video Sources | Note |
|---|---|---|---|---|---|
| Training | | 300 | 900 | Cam 1, 2, 3 | |
| Testing | Target set<br>Query set | 700 | 700<br>700 | Cam 1<br>Cam 2 | V2-V1 |
| | Target set<br>Query set | 700 | 700<br>700 | Cam 1<br>Cam 3 | V3-V1 |
| | Target set<br>Query set | 700 | 700<br>700 | Cam 2<br>Cam 3 | V3-V2 |
| | Target set<br>Query set | 700 | 700<br>700 | Cam 2<br>Cam 1 | V1-V2 |
| | Target set<br>Query set | 700 | 700<br>700 | Cam 3<br>Cam 1 | V1-V3 |
| | Target set<br>Query set | 700 | 700<br>700 | Cam 3<br>Cam 2 | V2-V3 |

*2) Evaluation Protocol for S2V Scenario:* Compared with V2S scenario, the target set of the S2V scenario conversely contains videos while the queries are still face images. Therefore, as shown in Table V, we can also form three different experiments, i.e., S-V1, S-V2, and S-V3, according to the source (/camcorder) of the video in the target set. Similarly, the 10 random partitions of the 300/700 subjects for training and testing are given in the "S2V partitions" folder of the released database.

*3) Evaluation Protocol for V2V Scenario:* To form V2V evaluations, for either target set or query set, we have 3 videos per subject respectively from Cam 1, Cam 2, and Cam 3. Therefore, they can mutually form 6 experiments, as shown in Table VI. Alternatively, we can also setup more experiments by taking one or two of the three videos to form the target set while keeping the remaining as the queries, which is not considered in this work in order to alleviate the evaluation burden. Similarly, the 10 random partitions of the 300/700 subjects for training and testing are given in the "V2V partitions" folder of the released database.

## III. REVIEW OF EXISTING SET-BASED METHODS FOR VIDEO-BASED FACE RECOGNITION

In this section, we briefly review previous set-based methods for video-based face recognition. These methods generally represent each still image with a feature vector (i.e., a point

TABLE VII

SELECTED SET-BASED METHODS FOR THE PROBLEM OF **V2S/S2V** FACE RECOGNITION. ACCORDING TO THE USED SET MODEL TYPE, THESE METHODS ARE GROUPED INTO FOUR CATEGORIZATIONS TO SOLVE THIS PROBLEM BY DEFINING DIFFERENT POINT-TO-SET DISTANCES

| Set model type | Literature source | Method abbr. | Point-to-set distance |
|---|---|---|---|
| Linear subspace | Chien *et al.*, PAMI'2002 [37] | NFS | Distance-from-feature-space |
| Nonlinear manifold | Wang *et al.*, CVPR'2008 [3] | PMD | Distance-from-feature-space |
| Affine/convex hull | Vincent *et al.*, NIPS'2001 [38] | HKNN | K-local-hyperplane distance |
| | Vincent *et al.*, NIPS'2001 [38] | CKNN | K-local-convex distance |
| | Cevicalp *et al.*, ICML'2008 [39] | NHD | Nearest-hyperdisk distance |
| | Zhu *et al.*, ICCV'2013 [40] | PSDML | Point-to-l2-affine-hull distance |
| Probabilistic model | Mahalanobis, NISI'1936 [41] | – | Mahalanobis distance |

TABLE VIII

SELECTED SET-BASED METHODS FOR THE PROBLEM OF **V2V** FACE RECOGNITION. ACCORDING TO THE USED SET MODEL TYPE, THESE METHODS ARE GROUPED INTO FOUR CATEGORIZATIONS TO SOLVE THIS PROBLEM BY DEFINING DIFFERENT SET-TO-SET DISTANCES

| Set model type | Literature source | Method abbr. | Set-to-set distance |
|---|---|---|---|
| Linear subspace | Yamaguchi *et al.*, FG'1998 [30] | MSM | Subspace canonical correlation |
| | Fukui *et al.*, ISRR'2003 [42] | CMSM | Subspace canonical correlation |
| | Kim *et al.*, PAMI'2007 [1] | DCC | Subspace canonical correlation |
| | Hamm *et al.*, ICML'2008 [32] | GDA | Subspace projection distance |
| | Harandi *et al.*, CVPR'2011 [43] | GGDA | Subspace canonical correlation/projection distance |
| Nonlinear manifold | Wang *et al.*, CVPR'2008 [3] | MMD | Subspace canonical correlation/exemplar distance |
| | Wang *et al.*, CVPR'2009 [44] | MDA | Exemplar distance |
| | Cui *et al.*, CVPR'2012 [45] | – | Subspace canonical correlation |
| | Chen *et al.*, CVPR'2013 [5] | SANS | Subspace canonical correlation |
| Affine/convex hull | Cevicalp *et al.*, CVPR'2010 [4] | AHISD | Inter-hull distance |
| | Cevicalp *et al.*, CVPR'2010 [4] | CHISD | Inter-hull distance |
| | Hu *et al.*, CVPR'2011 [46] | SANP | Inter-hull distance |
| | Yang *et al.*, FG'2013 [47] | RNP | Inter-hull distance |
| | Zhu *et al.*, ICCV'2013 [40] | SSDML | Inter-hull distance |
| | Zhu *et al.*, IFS'2014 [48] | ISCRC | Inter-hull distance |
| Probabilistic model | Shakhnarovich *et al.*, ECCV'2002 [49] | – | Gaussian distribution Kullback-Leibler divergence |
| | Arandjelović *et al.*, CVPR'2005 [50] | MDM | Gaussian distribution Kullback-Leibler divergence |
| | Wang *et al.*, CVPR'2012 [34] | CDL | SPD matrix Log-Euclidean distance |
| | Vemulapalli *et al.*, CVPR'2013 [51] | – | SPD matrix Log-Euclidean distance |
| | Lu *et al.*, ICCV'2013 [52] | LMKML | SPD matrix Euclidean distance |

in Euclidean space) while treating each video as an image set (i.e., a set of points in Euclidean space). According to the survey [10], in addition to some super-resolution techniques [53], [54], 3D modeling methods [55]–[57] and frame selection algorithms [9], [58], [59], a large proportion of set-based methods focus on the statistical modeling, which can efficiently capture the pattern variations (such as changes of poses, expressions and illumination) of face in videos. The core of this kind of set-based methods is how to model the set and how to measure the point-to-set and set-to-set distance respectively for V2S/S2V and V2V face recognition. As summarized in Table VII (V2S/S2V case) and Table VIII (V2V case), according to the corresponding set model type, we classify these methods into four categories, i.e., linear subspace, nonlinear manifold, affine/convex hull, and probabilistic model based methods.

### A. Linear Subspace Based Methods

As well recognized, face images form a low dimensional face subspace in the image space. Similarly, it should be true that face images of a specific person lie in a lower dimensional subspace in the face subspace. So, it is a natural choice to represent the face images from a video sequence as a linear subspace, as done by Yamaguchi *et al.* [30]. In the following, we briefly describe how it can be used to compute point-to-set distance and set-to-set distance respectively for V2S/S2V and V2V face recognition scenarios.

*1) For V2S/S2V Case:* For the problem of face recognition, Chien and Wu [37] proposed a Nearest Feature Space (NFS) classifiers to measure the point-to-subspace distance, which is equal to the well-known distance-from-feature-space (DFFS) [60]–[62]:

$$d(\boldsymbol{x}, \mathcal{S}) = \min_{\boldsymbol{y} \in \mathcal{S}} \|\boldsymbol{x} - \boldsymbol{y}\| = \|\boldsymbol{x} - \boldsymbol{x}'\|. \tag{1}$$

where $\boldsymbol{x}'$ is the projection of $\boldsymbol{x}$ in the subspace $\mathcal{S}$, also the nearest point to $\boldsymbol{x}$ in $\mathcal{S}$. With this defined point-to-set distance, NFS can be applied to V2S/S2V face recognition.

*2) For V2V Case:* Yamaguchi *et al.* [30] exploited the canonical correlations [63] between two linear subspaces to calculate the set-to-set distance. With the linear subspace modeling, Kim *et al.* [1] developed a Discriminant Canonical Correlations (DCC) method by maximizing the canonical correlations of within-class subspaces and minimizing those of between-class subspaces to perform subspace-based classification more discriminatively. Hamm and Lee [32] treated the subspaces of fixed dimensionality as elements on Grassmann manifold, and then applied the projection metric between subspaces for classification on Grassmann manifold in a kernel LDA framework. Based on this work, Harandi *et al.* [43] further improved the classification on Grassmann manifold

by simultaneously exploiting the canonical correlation and the projection distance between the subspaces in a more generalized graph-embedding learning framework.

### B. Nonlinear Manifold Based Methods

In most video-based face recognition scenarios, faces are recorded with rich appearance variations due to changes in pose, expression, lighting, etc. This will may make linear subspace methods lose their advantage in the aspect of set modeling. Therefore, several works [3], [5], [44], [45] explored the nonlinear manifold to approximate the nonlinear variations of a set of face images. Specifically, in these works, each nonlinear manifold is partitioned into a number of local linear models (i.e., linear subspaces) by employing the criterion of Maximal Linear Patch. In the following, we briefly introduce how manifold models can be used to compute point-to-set distance and set-to-set distance respectively for different scenarios of video-based face recognition.

*1) For V2S/S2V Case:* By applying DFFS [60]–[62], Wang *et al.* [3] defined a point-to-manifold distance (PMD) to match points against sets. With this nolinear manifold modeling, PMD is defined as the smallest point-to-subspace distance of any pairs of point and linear subspace as in the following:

$$d(\boldsymbol{x}, \mathcal{M}) = \min_{1 \le i \le m} d(\boldsymbol{x}, \mathcal{S}_i). \tag{2}$$

where $m$ is the number of local linear subspaces that constitute the manifold $\mathcal{M}$, $\mathcal{S}_i$ is the $i$-th linear subspace, and $d(\boldsymbol{x}, \mathcal{S}_i)$ is computed by Eq.1.

*2) For V2V Case:* Wang *et al.* [3] measured manifold-to-manifold distance by integrating the subspace-to-subspace distances between pair-wise subspaces respectively from two compared manifolds. In their method, the subspace-to-subspace distance is defined in a form of weighted average of canonical correlations between two subspaces and cosine similarities of exemplars in them. To further improve the manifold-to-manifold matching, Cui *et al.* [45] and Chen *et al.* [5] proposed effective and efficient algorithms for the problem of the alignment of the subspaces from two matching manifolds when calculating their canonical correlations.

### C. Affine/Convex Hull Based Methods

In real-world scenario, surveillance video sequences are commonly very short but cover large and complex data variations. In this kind of setting, linear subspace and nonlinear manifold methods typically fail to work. To handle this problem, several works [4], [38], [39] introduced the affine/convex hull set modeling, which approximates exemplars of each image set by the affine/convex combination of its samples. This typically gives a rather loose approximation to the data of image sets and is insensitive to the positions of the samples within the affine/convex hull [4].

*1) For V2S/S2V Case:* Vincent and Bengio [38] proposed K-local Hyperplane (Convex) Distance Nearest Neighbor algorithm (HKNN/CKNN) to measure the point-to-affine-hull/point-to-convex-hull distance. In addition, Cevikalp *et al.* [39] developed Nearest Hyperdisk method (NHD) to define the distance of point-to-hyperdisk, which models set with intersection of affine hull and bounding hypersphere. More recently, Zhu *et al.* [40] employed the l2-norm regularized affine hull defined in [47], and then proposed a novel metric learning framework to learn appropriate point-to-l2-affine-hull distance.

*2) For V2V Case:* Given two hulls, Cevikalp and Triggs [4] defined the inter-hull distance as the distance between the nearest virtual points in two comparing hulls via a convex optimization. For more robust classification, Hu *et al.* [46] and Yang *et al.* [47] extended this work by adding sparse (i.e., $l_1$ based) or collaborative (i.e., $l_2$ based) regularizations to calculate the inter-hull distance. By employing the basic inter-hull distance, Zhu *et al.* [40] proposed a novel distance metric learning for set-to-set classification. In addition, by modeling the query set as a convex or regularized hull, Zhu *et al.* [48] developed a novel image set based collaborative representation and classification (ISCRC) method to solve the problem of set-to-set classification.

### D. Probabilistic Model Based Methods

In the literature, probabilistic models have also been employed to represent image set due to their appealing property for characterizing the set data distribution. For instance, Shakhnarovich *et al.* [49] and Arandjelović *et al.* [50] represented image set with some well-studied probability density functions. More recently, Wang *et al.* [34] modeled each image set by its second-order statistics, i.e. covariance matrix, to fit the set data tightly and shows strong resistance to outliers. These models are exploited in video-based face recognition as described in the following.

*1) For V2S/S2V Case:* As can be seen from Table VII, few methods explicitly defined point-to-set distance for probabilistic models, but this does not mean they are not applicable. For instance, given Gaussian model of a set, point-to-set distance can be naturally defined based on the probability of the point. For covariance models, it is also natural to define the point-to-covariance distance with Mahalanobis distance:

$$d(\boldsymbol{x}, \mathcal{C}) = (\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{C}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}). \tag{3}$$

where the model $\mathcal{C}$ defines both the covariance matrix $\boldsymbol{C}$ and the mean $\boldsymbol{\mu}$ of the set.

*2) For V2V Case:* Shakhnarovich *et al.* [49] and Arandjelovic *et al.* [50] applied the Kullback-Leibler divergence (KLD) to measure the distance between Gaussian distributions. Wang *et al.* [34] employed the Log-Euclidean distance to measure the distances between covariance matrices and then classified them on Riemannian manifold, where they reside on. Based on the Log-Euclidean distance, Vemulapalli *et al.* [51] applied multiple kernel learning techniques to exploit more discriminant information between covariance matrices for more robust set-to-set classification.

## IV. A New Point-to-Set Correlation Learning Method for V2S/S2V Face Recognition

As it can be found in previous descriptions, relatively fewer works have been dedicated to V2S/S2V scenarios, which

does not match their importance in real-world applications. So, this paper further proposed a new point-to-set classification mathod especially for benchmarking V2S/S2V face recognition.

### A. Overview

As mentioned in the previous section, the V2S/S2V face recognition is actually a point-to-set classification task. To tackle this task, we study a Point-to-Set Correlation Learning (PSCL) framework to accommodate most of statistical models of set data. Inspired by our previous work [35], we formulate this task as a problem of learning correlations between Euclidean points and Riemannian elements, which respectively represent the images as points in Euclidean space $\mathbb{R}^D$ and the usual set models as elements on a specific Riemannian manifold $\mathcal{M}$. According to the type of set modeling, we follow [35] to study three different Euclidean-to-Riemannian matching cases, where the Riemannian elements are linear subspaces, affine subspaces and covariance matrices respectively. To more directly handle the three heterogeneous matching cases compared with [35], we exploit the well-known Kernel Canonical Correlation Analysis (KCCA) [36] to learn the maximum correlations between heterogeneous data for robust V2S/S2V face recognition.

### B. Proposed Method

Canonical Correlation Analysis (CCA), proposed by Hotelling [63], is a well-known method of correlating linear relationships between two sets of variables. Suppose we have two sets of data $\boldsymbol{X}$ and $\boldsymbol{Y}$, CCA learns two different projections $\boldsymbol{W}_x, \boldsymbol{W}_Y$ such that the correlation between the projections $\boldsymbol{W}_x^T\boldsymbol{X}$ and $\boldsymbol{W}_Y^T\boldsymbol{Y}$ of the variables onto these basis vectors is mutually maximized:

$$\rho = \max_{\boldsymbol{W}_x, \boldsymbol{W}_Y} \frac{\boldsymbol{W}_x^T \boldsymbol{X}\boldsymbol{Y}^T \boldsymbol{W}_Y}{\sqrt{\boldsymbol{W}_x^T \boldsymbol{X}\boldsymbol{X}^T \boldsymbol{W}_x \boldsymbol{W}_Y^T \boldsymbol{Y}\boldsymbol{Y}^T \boldsymbol{W}_Y}} \tag{4}$$

where the maximum canonical correlation is the maximum of $\rho$ with respect to $\boldsymbol{W}_x$ and $\boldsymbol{W}_Y$.

As a kernel extension of CCA, KCCA [36] has been applied to learn correlations in higher dimensional feature spaces. Here we provide a brief description on how it is applied in our case. Formally, given one collection of data for points $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m\}$ ($\boldsymbol{x}_i \in \mathbb{R}^D, 1 \leq i \leq m$) and one collection of data for set models $\boldsymbol{Y} = \{\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots, \boldsymbol{Y}_m\}$ ($\boldsymbol{Y}_j \in \mathcal{M}, 1 \leq j \leq m$), we consider the V2S/S2V face recognition scenario addressed in this paper.

Similar to other traditional kernel learning method, KCCA also employs the "kernel trick" to implicitly map the heterogeneous data into two higher dimensional Reproducing Kernel Hilbert Spaces (RKHS) $\mathcal{H}_x, \mathcal{H}_Y$ respectively. Here, we define the two implicit maps as $\phi_x : \mathbb{R}^D \mapsto \mathcal{H}_x$ and $\phi_Y : \mathcal{M} \mapsto \mathcal{H}_Y$. With these two maps, the corresponding kernels for points and set models are two functions, such that all $\boldsymbol{x}_i, \boldsymbol{x}_j \in \boldsymbol{X}$, $\boldsymbol{Y}_i, \boldsymbol{Y}_j \in \boldsymbol{Y}$ satisfy: $\boldsymbol{K}_x(\boldsymbol{x}_i, \boldsymbol{x}_j) = \langle \phi_x(\boldsymbol{x}_i), \phi_x(\boldsymbol{x_j}) \rangle$, $\boldsymbol{K}_Y(\boldsymbol{Y}_i, \boldsymbol{Y}_j) = \langle \phi_Y(\boldsymbol{Y}_i), \phi_Y(\boldsymbol{Y}_j)$. Then the two different directions $\boldsymbol{W}_x, \boldsymbol{W}_Y$ in Eq.4 can be rewritten as the projection of the data onto the direction $\boldsymbol{\alpha}, \boldsymbol{\beta}$:

$$\begin{aligned} \boldsymbol{W}_x &= \boldsymbol{X}^T\boldsymbol{\alpha}, \\ \boldsymbol{W}_Y &= \boldsymbol{Y}^T\boldsymbol{\beta}. \end{aligned} \tag{5}$$

Let $\boldsymbol{K}_x = \boldsymbol{X}\boldsymbol{X}^T$, $\boldsymbol{K}_Y = \boldsymbol{Y}\boldsymbol{Y}^T$ be the kernel matrices corresponding to the two representations and substitute Eq.5 into Eq.4, KCCA seeks to solve the optimization [36]:

$$\begin{aligned} \rho &= \max_{\boldsymbol{\alpha}, \boldsymbol{\beta}} \frac{\boldsymbol{\alpha}^T \boldsymbol{X}\boldsymbol{X}^T \boldsymbol{Y}\boldsymbol{Y}^T \boldsymbol{\beta}}{\sqrt{\boldsymbol{\alpha}^T \boldsymbol{X}\boldsymbol{X}^T \boldsymbol{X}\boldsymbol{X}^T \boldsymbol{\alpha}\boldsymbol{\beta}^T \boldsymbol{Y}\boldsymbol{Y}^T \boldsymbol{Y}\boldsymbol{Y}^T \boldsymbol{\beta}}} \\ &= \max_{\boldsymbol{\alpha}, \boldsymbol{\beta}} \frac{\boldsymbol{\alpha}^T \boldsymbol{K}_x \boldsymbol{K}_Y \boldsymbol{\beta}}{\sqrt{\boldsymbol{\alpha}^T \boldsymbol{K}_x^2 \boldsymbol{\alpha}\boldsymbol{\beta}^T \boldsymbol{K}_Y^2 \boldsymbol{\beta}}} \end{aligned} \tag{6}$$

As shown in the study [36], the optimal $\boldsymbol{\alpha}$ in Eq.6 is given by the leading eigenvectors of the generalized eigenproblem of the form $\boldsymbol{I}\boldsymbol{\alpha} = \lambda^2\boldsymbol{\alpha}$, where $\boldsymbol{I}$ is identity matrix. With $\boldsymbol{\alpha}$ solved, the optimal $\boldsymbol{\beta}$ in Eq.6 is then calculated by $\boldsymbol{\beta} = \frac{\boldsymbol{K}_Y^{-1}\boldsymbol{K}_x\boldsymbol{\alpha}}{\lambda}$. For more details to solve this optimization problem Eq.6, please refer to [36].

Now, we need to study how to define the kernel functions $\boldsymbol{K}_x, \boldsymbol{K}_Y$ for points and set models. For the Euclidean points, the Radial Basis Function (RBF) kernel, a positive definite kernel, is exploited. Formally, given a pair of points $\boldsymbol{x}_i, \boldsymbol{x}_j$ in Euclidean space, the kernel function is defined as:

$$\boldsymbol{K}_x(\boldsymbol{x}_i, \boldsymbol{x}_j) = exp(-\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2/2\sigma_x^2) \tag{7}$$

which actually makes use of the Euclidean distance between two data points $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$.

As to the set models, since they are usually defined on Riemannian manifold rather than Euclidean space [32]–[34], the above formulation will fail for them. So, it has to be generalized to Riemannian manifold. For this purpose, given two points $\boldsymbol{Y}_i, \boldsymbol{Y}_j$ on Riemannian manifold, we formally define a generalized kernel function for them as in the following:

$$\boldsymbol{K}_Y(\boldsymbol{Y}_i, \boldsymbol{Y}_j) = exp(-d^2(\boldsymbol{Y}_i, \boldsymbol{Y}_j)/2\sigma_Y^2) \tag{8}$$

It is clear that this kernel actually takes the form of Gaussian function. In the formulation, the most important component is $d(\boldsymbol{Y}_i, \boldsymbol{Y}_j)$, which defines a distance on the Riemannian manifold that the usual set models $\boldsymbol{Y}$ lies. Specifically, for three typical set models, i.e., linear subspace, affine subspace and covariance matrix, this distance is discussed individually in the following.

*1) For Linear Subspaces:* A number of works [32], [33], [43], [51], [64] studied that the space of $d$-dimensional linear subspaces of the $\mathbb{R}^D$ is a Grassmann manifold $\mathcal{G}(d, D)$. In other words, an element of $\mathcal{G}(d, D)$ is a linear subspace, which can be represented by its orthonormal basis matrix $\boldsymbol{U}$. $\boldsymbol{U}$ is formed by the $d$ leading eigenvectors corresponding to the $d$ largest eigenvalues of the covariance matrix of an image set, where each image is represented as a $D$ dimensional feature vector. For two points $\boldsymbol{Y}_i, \boldsymbol{Y}_j$ on the manifold, their distance is usually measured by the Projection metric [32]:

$$d(\boldsymbol{Y}_i, \boldsymbol{Y}_j) = 2^{-1/2}\|\boldsymbol{U}_i\boldsymbol{U}_i^T - \boldsymbol{U}_j\boldsymbol{U}_j^T\|_F. \tag{9}$$

where $\|\cdot\|_F$ denotes the matrix Frobenius norm.

*2) For Affine Subspaces:* An affine subspaces is actually a linear subspace with an offset. Accordingly, the work [33] defined the space of $d$-dimensional affine subspaces as an Affine Grassmann manifold $\mathcal{AG}(d, D)$. Formally, each point on $\mathcal{AG}(d, D)$ is an affine subspace spanned with an orthonormal matrix $\boldsymbol{U}$ adding the offset $\boldsymbol{\mu}$ (i.e., the mean of a set of Euclidean points) from the origin. In this work, we simply extend the similarity proposed in [33] to the distance of two points $\boldsymbol{Y}_i, \boldsymbol{Y}_j$ on the affine manifold as:

$$d(\boldsymbol{Y}_i, \boldsymbol{Y}_j) = 2^{-1/2}(\|\boldsymbol{U}_i\boldsymbol{U}_i^T - \boldsymbol{U}_j\boldsymbol{U}_j^T\|_F + \|(\boldsymbol{I} - \boldsymbol{U}_i\boldsymbol{U}_i^T)\boldsymbol{\mu}_i - (\boldsymbol{I} - \boldsymbol{U}_j\boldsymbol{U}_j^T)\boldsymbol{\mu}_j\|_F), \quad (10)$$

where $\boldsymbol{I} \in \mathbb{R}^{D\times D}$ is the identity matrix.

*3) For Covariance Matrices:* As studied in [34] and [65]–[67], when the covariance matrices are nonsingular (frequently the case), they are Symmetric Positive Definite (SPD) matrices $\boldsymbol{C}$ of size $d \times d$, and thus typically reside on a specific Riemannian manifold $\mathbb{S}_+^d$. To define the distance $d(\boldsymbol{Y}_i, \boldsymbol{Y}_j)$ for Eq.8 on the SPD manifold, we exploit Log-Euclidean distance (LED) [68], which overcomes the computational limitations of the affine-invariant distance (AID) [66], [69] framework while retaining favorable Riemannian geometry. Formally, LED is achieved by classical Euclidean computations in the domain of SPD matrix logarithms:

$$d(\boldsymbol{Y}_i, \boldsymbol{Y}_j) = \|\log(\boldsymbol{C}_i) - \log(\boldsymbol{C}_j)\|_F. \quad (11)$$

where $\log(\boldsymbol{C}) = \boldsymbol{U}\log(\boldsymbol{\Sigma})\boldsymbol{U}^T$ with $\boldsymbol{C} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{U}^T$ being the eigen-decomposition of covariance matrix $\boldsymbol{C}$.

Based on the distances defined in Eq.9, Eq.10 and Eq.11, the corresponding kernel function on the specific type of Riemannian manifold can be yielded by using Eq.8. However, according to Mercer's theorem, only positive definite kernels define valid RKHS. By employing the approach developed in the work [67], we can easily prove the positive definiteness of these Gaussian kernels defined on the corresponding Riemannian manifold. As for details, readers are refered to [67].

Similar to our previous work [35], according to the type of Riemannian manifold where the employed set models reside on, this work can actually deal with three heterogeneous matching cases: Euclidean-to-Grassmannian (EG), Euclidean-to-AffineGrassmannian (EA) and Euclidean-to-SPD (ES) matching cases. Accordingly, we denote the proposed PSCL working in the three different cases respectively by PSCL-EG, PSCL-EA and PSCL-ES.

### C. Discussion

To address the problem of heterogeneous matching, our previous approach [35], called Learning Euclidean-to-Riemannian Metric (LERM), attempts to learn a proper cross-view distance metric across heterogeneous spaces (i.e., Euclidean space and Riemannian manifold). In contrast, our new method PSCL exploits the well-known KCCA to more directly learn correlations between heterogeneous data. Actually, both of these two methods can be viewed as a two-stage procedure, which firstly embeds the heterogeneous spaces into two high dimensional Hilbert spaces, and then learn a couple of transformations from the two Hilbert spaces to a common subspace. Nevertheless, their inputs and objectives are quite different: LERM is designed to learn an optimal distance metric by taking both positive and negative pairs as constraints, while PSCL aims to learn the maximum correlations between positive pairs. Therefore, intuitively, LERM seems to be more sophisticated than PSCL. However, under the classical KCCA framework, PSCL is definitely a good benchmark tool to study the cross-view learning problem of V2S/S2V face recognition.

## V. Experiments

In this section, we evaluate a number of representative set-based face recognition methods listed in Table VII and Table VIII for benchmarking both V2S/S2V and V2V face recognition on our released COX Face DB.

### A. Evaluations on V2S/S2V Face Recognition

In the V2S/S2V face recognition evaluation, either the still face images of high resolution or the smaller size face frames from video sequences are all normalized to the same size grayscale images of $48 \times 60$. In addition, we adopt histogram equalization to eliminate the illumination effects on normalized facial images.

For fair comparison, most of the parameters of the comparative methods are empirically tuned according to the recommendations in the original work as well as the source codes provided by the original authors: For HKNN, the regularization parameter $\lambda$ is set to 50. For PSDML, we set its parameters $\nu = 1$, $\lambda = 0.8$. For our method PSCL and LERM, we implement three Euclidean-to-Riemannian cases, i.e., PSCL-EG/PSCL-EA/PSCL-ES and LERM-EG/LERM-EA/LERM-ES. In LERM, the parameters $\lambda_1 = 0.01, \lambda_2 = 0.1$, the neighborhood number $k_1 = 1, k_2 = 20$ and the number of iterations is set to 30. For PSCL-EG/PSCL-EA/LERM-EG/LERM-EA, the dimension of (affine) Grassmann manifold is set to 10. For PSCL-ES and LERM-ES, in order to deal with possible singularity of covariance matrices, following [34], regularization technique is applied to the original covariance matrix $\boldsymbol{C}$ as $\boldsymbol{C}^* = \boldsymbol{C} + \lambda\boldsymbol{I}$, where $\boldsymbol{I}$ is the identity matrix and $\lambda$ is set to $10^{-3} \times trace(\boldsymbol{C})$. For all the cases in PSCL and LERM, the kernel widths $\sigma_x, \sigma_Y$ are specified from the mean of distances. The final dimensions of the PSCL and LERM are both set to $c - 1$ where $c$ is the number of subjects in training.

The rank-1 recognition rates on V2S and S2V face recognition evaluations, defined in Table IV and Table V, are both reported in Table IX. In this table, each result includes both mean and standard deviation over 10 runs of the evaluation, as detailed in the corresponding evaluation protocol in Section II.D. For the verification evaluation, the verification results are shown by depicting ROC curves in Fig. 4. In this figure, we present the ROC curves of six representative methods, each of which reached the (second) highest rank-1 recognition rates respectively with different set modelings. In addition to these results, we also compare their efficiencies in Table X. This table tabulates the training and testing time of different methods working in the V2S/S2V scenario on an

TABLE IX

RANK-1 RECOGNITION RATES (%) OF **V2S/S2V** FACE RECOGNITION EVALUATIONS SPECIFIED IN TABLE IV AND TABLE V ON COX FACE DB. THE BOLD VALUES INDICATE THE FIRST THREE HIGHEST PERFORMANCES IN THE CORRESPONDING TESTINGS

| Methods | V2S face recognition rate | | | S2V face recognition rate | | |
|---|---|---|---|---|---|---|
| | V1-S | V2-S | V3-S | S-V1 | S-V2 | S-V3 |
| NFS [37] | 9.99 ± 1.17 | 5.90 ± 0.92 | 22.23 ± 1.60 | 11.64 ± 0.81 | 6.51 ± 0.57 | 31.67 ± 1.48 |
| **PSCL-EG** | 32.51 ± 0.97 | 28.87 ± 1.70 | 48.43 ± 1.41 | 30.16 ± 1.00 | 27.34 ± 1.44 | 44.91 ± 1.03 |
| **LERM-EG** [35] | 34.13 ± 1.83 | 31.10 ± 1.23 | 47.87 ± 2.65 | 34.53 ± 1.89 | **32.20 ± 1.85** | 50.26 ± 1.77 |
| PMD [3] | 6.46 ± 0.44 | 4.49 ± 0.51 | 12.36 ± 0.87 | 6.10 ± 0.45 | 4.06 ± 0.50 | 12.83 ± 0.64 |
| HKNN [38] | 4.70 ± 0.33 | 3.70 ± 0.44 | 12.70 ± 1.00 | 6.34 ± 0.50 | 4.64 ± 0.52 | 20.41 ± 0.80 |
| CKNN [38] | 7.93 ± 0.56 | 5.47 ± 0.52 | 14.83 ± 1.06 | 8.89 ± 0.75 | 5.67 ± 0.67 | 26.24 ± 0.82 |
| PSDML [40] | 12.14 ± 1.04 | 9.43 ± 1.41 | 25.43 ± 1.29 | 7.04 ± 0.60 | 4.14 ± 0.52 | 29.86 ± 1.69 |
| **PSCL-EA** | 30.33 ± 1.17 | 28.39 ± 1.30 | 47.74 ± 1.92 | 28.49 ± 1.12 | 26.49 ± 1.01 | 45.21 ± 1.40 |
| **LERM-EA** [35] | **45.03 ± 1.22** | **42.53 ± 1.50** | **59.76 ± 1.40** | **43.17 ± 0.87** | **41.51 ± 1.76** | **60.26 ± 1.68** |
| Mahalanobis [41] | 12.81 ± 0.85 | 10.53 ± 0.79 | 22.06 ± 1.20 | 8.34 ± 0.55 | 6.74 ± 0.42 | 31.40 ± 1.04 |
| **PSCL-ES** | **38.60 ± 1.39** | **33.20± 1.77** | **53.26 ± 0.80** | **36.39 ± 1.61** | 30.87 ± 1.77 | **50.96 ± 1.44** |
| **LERM-ES** [35] | **45.71 ± 2.05** | **42.80 ± 1.86** | **58.37 ± 3.31** | **49.07 ± 1.53** | **44.16 ± 0.94** | **63.83 ± 1.58** |

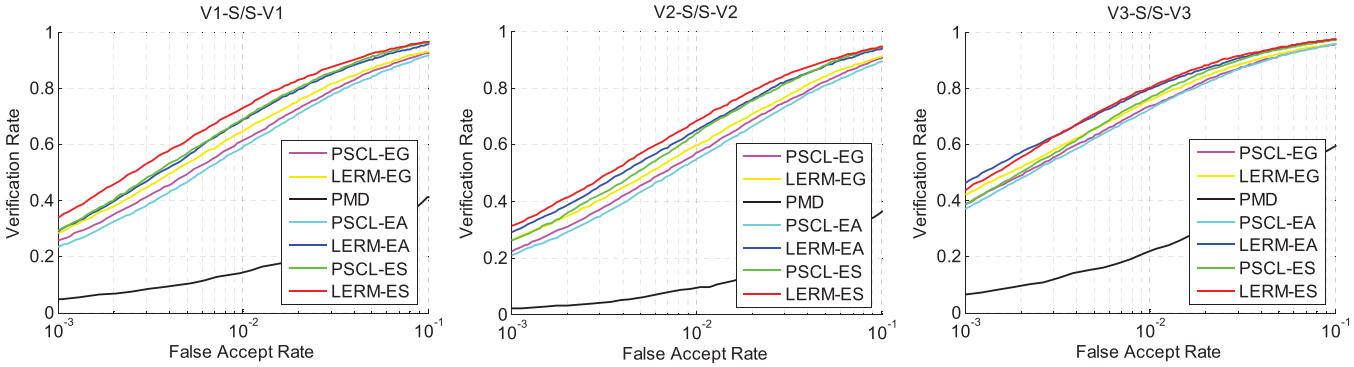



Fig. 4. ROC curve of different methods working in **V2S/S2V** scenario. Here, we show the verification rates (%) of several representative set-based methods from FAR = 0.001 to FAR = 0.1.

TABLE X

RUNNING TIME (IN SECONDS) COMPARISON OF DIFFERENT METHODS IN THE **V2S/S2V** FACE RECOGNITION

| **Methods** | NFS | PMD | HKNN | CKNN | PSDML | Mahalanobis | **PSCL-ES** | **LERM-ES** |
|---|---|---|---|---|---|---|---|---|
| Train | N/A | N/A | N/A | N/A | 12388.50 | N/A | 822.94 | 991.57 |
| Test | 0.31 | 12.47 | 2.47 | 5.42 | 3.27 | 0.19 | 1.07 | 0.92 |

Intel(R) Core(TM) i7-37700M (3.40GHz) PC. In this table, training time is only needed by supervised methods. For testing, we report the classification time for recognizing 1 probe subject from 700 gallery subjects.

From the results in Table IX, we can clearly observe that the rank-1 recognition rates of most of the comparative methods are quite low. Specifically, the highest recognition rates for Exp.V1-S, Exp.V2-S, and Exp.V3-S, are only 45.71%, 42.80%, and 58.37% respectively; and for Exp.S-V1, Exp.S-V2, and Exp.S-V3, they are respectively 49.07%, 44.16%, and 63.83%. The top verification rates for Exp.V1-S/S-V1, Exp.V2-S/S-V2, and Exp.V3-S/S-V3 are 34.30%, 31.12%, 45.97% at false accept rate (FAR)= 0.001, and are 72.63%, 68.37%, 80.46% at FAR= 0.01. It is clear that the experimental results with V1/V2 videos is much worse than those with V3 videos, which can be attributed to the low quality, low resolution, and large pose variations of the faces in V1/V2. Experiments with V3 videos are easier mainly because V3 videos record more frontal face images of higher resolution, as the subjects walked directly to the camera. Overall speaking, these relatively low recognition rates prove that the proposed COX Face DB and corresponding evaluations are very suitable for validating and advancing future V2S/S2V face recognition technologies.

Since relatively few set-based methods can be directly applied to V2S/S2V face recognition, most of comparative methods, except PSDML and our method LERM-EG/LERM-EA/LERM-ES, are unsupervised. As can be seen from Table IX, among the unsupervised methods, NFS, Mahalanobis slightly outperform PMD, HKNN, and CKNN. This may demonstrate that NFS and Mahalanobis are more robust point-to-set distances than others for S2V/V2V face recognition on this dataset. Another observation is that the performances of all these unsupervised methods are very low due to the challenging setting of this dataset and their failure to consider the large heterogeneity between still images

TABLE XI

RANK-1 RECOGNITION RATES (%) OF **V2V** FACE RECOGNITION EVALUATIONS SPECIFIED IN TABLE VI ON COX FACE DB. THE BOLD VALUES INDICATE THE FIRST THREE HIGHEST PERFORMANCES IN THE CORRESPONDING TESTINGS

| Methods | V2V face recognition rate | | | | | |
|---|---|---|---|---|---|---|
| | V2-V1 | V3-V1 | V3-V2 | V1-V2 | V1-V3 | V2-V3 |
| MSM [30] | 45.53 ± 0.46 | 21.47 ± 1.87 | 11.00 ± 0.85 | 39.83 ± 0.67 | 19.36 ± 0.67 | 9.50 ± 0.67 |
| DCC [1] | 62.53 ± 6.13 | 66.10 ± 3.52 | 50.56 ± 4.21 | 56.09 ± 11.27 | 53.84 ± 11.37 | 45.19 ± 9.81 |
| GDA [32] | 68.61 ± 1.96 | **77.70 ± 1.58** | 71.59 ± 1.11 | **65.93 ± 1.88** | 76.11 ± 0.98 | 74.83 ± 1.80 |
| GGDA [43] | **70.80 ± 1.24** | 76.23 ± 1.25 | **71.99 ± 1.05** | **69.17 ± 1.01** | **76.77 ± 1.57** | **77.43 ± 1.41** |
| **PSCL-GG** | 55.16 ± 1.61 | 66.71 ± 1.27 | 64.53 ± 1.75 | 58.16 ± 1.83 | 74.80 ± 1.04 | 68.51 ± 1.72 |
| **LERM-GG** [35] | 60.90 ± 2.06 | 73.61 ± 1.54 | 69.56 ± 0.99 | 61.73 ± 1.52 | 75.80 ± 1.67 | 70.49 ± 1.89 |
| MMD [3] | 38.29 ± 0.23 | 30.34 ± 1.43 | 15.24 ± 1.28 | 34.86 ± 0.47 | 22.21 ± 0.63 | 11.44 ± 0.60 |
| MDA [44] | 65.83 ± 0.92 | 62.96 ± 2.12 | 36.20 ± 0.95 | 55.53 ± 1.41 | 43.24 ± 1.60 | 29.94 ± 1.56 |
| AHISD [4] | 53.03 ± 2.05 | 36.13 ± 1.00 | 17.50 ± 0.81 | 43.51 ± 0.66 | 34.99 ± 0.83 | 18.80 ± 0.69 |
| CHISD [4] | 56.90 ± 0.64 | 30.13 ± 0.79 | 15.03 ± 0.77 | 44.36 ± 0.51 | 26.40 ± 0.72 | 13.69 ± 0.76 |
| RNP [47] | 58.07 ± 0.97 | 37.89 ± 0.83 | 14.56 ± 0.79 | 52.53 ± 1.71 | 33.34 ± 1.03 | 14.80 ± 0.80 |
| SSDML [40] | 60.13 ± 0.23 | 53.14 ± 0.72 | 28.73 ± 0.53 | 47.91 ± 0.38 | 44.42 ± 0.74 | 27.34 ± 0.85 |
| ISCRC [48] | **69.74 ± 0.69** | 60.71 ± 1.53 | 37.66 ± 1.14 | 61.09 ± 4.05 | 64.96 ± 6.34 | 37.71 ± 4.21 |
| **PSCL-AA** | 51.91 ± 2.20 | 64.27 ± 2.10 | 61.96 ± 1.54 | 55.53 ± 1.81 | 72.39 ± 1.58 | 65.94 ± 1.86 |
| **LERM-AA** [35] | 57.17 ± 1.86 | 72.13 ± 1.51 | 68.36 ± 1.49 | 57.74 ± 1.66 | 74.60 ± 1.25 | 69.04 ± 1.85 |
| LMKML [52] | 56.14 ± 1.78 | 44.26 ± 3.54 | 33.14 ± 4.63 | 55.37 ± 3.06 | 39.83 ± 3.35 | 29.54 ± 3.94 |
| CDL [34] | **78.43 ± 1.01** | **85.31 ± 0.97** | **79.71 ± 1.47** | **75.56 ± 1.95** | **85.84 ± 0.86** | **81.87 ± 1.14** |
| **PSCL-SS** | 57.70 ± 1.40 | 73.17 ± 1.44 | 67.70 ± 1.70 | 62.77 ± 1.02 | 78.26 ± 0.97 | 68.91 ± 2.28 |
| **LERM-SS** [35] | 65.94 ± 1.97 | **78.24 ± 1.32** | **70.67 ± 1.88** | 64.44 ± 1.55 | **80.53 ± 1.36** | **72.96 ± 1.99** |

(i.e., points) and video sequences (i.e., sets) in the V2S/S2V scenarios. As a supervised method, PSDML outperforms HKNN with the same set modeling (i.e., affine hull modeling) in most tests but does not beat other unsupervised methods with different set models. In contrast, our PSCL and LERM methods with different set modelings achieve much higher performances than other methods including PSDML. This is possibly because PSDML fails to respect the non-Euclidean geometry of the space of set models. As is well known, neglecting the underlying data structure in the learning stage will probably lead to undesirable metrics. So, by exploiting the Riemannian geometry of the manifold of set models to learn the correlation of heterogeneous data, our LERM always significantly outperforms the competitive methods in the V2S/S2V benchmark tests.

### *B. Evaluations on V2V Face Recognition*

In the V2V face recognition testing, each video face frames is normalized to a grayscale image with a size of $32 \times 40$. Similar to the V2S/S2V evaluation, histogram equalization is employed to eliminate the lighting effects on the normalized facial images.

According to the original works of the competing methods, we carefully tune their parameters for fair comparison. For MSM/DCC/MMD/AHISD, PCA is performed to learn the linear subspaces by preserving 95% of data energy. For GDA/GGDA, we set the dimension of (affine) Grassmann manifold as 10. For GGDA, the number of neighborhood is set to 10. In MMD/MDA, the number of between-class Nearest Neighbor local models and the subspace dimension are specified as [3]/[44]. For both AHISD and CHISD, we use their linear versions with the default setting. For SSDML, we set the parameters $\lambda_1 = 0.001, \lambda_2 = 0.5$, the numbers of positive and negative pairs per set is set to 10 and 20. For RNP, we set the parameters $\lambda_1 = 0.001$ and $\lambda_2 = 0.1$. For ISCRC, we implement the RH-ISCRC version with l2 regularization due to its relatively high efficiency. In this method, we set $\lambda_1 = 0.001, \lambda_2 = 0.001$. For CDL, since the kernel Partial Least Squares (KPLS)-based CDL works only when the gallery data is used for training, it is ineligible for our evaluation protocol on COX Face DB. So, we only implement the KLDA-based CDL. To yield non-singular covariance matrices in CDL, we adopt the same regularization technique on covariance matrices mentioned before. The final dimension of GDA/GGDA/CDL are all set to $c-1$ where $c$ is the number of classes in training. For LMKML, we use median distance heuristic to tune the widths of Gaussian kernels.

Note that, our method PSCL and LERM are both designed for cross-view based classification. Since there are three different sets of videos for training and testing as designed in Table VI, we also evaluate the performances of PSCL and LERM. According to different set modeling (i.e., linear subspace, affine subspace and covariance matrix modeling), we denote their corresponding instances as PSCL-GG/PSCL-AA/PSCL-SS and LERM-GG/LERM-AA/LERM-SS, where GG, AA, SS respectively correspond to different manifold-to-manifold matching cases: Grassmann to Grassmann (GG), Affine Grassmann to Affine Grassmann (AA), SPD to SPD (SS). Their parameter settings are the same as those working in the V2S/S2V scenario.

As defined in Table VI in Section II.D, we report the rank-1 identification rates and verification rates in the V2V face recognition scenario. As shown in Table XI, following the protocol designed in section II.D, the identification results are consist of both mean and standard deviation over 10 random runs. The verification results of several representative methods are represented by the ROC curves depicted in Fig. 5. To study the efficiencies of the comparative methods, Table XII lists

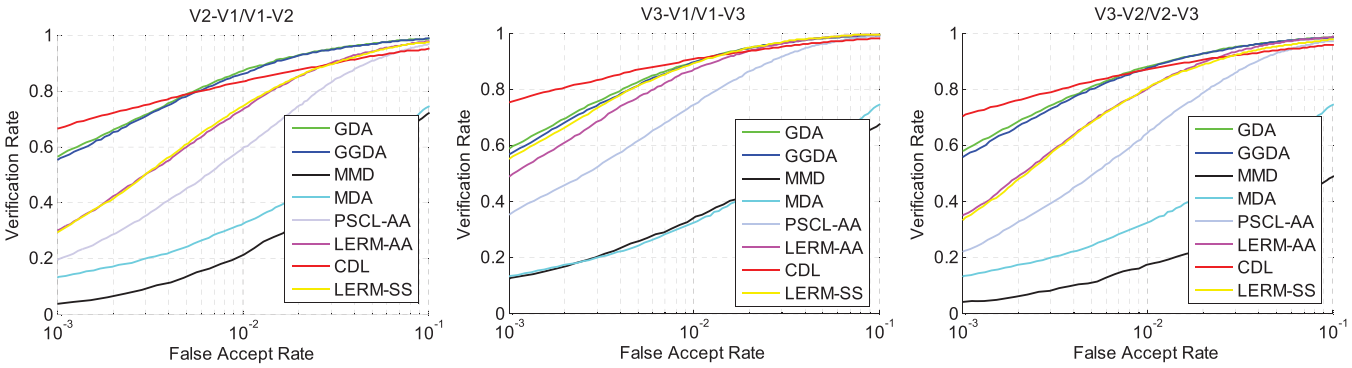



Fig. 5. ROC curve of different methods working in **V2V** scenario. Here, we show the verification rates (%) of several representative set-based methods from FAR = 0.001 to FAR = 0.1.

TABLE XII

RUNNING TIME (IN SECONDS) COMPARISON OF DIFFERENT METHODS IN THE **V2V** FACE RECOGNITION

| **Methods** | MSM | DCC | GDA | GGDA | MMD | MDA | AHISD | CHISD | RNP | SSDML | ISCRC | LMKML | CDL | **PSCL-SS** | **LERM-SS** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Train | N/A | 189.79 | 118.06 | 168.67 | N/A | 107.07 | N/A | N/A | 17.37 | 7012.80 | 151.03 | 92968.91 | 431.79 | 96.35 | 102.74 |
| Test | 0.75 | 8.79 | 0.29 | 0.36 | 18.03 | 14.11 | 8.85 | 27.46 | 3.13 | 3.23 | 44.42 | 1.07 | 5.77 | 0.19 | 0.22 |

their the running time on an Intel(R) Core(TM) i7-37700M (3.40GHz) PC.

From Table XI, the first clear impression is that the V2V recognition rates are much higher than those of S2V/V2S (as reported in Table IX). In addition, the top verification rates in this scenario are also higher: for Exp.V2-V1/V1-V2, Exp.V3-V1/V1-V3, and Exp.V3-V2/V2-V3 are 66.67%, 75.69%, 70.19% at false accept rate (FAR)= 0.001, and are 87.26%, 90.77%, 87.14% at FAR= 0.01. Two reasons are enumerated to explain this phenomenon: 1) In the V2V face recognition scenario, both the query and the target are videos containing multiple frames, thus providing more discriminative information for classification. 2) For the V2V scenario, the query-target matching performs in homogenous space, which is easier than matching between heterogeneous spaces (i.e., the space of still images and the space of usual set models for video sequences).

The second observation is that, among all the non-discriminant methods, nonlinear manifold based methods (i.e. MMD) outperforms linear subspace based method (i.e. MSM) in 4 tests out of all the 6 ones. However, both of the two methods are beaten by affine/convex hull based methods (i.e., AHISD and CHISD). We mildly attribute the superiority of the affine/convex hull based methods to their generating virtual samples from the relatively sparse original samples in the high-dimensional image space.

Thirdly, it is also clear that the discriminant methods (e,g., DCC, GDA, GGDA, MDA, SSDML, CDL, PSCL, LERM) consistently outperform those non-discriminative methods (e.g., MSM, MMD, AHISD, CHISD). As analyzed previously, it is not surprising that discriminant methods have learned better models for the purpose of classification from the training data with distribution probabilistically similar to that of the testing data.

Fourthly, among the discriminant methods, the kernel-based discriminant methods (e.g., GDA, GGDA, CDL, and LERM-SS) generally achieve better performances than others in the V2V face recognition scenarios. The reason is that the kernel-based discriminant methods usually treat the set models as elements on a specific Riemannian manifold and perform feature extraction and classification in the same space. In contrast, as presented in the existing work [32], the other discriminant methods typically extract features and classify subjects in two different spaces, and thus make the latter benefit less from the former.

Lastly, we also find that, different from the results in the V2S/S2V evaluation, both our methods LERM and PSCL can not reach the best performances in the V2V face recognition scenario. This is because LERM and PSCL is originally designed for the cross-view problem of V2S/S2V face recognition. In order to adapt them to the V2V scenario, they are exploited to learn cross-view models on two different video sources by assuming the videos captured by different camcorders belongs to different views. With this strong assumption to learn the discriminant information on training data, it is reasonable that LERM and PSCL can not achieve the-state-of-the-art in the V2V evaluation.

## VI. CONCLUSION

This paper released a large scale face database with 1,000 subjects, titled COX Face DB, which contains both still images taken by DC with seated subjects and surveillance-like videos captured by camcorders with walking subjects. The dataset and its evaluation protocol are designed for evaluating all the three video-based face recognition scenarios: V2S, S2V and V2V. Evaluation results show that, on our COX Face DB, the best identification rates for V2S/S2V scenarios are around 40%∼60%, while those for V2V scenario are around 75%∼ 85%, which indicates that the new dataset is appropriate as a new benchmark dataset for video-based face recognition.

As a comparative study paper, we also reviewed existing set-based methods recently proposed for video-based face recognition in terms of their representation for face image set and their point-to-set or set-to-set metrics. Most of these

methods are evaluated based on the COX Face DB with the accompanying evaluation protocols.

To advance the relatively under-studied V2S/S2V tasks, we additionally proposed a novel baseline method named Point-to-Set Correlation Learning (PSCL). In consideration of the heterogeneity between still image (i.e., Euclidean point) and video sequence (i.e., Riemannian elements), PSCL exploits the well-know KCCA to learn the maximum correlations between these heterogeneous data. On COX Face DB, our PSCL impressively outperforms most of the existing methods for V2S/S2V face recognition tasks. In addition, this method can be also adapted to the V2V task and finally achieves promising performances for this task.

Overall speaking, the studies in this paper show that, video-based face recognition is far from mature especially compared with face recognition from still images. We suggest more efforts should be made to advance the real-world video-based face recognition applications.

## ACKNOWLEDGMENT

The authors would like to thank Prof. TurgLm IBRAHIM, Dr. Luhong Liang and Mr. Yan Li for giving us great helps in collecting and processing COX face database.

## REFERENCES

[1] T.-K. Kim, J. Kittler, and R. Cipolla, "Discriminative learning and recognition of image set classes using canonical correlations," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 6, pp. 1005–1018, Jun. 2007.

[2] M. Kim, S. Kumar, V. Pavlovic, and H. Rowley, "Face tracking and recognition with visual constraints in real-world videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1787–1794.

[3] R. Wang, S. Shan, X. Chen, and W. Gao, "Manifold-manifold distance with application to face recognition based on image set," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 2940–2947.

[4] H. Cevikalp and B. Triggs, "Face recognition based on image sets," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 2567–2573.

[5] S. Chen, C. Sanderson, M. T. Harandi, and B. C. Lovell, "Improved image set classification via joint sparse approximated nearest subspaces," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 452–459.

[6] S. Zhou, V. Krueger, and R. Chellappa, "Probabilistic recognition of human faces from video," *Comput. Vis. Image Understand.*, vol. 91, nos. 1–2, pp. 214–245, Jul. 2003.

[7] S. Biswas, G. Aggarwal, and P. J. Flynn, "Pose-robust recognition of low-resolution face images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 601–608.

[8] Z. Huang, S. Shan, H. Zhang, H. Lao, A. Kuerban, and X. Chen, "Benchmarking still-to-video face recognition via partial and local linear discriminant analysis on COX-S2V dataset," in *Proc. Asian Conf. Comput. Vis.*, 2012, pp. 589–600.

[9] Z. Huang, X. Zhao, S. Shan, R. Wang, and X. Chen, "Coupling alignments with recognition for still-to-video face recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 3296–3303.

[10] J. H. Barr, K. W. Bowyer, P. J. Flynn, and S. Biswas, "Face recognition from video: A review," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 26, no. 5, p. 1266002, 2012.

[11] M. Grgic, K. Delac, and S. Grgic, "SCface—Surveillance cameras face database," *Multimedia Tools Appl.*, vol. 51, no. 3, pp. 863–879, Feb. 2011.

[12] A. J. O'Toole, P. Phillips, S. Weimer, D. Roark, J. Ayyad, R. Barwick, and J. Dunlop, "Recognizing people from dynamic and static faces and bodies: Dissecting identity with a fusion approach," *Vis. Res.*, vol. 51, no. 1, pp. 74–83, Jan. 2011.

[13] L. Wolf, T. Hassner, and I. Maoz, "Face recognition in unconstrained videos with matched background similarity," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 529–534.

[14] J. R. Beveridge, P. Phillips, D. Bolme, B. A. Draper, G. Givens, Y. Lui, M. N. Teli, H. Zhang, W. T. Scruggs, K. W. Bowyer, P. J. Flynn, and S. Cheng, "The challenge of face recognition from digital point-and-shoot cameras," in *Proc. IEEE 6th Int. Conf. Biometrics, Theory, Appl. Syst.*, Sep./Oct. 2013, pp. 1–8.

[15] R. Gross and J. Shi, "The CMU motion of body (MoBo) database," Tech. Rep. CMU-RI-TR-01-18, Robot. Inst., Carnegie Mellon Univ., Pittsburgh, PA, USA, 2001.

[16] K.-C. Lee, J. Ho, M.-H. Yang, and D. Kriegman, "Video-based face recognition using probabilistic appearance manifolds," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2003, pp. 313–320.

[17] K.-C. Lee, J. Ho, M.-H. Yang, and D. Kriegman, "Visual tracking and recognition using probabilistic appearance manifolds," *Comput. Vis. Image Understand.*, vol. 99, no. 3, pp. 303–331, Sep. 2005.

[18] R. Goh, L. Liu, X. Liu, and T. Chen, "The CMU face in action (FIA) database," in *Proc. Int. Conf. Autom. Face Gesture Recognit.*, 2005, pp. 255–263.

[19] *University of Cambridge Face Database*. [Online]. Available: http://mi.eng.cam.ac.uk/~oa214/academic/, accessed Oct. 24, 2015.

[20] *Faces96 Database*. [Online]. Available: http://cswww.essex.ac.uk/mv/allfaces/faces96.html, accessed Oct. 24, 2015.

[21] C. Sanderson. (2008). *Biometric Person Recognition: Face, Speech and Fusion*. [Online]. Available: http://www.itee.uq.edu.au/~onrad/vidtimit/

[22] J. Phillips, "Video challenge problem multiple biometric grand challenge: Preliminary results of version 2," *Nat. Inst. Standards Technol.*, 2009. [Online]. Available: http://biometrics.nist.gov/cs_links/face/mbgc/FINAL_VIDEO_MBGC_Workshop2_v1.pdf., accessed Oct. 24, 2015.

[23] J. R. Barr, K. W. Bowyer, and P. J. Flynn, "Detecting questionable observers using face track clustering," in *Proc. IEEE Workshop Appl. Comput. Vis.*, Jan. 2011, pp. 182–189.

[24] Y. Wong, S. Chen, S. Mau, C. Sanderson, and B. C. Lovell, "Patch-based probabilistic image quality assessment for face selection and improved video-based face recognition," in *Proc. IEEE Conf. Workshop Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 74–81.

[25] R. Chellappa, J. Ni, and V. M. Patel, "Remote identification of faces: Problems, prospects, and progress," *Pattern Recognit. Lett.*, vol. 33, no. 14, pp. 1849–1859, Oct. 2012.

[26] L. Liu, L. Zhang, H. Liu, and S. Yan, "Toward large-population face identification in unconstrained videos," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 11, pp. 1874–1884, Nov. 2014.

[27] X. Liu and T. Chen, "Video-based face recognition using adaptive hidden Markov models," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2003, pp. 340–345.

[28] S. K. Zhou, R. Chellappa, and B. Moghaddam, "Visual tracking and recognition using appearance-adaptive models in particle filters," *IEEE Trans. Image Process.*, vol. 13, no. 11, pp. 1491–1506, Nov. 2004.

[29] N. Ye and T. Sim, "Towards general motion-based face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 2598–2605.

[30] O. Yamaguchi, K. Fukui, and K. Maeda, "Face recognition using temporal image sequence," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit.*, Apr. 1998, pp. 318–323.

[31] M. Hayat, M. Bennamoun, and S. An, "Learning non-linear reconstruction models for image set classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1915–1922.

[32] J. Hamm and D. D. Lee, "Grassmann discriminant analysis: A unifying view on subspace-based learning," in *Proc. Int. Conf. Mach. Learn.*, 2008, pp. 376–383.

[33] J. Hamm and D. D. Lee, "Extended Grassmann kernels for subspace-based learning," in *Proc. Adv. Neural Inf. Process.*, 2009, pp. 601–608.

[34] R. Wang, H. Guo, L. S. Davis, and Q. Dai, "Covariance discriminative learning: A natural and efficient approach to image set classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2496–2503.

[35] Z. Huang, R. Wang, S. Shan, and X. Chen, "Learning Euclidean-to-Riemannian metric for point-to-set classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1677–1684.

[36] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical correlation analysis: An overview with application to learning methods," Dept. Comput. Sci., Roy. Holloway, Univ. London, London, U.K., Tech. Rep. CSD-TR-03-02, 2003.

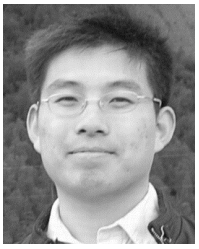
[37] J.-T. Chien and C.-C. Wu, "Discriminant waveletfaces and nearest feature classifiers for face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 12, pp. 1644–1649, Dec. 2002.

[38] P. Vincent and Y. Bengio, "K-local hyperplane and convex distance nearest neighbor algorithms," in *Proc. Adv. Neural Inf. Process. Syst.*, 2001, pp. 985–992.
[39] H. Cevikalp, B. Triggs, and R. Polikar, "Nearest hyperdisk methods for high-dimensional classification," in *Proc. Int. Conf. Mach. Learn.*, 2008, doi: 10.1145/1390156.1390172.
[40] P. Zhu, L. Zhang, W. Zuo, and D. Zhang, "From point to set: Extend the learning of distance metrics," in *Proc. Int. Conf. Comput. Vis.*, Dec. 2013, pp. 2664–2671.
[41] P. C. Mahalanobis, "On the generalized distance in statistics," in *Proc. Nat. Inst. Sci. India*, vol. 2, no. 1, pp. 49–55, 1936.
[42] K. Fukui and O. Yamaguchi, "Face recognition using multi-viewpoint patterns for robot vision," in *Proc. Int. Symp. Robot. Res.*, 2003, pp. 192–201.
[43] M. T. Harandi, C. Sanderson, S. Shirazi, and B. C. Lovell, "Graph embedding discriminant analysis on Grassmannian manifolds for improved image set matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 2705–2712.
[44] R. Wang and X. Chen, "Manifold discriminant analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 429–436.
[45] Z. Cui, S. Shan, H. Zhang, S. Lao, and X. Chen, "Image sets alignment for video-based face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2012, pp. 2626–2633.
[46] Y. Hu, A. S. Mian, and R. Owens, "Sparse approximated nearest points for image set classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 121–128.
[47] M. Yang, P. Zhu, L. Van Gool, and L. Zhang, "Face recognition based on regularized nearest points between image sets," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit.*, Apr. 2013, pp. 1–7.
[48] P. Zhu, W. Zuo, L. Zhang, S. C.-K. Shiu, and D. Zhang, "Image set-based collaborative representation for face recognition," *IEEE Trans. Inf. Forensics Security*, vol. 9, no. 7, pp. 1120–1132, Jul. 2014.
[49] G. Shakhnarovich, J. W. Fisher, and T. Darrell, "Face recognition from long-term observations," in *Proc. Eur. Conf. Comput. Vis.*, 2002, pp. 851–865.
[50] O. Arandjelović, G. Shakhnarovich, J. Fisher, R. Cipolla, and T. Darrell, "Face recognition with image sets using manifold density divergence," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2005, pp. 581–588.
[51] R. Vemulapalli, J. K. Pillai, and R. Chellappa, "Kernel learning for extrinsic classification of manifold features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 1782–1789.
[52] J. Lu, G. Wang, and P. Moulin, "Image set classification using holistic multiple order statistics features and localized multi-kernel metric learning," in *Proc. IEEE Conf. Int. Conf. Comput. Vis.*, Dec. 2013, pp. 329–336.
[53] B. K. Gunturk, A. U. Batur, Y. Altunbasak, M. H. Hayes, and R. M. Mersereau, "Eigenface-domain super-resolution for face recognition," *IEEE Trans. Image Process.*, vol. 12, no. 5, pp. 597–606, May 2003.
[54] O. Arandjelović and R. Cipolla, "A manifold approach to face recognition from low quality video across illumination and pose using implicit super-resolution," in *Proc. Int. Conf. Comput. Vis.*, 2007, pp. 1–8.
[55] U. Park and A. K. Jain, "3D model-based face recognition in video," in *Proc. Int. Conf. Biometrics*, 2007, pp. 1085–1094.
[56] D. Thomas, K. W. Bowyer, and P. J. Flynn, "Multi-factor approach to improving recognition performance in surveillance-quality video," in *Proc. IEEE Conf. Biometrics, Theory, Appl. Syst.*, Sep./Oct. 2008, pp. 1–7.
[57] X. Liu and T. Chen, "Face mosaicing for pose robust video-based recognition," in *Proc. Asian Conf. Comput. Vis.*, 2007, pp. 662–671.
[58] Y. Zhang and A. M. Martínez, "A weighted probabilistic approach to face recognition from multiple images and video sequences," *Image Vis. Comput.*, vol. 24, no. 6, pp. 626–638, Jun. 2006.
[59] J. Stallkamp, H. K. Ekenel, and R. Stiefelhagen, "Video-based face recognition on real-world data," in *Proc. Int. Conf. Comput. Vis.*, Oct. 2007, pp. 1–8.
[60] M. A. Turk and A. P. Pentland, "Face recognition using eigenfaces," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 1991, pp. 586–591.
[61] B. Moghaddam and A. Pentland, "Probabilistic visual learning for object representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 696–710, Jul. 1997.
[62] X. Wang and X. Tang, "Dual-space linear discriminant analysis for face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun./Jul. 2004, pp. 564–569.
[63] H. Hotelling, "Relations between two sets of variates," *Biometrika*, vol. 28, nos. 3–4, pp. 312–377, 1936.
[64] M. T. Harandi, M. Salzmann, S. Jayasumana, R. Hartley, and H. Li, "Expanding the family of Grassmannian kernels: An embedding perspective," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 408–423.
[65] V. Arsigny, P. Fillard, X. Pennec, and N. Ayache, "Fast and simple computations on tensors with Log–Euclidean metrics," French Inst. Res. Comput. Sci. Autom., Rocquencourt, France, Tech. Rep. RR-5584, 2005.
[66] X. Pennec, P. Fillard, and N. Ayache, "A Riemannian framework for tensor computing," *Int. J. Comput. Vis.*, vol. 66, no. 1, pp. 41–66, Jan. 2006.
[67] S. Jayasumana, R. Hartley, M. Salzmann, H. Li, and M. T. Harandi, "Kernel methods on the Riemannian manifold of symmetric positive definite matrices," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 73–80.
[68] V. Arsigny, P. Fillard, X. Pennec, and N. Ayache, "Geometric means in a novel vector space structure on symmetric positive-definite matrices," *SIAM J. Matrix Anal. Appl.*, vol. 29, no. 1, pp. 328–347, 2007.
[69] W. Förstner and B. Moonen. (2003). A metric for covariance matrices. Geodesy-The Challenge of the 3rd Millennium, pp. 299–309. [Online]. Available: http://link.springer.com/chapter/10.1007/978-3-662-05296-9_31

**Zhiwu Huang** (S'13) received the B.S. degree in computer science and technology from Huaqiao University, Quanzhou, China, in 2007, the M.S. degree in computer software and theory from Xiamen University, Xiamen, China, in 2010, and the Ph.D. degree in computer science and technology from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, in 2015.

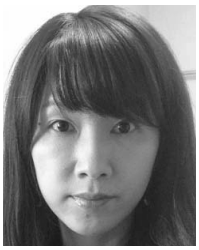
He has been a Post-Doctoral Researcher with the Computer Vision Laboratory, Swiss Federal Institute of Technology, Zurich, Switzerland, since 2015. His current research interests include Riemannian geometry, kernel learning, metric learning and feature learning with application to face recognition, face retrieval, facial expression recognition, gesture recognition, action recognition, and image set classification in computer vision.

**Shiguang Shan** (M'04–SM'15) received the M.S. degree in computer science from the Harbin Institute of Technology, Harbin, China, in 1999, and the Ph.D. degree in computer science from the Institute of Computing Technology (ICT), Chinese Academy of Sciences (CAS), Beijing, China, in 2004. He joined ICT, CAS, in 2002, where he has been a Professor since 2010. He is currently the Deputy Director of the Key Laboratory of Intelligent Information Processing with CAS. He has authored over 200 papers in refereed journals and proceedings in the areas of computer vision and pattern recognition. His research interests cover computer vision, pattern recognition, and machine learning. He especially focuses on face recognition related research topics. He received the China's State Scientific and Technological Progress Awards in 2005 for his work on face recognition technologies. He has served as the Area Chair for many international conferences, including ICCV'11, ICPR'12, ACCV'12, FG'13, ICPR'14, and ICASSP'14. He served as the Workshop Co-Chair of ACCV14 and the Website Co-Chair of ICCV15. He is an Associate Editor of the IEEE TRANSACTIONS ON IMAGE PROCESSING, *Neurocomputing*, and the *EURASIP Journal of Image and Video Processing*.



**Ruiping Wang** (S'08–M'11) received the B.S. degree in applied mathematics from Beijing Jiaotong University, Beijing, China, in 2003, and the Ph.D. degree in computer science from the Institute of Computing Technology, Chinese Academy of Sciences (CAS), Beijing, in 2010.

He was a Post-Doctoral Researcher with the Department of Automation, Tsinghua University, Beijing, from 2010 to 2012. He spent one year working as a Research Associate with the Computer Vision Laboratory, Institute for Advanced Computer Studies (UMIACS), University of Maryland, College Park, from 2010 to 2011. He has been with the Institute of Computing Technology, CAS, since 2012, where he is currently an Associate Professor. His research interests include computer vision, pattern recognition, and machine learning.



**Haihong Zhang** (M'01) received the B.E. degree from Shanghai Jiao Tong University, Shanghai, China, in 1997, and the M.E. and Ph.D. degrees from Osaka City University, Osaka, Japan, in 2001 and 2004, respectively. She is currently a Leader Research Engineer with OMRON Social Solutions Company, Ltd., Tokyo, Japan. Her current research interests include computer vision, machine learning, and image processing.

**Shihong Lao** (M'05) received the B.S. degree in electrical engineering from Zhejiang University, Hangzhou, China, in 1984, and the M.S. degree in electrical engineering from Kyoto University, Kyoto, Japan, in 1988.

He has been with OMRON Social Solutions Company, Ltd., Kyoto, Japan, since 1992, where he is currently an Advisory Technology Specialist with the Core Technology Center. His current interests include facial image processing, visual surveillance, and robot vision.

**Alifu Kuerban** was born in Korla, Xinjiang, China, in 1967. He received the master's degree in software engineering from the Institute of Software Engineering, University of Electronic Science and Technology of China, Chengdu, China, in 2011. From 1993 to 2012, he was a Lecturer with Xinjiang University, where he has been a Professor since 2012. His research areas are ethnic minority language processing and database application.

He has authored over 20 articles. In 2008, he won the second prize of the Science and Technology Progress, China, and the Excellent Teacher Award in Xinjiang Uyghur autonomous region, China, in 2009.

**Xilin Chen** (M'00–SM'09) received the B.S., M.S., and Ph.D. degrees from the Harbin Institute of Technology, Harbin, China, in 1988, 1991, and 1994, respectively, all in computer science. He was a Professor with the Harbin Institute of Technology from 1999 to 2005. He was a Visiting Scholar with Carnegie Mellon University, Pittsburgh, PA, from 2001 to 2004. He has been a Professor with the Institute of Computing Technology, Chinese Academy of Sciences (CAS), since 2004. He is currently the Director of the Key Laboratory of Intelligent Information Processing at CAS. He has authored one book and over 200 papers in refereed journals and proceedings in the areas of computer vision, pattern recognition, image processing, and multimodal interfaces. He is a fellow of the China Computer Federation. He served as an Organizing Committee/Program Committee Member of more than 50 conferences. He was a recipient of several awards, including the China's State Scientific and Technological Progress Award in 2000, 2003, 2005, and 2012 for his research work. He is a Leading Editor of the *Journal of Computer Science and Technology*, and an Associate Editor-in-Chief of the *Chinese Journal of Computers*.