Eigen-PEP for Video Face Recognition

Haoxiang Li[†], Gang Hua[†], Xiaohui Shen[‡], Zhe Lin[‡], Jonathan Brandt[‡]

[†]Stevens Institute of Technology   [‡]Adobe Systems Inc.

Abstract. To effectively solve the problem of large scale video face recognition, we argue for a comprehensive, compact, and yet flexible representation of a face subject. It shall comprehensively integrate the visual information from all relevant video frames of the subject in a compact form. It shall also be flexible to be incrementally updated, incorporating new or retiring obsolete observations. In search for such a representation, we present the Eigen-PEP that is built upon the recent success of the probabilistic elastic part (PEP) model. It first integrates the information from relevant video sources by a part-based average pooling through the PEP model, which produces an intermediate high dimensional, part-based, and pose-invariant representation. We then compress the intermediate representation through principal component analysis, and only a number of principal eigen dimensions are kept (as small as 100). We evaluate the Eigen-PEP representation both for video-based face verification and identification on the YouTube Faces Dataset and a new Celebrity-1000 video face dataset, respectively. On YouTube Faces, we further improve the state-of-the-art recognition accuracy. On Celebrity-1000, we lead the competing baselines by a significant margin while offering a scalable solution that is linear with respect to the number of subjects.

1 Introduction

With the proliferation of videos accumulated in online social multimedia, *e.g.*, hundreds of hours videos are uploaded to YouTube every minute, the problem of video face recognition in the wild has caught more and more attention in recent years. Compared with image-based face recognition, face recognition from videos not only presents new challenges, but also offers new opportunities. As shown in Figure 1, faces in the videos are generally in lower quality, present more pose variations, and often suffer from motion blur. These factors can induce more visual variations of the faces and negatively influence the recognition accuracy. On the other hand, a video clip of a face usually contains hundreds of frames which present varied appearance of the same subject. This obviously offers additional opportunities to better model the visual variations for more robust face recognition by integrating the information from all the frames.

A naive approach to video face recognition would be applying existing image based face recognition algorithms [1], such as those top performers on the Labeled Faces in the Wild (LFW) [2–7], to conduct frame-to-frame matching and then fusing the matching results across all the frame pairs together when comparing

(a) LFW   (b) YouTube Faces   (c) Celebrity-1000

Fig. 1. Sample images in three unconstrained face recognition datasets: the image-based Labeled Faces in the Wild (LFW), video-based YouTube Faces Database, and video-based Celebrity-1000 dataset.
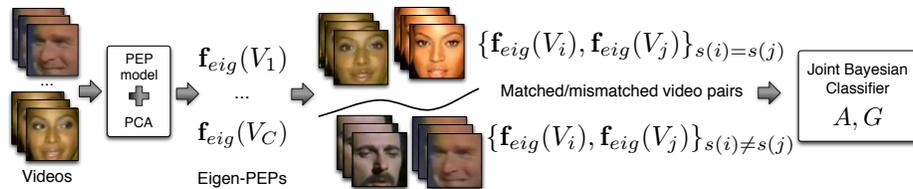
Fig. 2. The high-level training work-flow of our method.

two video faces. This is obviously not a scalable solution as the complexity of a single match is already $O(n^2)$ with respect to the number of frames n each video possesses.

Previous work on the video face recognition includes methods representing the video data by linear combination of the training data [8, 9], utilizing probabilistic methods to exploit the intrinsic manifolds [10–12], etc. We refer readers to Zhao *et al.* [13] for a more comprehensive survey of earlier literatures. Notwithstanding the demonstrated efficacy of these methods, the computational expense is a hurdle when applied to large-scale video face recognition.

We argue that, to effectively solve the problem of large scale video face recognition, we need a comprehensive, compact, and yet flexible representation of a face subject. By comprehensive, we mean that it shall integrate the visual information from all relevant video frames (even those from multiple videos) of a subject to better model the visual variations. By compact, we mean that it is scalable both in terms of computing and storage. By flexible, we mean that it can be incrementally updated, either incorporating new observations, or retiring obsolete observations, without the need to revisit all the video frames used to build the original representation.

To address these requirements, we propose a new video face representation named Eigen-PEP for video face recognition in the wild. The Eigen-PEP representation is built upon the recent success of the probabilistic elastic part (PEP) model proposed by Li *et al.* [14, 15] The Eigen-PEP integrates information from all the video frames by a part-based average pooling through the PEP model,

which produces an intermediate high dimensional, part-based, pose-invariant representation. It then compacts the high dimensional intermediate representation by principal component analysis (PCA), after which only a small number (as small as 100) of principal eigen dimensions is retained. This compact video representation maintains the flexibility from the nature of average pooling to incorporate or exclude frames incrementally. We then adopt the joint Bayesian classifier [16] to implement face recognition based on the Eigen-PEP representation. The high-level work-flow of the video face recognition system based on Eigen-PEP is summarized in Figure 2. We utilize the PEP model [14, 15] and PCA to construct the Eigen-PEPs for videos (Section 3.2). In the training stage, the joint Bayesian classifier [16] is trained from a set of matched and mismatched video pairs (Section 3.3) represented in the Eigen-PEPs. The classifier is then applied to compare two face videos in the testing stage for either video face verification or identification. In practice, the storage size of an Eigen-PEP produced from one video or multiple videos of a subject can be less than 400 bytes. Hence a system based on Eigen-PEP is highly scalable and the matching process can be very efficient.

In particular, without resorting to more advanced indexing scheme, a video face identification system based on the proposed Eigen-PEP representation would have a run-time that is linear to the number of subjects presented in the gallery database. This is achieved by generating one Eigen-PEP representation per subject, from all videos associated with that subject. Another advantage of the proposed Eigen-PEP representation is that its size is invariant to the length of the input video. Hence, the Eigen-PEP representation can be readily used with more advanced indexing methods, such as tree-based indexing, to further reduce the run-time complexity for identification to be $O(log(n))$, which we defer to our future work.

We evaluate our method on two large-scale video face recognition databases, and an image face recognition dataset, both for face verification and identification. We also participated the recent Point-and-Shoot Face Recognition Challenge (PaSC) [1] and our method significantly outperforms other competitors under the video-to-video face recognition setting [17]. Note the proposed method can be applied to image face recognition naturally by processing an image as a one-frame video. We can also flip the image horizontally to generate a two-frame video, from which we built the Eigen-PEP representation. Therefore, our research contributes to video face recognition in the following aspects:

- We propose a comprehensive, compact, and flexible Eigen-PEP video face representation with superb recognition accuracy.
- We present a highly scalable video face recognition system based on the Eigen-PEP representation.
- We outperform the state-of-the-art recognition accuracy over three challenging face recognition datasets.

---

[1] http://www.cs.colostate.edu/~vision/pasc/ijcb2014/

2   Review of the PEP Model

As we have mentioned, the proposed Eigen-PEP representation is built upon the PEP representation proposed by Li et al. [14, 15]. The PEP representation can deal with a single face or a face set. The PEP representation itself has been shown to be robust to pose variations. When it is applied to video face recognition, a PEP model selects a set of image patches out of all video frames and concatenates the descriptors of the selected image patches into a single vector as the PEP representation.

Although the PEP model presents great potential in modeling human faces, there are several issues when applying it to more practical and large-scale video face recognition. First, the PEP representation is high dimensional (*e.g.*, 1024×128 dimensional using SIFT) which is memory demanding. Second, Li et al. [14] used a kernel Support Vector Machine (SVM) to match two PEP representations for recognition, which is not scalable (Section 4.2).

Third, for modeling video faces, the PEP representation may lose valuable information from appearance variations presented in the video, since it keeps only a small portion of feature descriptors by a part-based probabilistic max pooling. Because of this, although the PEP representation can be incrementally updated to incorporate new observations, it cannot be incrementally updated to remove obsolete observations.

Compared to the PEP representation, the Eigen-PEP representation is more compact, flexible, and comprehensive. We integrate the information from all video frames (even those from multiple videos) by introducing a part-based average pooling to the PEP model. Since we build PEP representation for every frame, the appearance variations under different poses, expressions, and illuminations etc., are integrated.

Because the PEP representations are part-based and robust to pose variations, the corresponded selected descriptors consistently come from the same facial part. Intuitively, the mean of the descriptors from each part can naturally suppress the appearance variations, leading to a robust representation. To address the high-dimensionality problem, we apply PCA over all video-level PEP representations and only retain a small number of principal eigen dimensions.

Since an Eigen-PEP integrates the appearance of different poses and expressions, it is very suitable to represent a subject in a large-scale video face identification system by building a single representation from all videos associated with that subject. Once each gallery person has a single vector representation to incorporate all available videos of him/her, even the brute-force complexity in the testing stage will be linear to the number of gallery identities, instead of the number of gallery videos.

In addition to its compactness and comprehensiveness, the Eigen-PEP benefits from the nature of the average pooling to be flexible for incremental modifications (Section 3.2). For example, we can update the Eigen-PEP incrementally to incorporate new video frames of the same subject, or to remove obsolete video frames, without the need to access all the other video frames used to build the initial representation.

3   The Eigen-PEP Representation

3.1   The PEP Representation

The PEP representation has been shown to be effective in modeling human faces [14, 15]. We refer the readers to Li et al. [14] for the details. To build the PEP representation for a video, all the video frames are firstly processed into a set of descriptors $\{\mathbf{f}\} = \{[\mathbf{a}_i \ \mathbf{l}_i]\}_{i=1}^{M}$, where $[\mathbf{a}_i \ \mathbf{l}_i]$ denotes one spatial-appearance descriptor; $\mathbf{a}$ is the appearance part and $\mathbf{l}$ is the spatial part.

The training stage builds a PEP model (or Universal Background Model in [14]) parameterized by Θ over training descriptors with the Expectation-Maximization (EM) algorithm. The PEP model is a Gaussian mixture model with K spherical Gaussian components,

$$P([\mathbf{a} \ \mathbf{l}]|\Theta) = \sum_{k=1}^{K} \omega_k \mathcal{G}([\mathbf{a} \ \mathbf{l}]|\boldsymbol{\mu}_k, \sigma_k^2 \mathbf{I}), \tag{1}$$

where $\Theta = (\omega_1, \boldsymbol{\mu}_1, \sigma_1, \ldots, \omega_K, \boldsymbol{\mu}_K, \sigma_K)$; $\mathbf{I}$ is an identity matrix; ω_k is the mixture weight of the k-th Gaussian component; $\mathcal{G}(\mu_k, \sigma_k^2 \mathbf{I})$ is a spherical Gaussian with mean μ_k and variance $\sigma_k^2 \mathbf{I}$. Each one of the K Gaussian components commits one descriptor with the highest generative probability and the PEP representation of $\{\mathbf{f}\}$ is the concatenation of the appearance part of the K selected descriptors, i.e,

$$\mathcal{F} = [\mathbf{a}_{g_1} \ \mathbf{a}_{g_2} \ \ldots \ \mathbf{a}_{g_K}], \quad g_k = \arg\max_i \omega_k \mathcal{G}([\mathbf{a}_i \ \mathbf{l}_i]|\boldsymbol{\mu}_k, \sigma_k^2 \mathbf{I}). \tag{2}$$

3.2   The Eigen-PEP Extension

Given a video of multiple frames, we process each single video frame into its PEP representation. Because the PEP representation is part-based and pose-invariant, the PEP representations from the video frames are aligned facial part descriptors. Since the PEP representation is concatenated local descriptors of facial parts, the mean of the corresponding descriptors naturally suppresses the appearance variations across all video frames. Hence the mean of the PEP representations over all the frames is an intermediate high-dimensional part-based video-level representation.

To reduce its dimensionality, we apply Principle Component Analysis (PCA) and keep d principal eigen dimensions. The PCA is trained over all the video-level intermediate PEP representations from the training data. We hence name the video-level representation after PCA the Eigen-PEP.

The workflow for building the Eigen-PEP is shown in Figure 3. Formally, let $\mathcal{F}_1, \mathcal{F}_2, \ldots, \mathcal{F}_N$ denote the PEP representations for video V with N frames; and P denotes the PCA projection. The Eigen-PEP for the video V is

$$f_{eig}(V) = P^T \frac{1}{N} \sum_{n=1}^{N} \mathcal{F}_n. \tag{3}$$

Training video frames and video-level representations
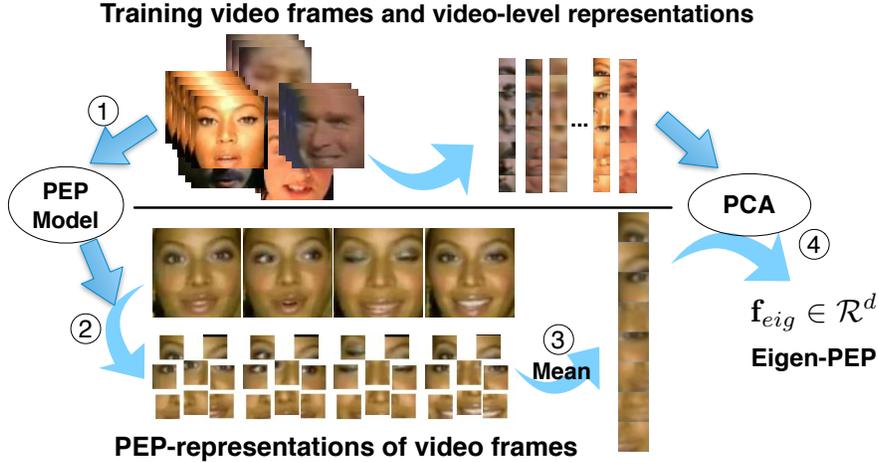
Fig. 3. Workflow for building the Eigen-PEP of a video: 1) the PEP model is learned from training video frames; 2) for each frame in the testing video, the PEP representation is partially visualized as the selected image patches of which the patches at the same location are consistent in semantics but varied in appearance across the video frames; 3) visualization of the intermediate video level representation as the pixel-level mean; 4) apply PCA to project the intermediate video level representation into a low-dimensional space to build the Eigen-PEP; the PCA is trained over all video level intermediate representations.

Compared with the PEP representation, the Eigen-PEP is more comprehensive and compact. In building the intermediate video level representation of the video V, each Gaussian component of the PEP model actually commits N descriptors (one from each video frame), and therefore encodes more appearance variations. The intermediate representation is then built by average pooling per Gaussian component over the N descriptors it selected.

Besides that, benefiting from the nature of this part-based average pooling, the intermediate representation is flexible to incremental modification. Furthermore, the linear nature of the PCA allows the Eigen-PEP to maintain this flexibility. Specifically, with a new video frame $\mathcal{F}_{N+1}$, the Eigen-PEP can be updated incrementally without the need of accessing other video frames, *i.e.*,

$$f_{eig}(V) \leftarrow \frac{N}{N+1} f_{eig}(V) + \frac{1}{N+1} P^T \mathcal{F}_{N+1}. \tag{4}$$

Similarly, to retire the n-th frame it can be deducted from the representation without accessing the other video frames by

$$f_{eig}(V) \leftarrow \frac{N}{N-1} f_{eig}(V) - \frac{1}{N-1} P^T \mathcal{F}_n. \tag{5}$$

3.3   Joint Bayesian Classifier

Chen *et al.* [16] propose the joint Bayesian classifier to explicitly model the intra-person and extra-person variations as zero-mean Gaussians with covariance

matrices Σ_I and Σ_E respectively. The similarity of a face pair (x_1, x_2) is then measured by the likelihood ratio

$$r(x_1, x_2) = log \frac{P(x_1, x_2|H_I)}{P(x_1, x_2|H_E)} = x_1^T A x_1 + x_2^T A x_2 - 2x_1^T G x_2, \qquad (6)$$

where

$$\begin{pmatrix} F & G \\ G & F \end{pmatrix} = \Sigma_I^{-1}, \quad \begin{pmatrix} A & 0 \\ 0 & A \end{pmatrix} = \Sigma_E^{-1} - \begin{pmatrix} F & 0 \\ 0 & F \end{pmatrix}. \qquad (7)$$

H_I and H_E denote the intra-person and extra-person hypothesis parameterized by the covariance matrices Σ_I and Σ_E respectively.

Chen et al. [16] utilizes an EM algorithm relying on identity information to estimate the matrices A and G. In practice, when the identity information is not available, we can estimate Σ_I and Σ_E from matched and mismatched face pairs directly i.e.,

$$\Sigma_I = cov(X_I, X_I), \quad \Sigma_E = cov(X_E, X_E), \qquad (8)$$

where X_I and X_E are the sets of concatenated Eigen-PEP pairs of the matched and mismatched face pairs respectively.

In face verification, we use the joint Bayesian classifier without EM to bypass the necessity of identity information. In face identification, since the identity information is available, we follow the one with EM for better recognition accuracy. Note that only the training time complexity is different in these two cases, the run-time efficiency is the same.

4 Experiments

We perform extensive experiments to evaluate the effectiveness of the proposed representation under different scenarios including video face verification on the YouTube Faces Database [1], large-scale video based face identification on the Celebrity-1000 dataset [18][2], and image face verification on the Labeled Face in the Wild (LFW) dataset [19]. Over all three datasets, our method achieves superior performance compared to the state-of-the-art algorithms.

4.1 Video face verification on YouTube Faces database

In video face verification, the training data is given in the form of matched and mismatched video face pairs. We follow Equation 7 and Equation 8 to learn the matrices A and G.

In the testing stage, the input is a pair of videos V_1 and V_2. After processing the video face pair into Eigen-PEPs $f_{eig}(V_1)$ and $f_{eig}(V_2)$, the joint Bayesian classifier is applied following Equation 6 to assign the similarity score to this video pair, i.e.,

$$r(f_{eig}(V_1), f_{eig}(V_2)) = f_{eig}(V_1)^T A f_{eig}(V_1) + f_{eig}(V_2)^T A f_{eig}(V_2)$$
$$- 2f_{eig}(V_1)^T G f_{eig}(V_2).$$

---

[2] http://www.lv-nus.org/facedb/

We evaluate our method on the YouTube Faces Dataset (YTFaces) published by Wolf *et al.* [1], and compare the result with the state-of-the-art. This dataset contains 3,425 videos of 1,595 different people. Each video consists of 181.3 frames on average. Faces are detected by the Viola-Jones detector and aligned by fixing the coordinates of automatically detected facial feature points [1]. We follow the standard protocol to report the average accuracy over 10-folds evaluation.

In our experiments, video frames are center cropped to 100×100 before feature extraction. To leverage the left-right facial symmetry in the Eigen-PEP, we flip the original video frames horizontally as additional new video frames. We report the recognition accuracy with and without the flipped frames separately.

For the parameters in our system, the SIFT descriptors are extracted over a 3-scale Gaussian image pyramid with scaling factor 0.9, densely from a 8×8 sliding window with 2-pixel spacing. The PEP model consists of 1024 Gaussian components and we keep top 100 eigen vectors in the PCA. Hence the dimensionality of Eigen-PEPs is 100. The storage size of Eigen-PEP for a single video is hence only 400 bytes (100 float values).

As shown in Table 1 and Figure 4, our method outperforms the state-of-the-art algorithms on the YouTube Faces Database under the restricted protocol. Although the Parkhi *et al.* [20] achieves comparable performance to our method, their method relies on large amount of training data in the discriminative dimensionality reduction. As a result, under the restricted protocol, their method produces very high dimensional video representations. Note that on the same dataset, Taigman *et al.* [21] pushed the accuracy as high as $91.4 \pm 1.1\%$. However they leveraged massive outside training data (4 million) while we only use the provided $4,500$ pairs of face tracks for training. Note that there is a list of label errors uploaded to the YouTube Faces webpage recently [21], we also report our result with the corrected labels.

Table 1. Performance comparison over YouTube Faces

Algorithm	Accuracy $\pm$ Error(%)
MBGS [1]	76.4 ± 1.8
MBGS+SVM- [22]	78.9 ± 1.9
STFRD+PMML [23]	79.5 ± 2.5
VSOF+OSS(Adaboost) [24]	79.7 ± 1.8
APEM (fusion) [14]	79.1 ± 1.5
VF^2 [20]	84.7 ± 1.4
DDML (combined) [25]	82.3 ± 1.5
Our method	82.40 ± 1.7
Our method (with flipped frames)	$\mathbf{84.80 \pm 1.4}$
Our method (with flipped frames, corrected labels)	$\mathbf{85.04 \pm 1.49}$

4.2   Video face identification on Celebrity-1000

In terms of video face identification on Celebrity-1000 dataset, there are two categories of protocols: the open-set face identification and the close-set face identification. In both protocols, the task is to identify the identity of the probe
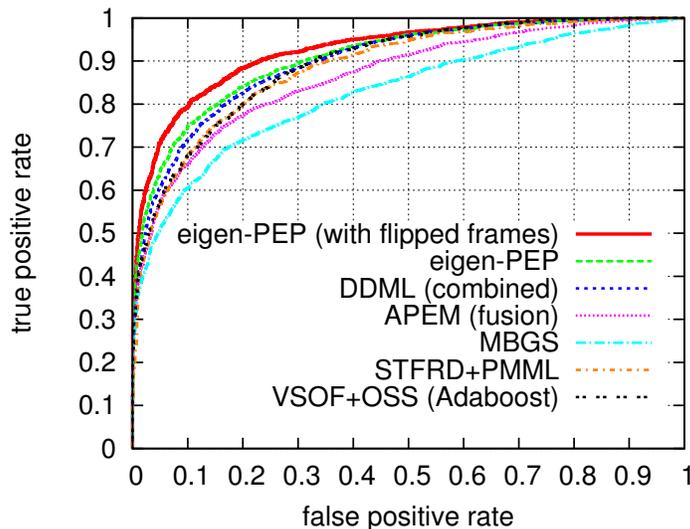
Fig. 4. Performance comparison over YouTube Faces.

face video given a set of gallery face videos. In the open-set protocol, the gallery face videos are not in the training data.

In the training stage, we use the training data to learn the PEP model and PCA projections for the Eigen-PEP representation. After that, for each gallery subject, we build one Eigen-PEP from all his/her videos as the representation of the subject. Since the identity information is available, instead of following Equation 8, we follow Chen *et al.* [16] to train the joint Bayesian classifier.

In the testing stage, the probe face video is firstly processed into Eigen-PEP. Then the similarity between the probe face and each gallery face is measured by the joint Bayesian classifier. The performance of the identification is measured by the cumulative match characteristic curve (CMC) [26] which reports the top k recognition accuracy with varying k.

Liu *et al.* [18] published the Celebrity-1000 dataset to study the large-scale unconstrained video-based face identification problem. This dataset contains 159,726 video sequences of 1,000 human subjects. Faces are detected by the OMRON face detector. We evaluate our method under both the open-set and close-set protocols.

In the open-set protocol, 200 subjects are used for training. In the testing stage, videos are provided as the gallery set and probe set. There are 4 different experimental settings with different number of probe and gallery subjects: 100, 200, 400 and 800. In the close-set protocol, dataset is divided into training (gallery) subset and testing (probe) subset. Similarly, there are 4 settings for close-set: 100, 200, 500 and 1000 subjects.

Considering the relatively low-resolution of the video frames (80×64) in Celebrity-1000, we extract SIFT descriptors in a 8×8 sliding windows with 1-pixel spacing. The PEP model consists of 200 components. In the PCA, we keep 90% accumulated eigen values. We use a maximum of $20,000$ training videos in the PCA. As a result, the dimensionality of Eigen-PEPs varies from 100 to 400

(a) close-set 100 subjects

(b) close-set 200 subjects

(c) close-set 500 subjects

(d) close-set 1000 subjects
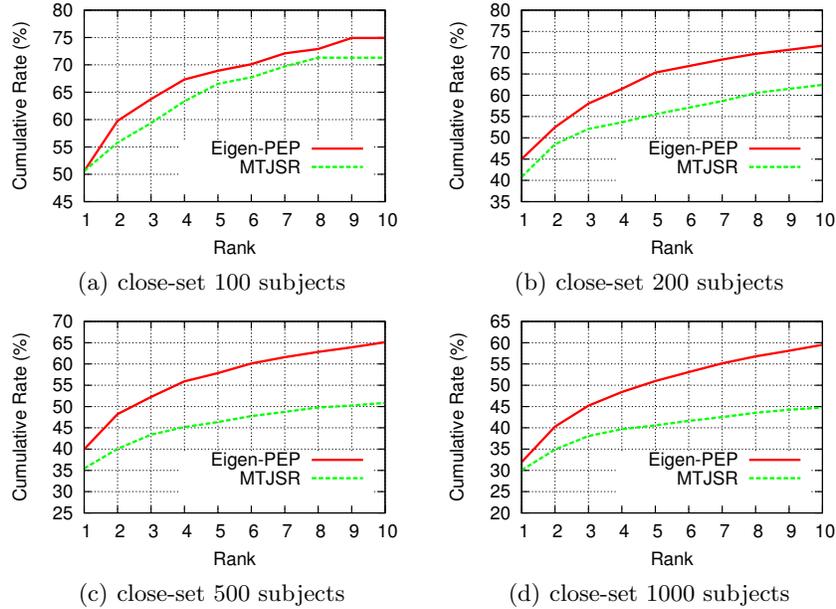
Fig. 5. Performance comparison over Celebrity-1000 dataset (close-set): the curve describes the rank K recognition accuracy.

for different settings. For the open-set protocol, the Eigen-PEP dimension is set to 500. Hence, the storage size of Eigen-PEP for a single gallery subject is no more than 2 kilobytes.
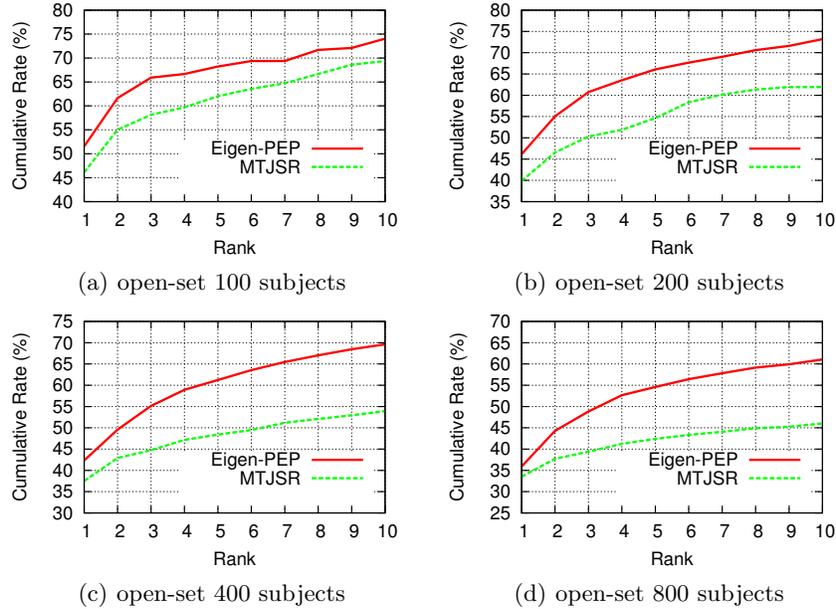
We compare our method with the Multi-task Joint Sparse Representation (MTJSR) [8] which is the current state-of-the-art on Celebrity-1000 [18] [3]. As shown in Figure 5 and Figure 6, and Table 2, our method outperforms the MTJSR algorithm under both the open-set and close-set protocols.

In addition to the superior accuracy, our system is more efficient than the MTJSR. In the testing stage of MTJSR, it solves an optimization problem to represent every frame of the probe video sequence as a sparse linear combination of video sequences of a gallery subject. The classification is then based on the accumulated reconstruction error.

Denote the number of gallery subjects as M, the number of frames of the probe video is N; the number of matching times of MTJSR is generally $N \times M$. Moreover, each matching needs to solve an optimization problem of a sparse representation which by itself is a complex computation. Given the same probe video, the number of matching times in our system, after processing the probe video into Eigen-PEP, is only M. Besides, each matching operation in our system is exactly three times of vector-matrix multiplications and two times of add operations of scalar values. Considering the typical dimension of Eigen-PEP is only a few hundred, our matching operation is far faster.

Specifically, in the close-set protocol with 1000 gallery subjects, the run-time of evaluating one probe video in our system is about 2 seconds, most of which

---

[3] We thank the authors for sharing theirs results.

(a) open-set 100 subjects

(b) open-set 200 subjects

(c) open-set 400 subjects

(d) open-set 800 subjects

Fig. 6. Performance comparison over Celebrity-1000 dataset (open-set): the curve describes the rank K recognition accuracy.

is for building the Eigen-PEP, and the matching time is only 0.05 second. In comparison, the evaluation time of MTJSR for one test sequence, as reported by [18], is 1.6×10^3 seconds. Besides, our experiment is conducted on a single machine with 12 CPU cores (2.4GHz) while Liu *et al.* [18] used a cluster with 14 workstations each of which has 8 CPU cores (3GHz). On average, their run-time is roughly 6 orders of magnitude greater than ours.

We also evaluate the performance of using SVM with the PEP-representations as described in Li *et al.* [14] which takes 41 seconds (matching time) for one query in the 1000 gallery subjects face identification task. Hence in terms of the matching time, our system is 800 times faster than theirs and our video representations are far more storage-efficient.

To further explore how the number of dimensions influence the effectiveness of the Eigen-PEP. We perform an experiment on the open-set 800 subjects setting to evaluate the identification accuracy with Eigen-PEPs of differing dimensions, *i.e.*, 200, 300, 500 and 1000. As shown in Figure 7, except for the rank-1 accuracy when Eigen-PEP is of only 200 or 300 dimensions, all Eigen-PEPs outperforms the MTJSR by a significant margin. This observation also suggests that the dimension of the Eigen-PEP can be a trade-off parameter to balance the accuracy and efficiency.

We present an example result in Figure 9. As observed, different video sequences of the same gallery subject present varied appearance. Nevertheless by representing a subject as one Eigen-PEP representation our system can successfully identify the probe video. This observation demonstrates the Eigen-PEP as a comprehensive video representation.

Table 2. Performance comparison on Celebrity-1000 dataset: showing the rank-K accuracy.

		rank-1 (%)	rank-2 (%)	rank-5 (%)	rank-10 (%)
close-set 100	Eigen-PEP	50.60	**59.76**	**68.92**	**74.90**
	MTJSR	50.60	55.78	66.53	71.31
close-set 200	Eigen-PEP	**45.02**	**52.49**	**65.33**	**71.65**
	MTJSR	40.80	48.47	55.56	62.45
close-set 500	Eigen-PEP	**39.97**	**48.21**	**57.85**	**65.09**
	MTJSR	35.46	40.05	46.35	50.86
close-set 1000	Eigen-PEP	**31.94**	**40.27**	**51.01**	**59.50**
	MTJSR	30.04	34.88	40.58	44.77
open-set 100	Eigen-PEP	**51.55**	**61.63**	**68.22**	**74.03**
	MTJSR	46.12	55.04	62.02	69.38
open-set 200	Eigen-PEP	**46.15**	**55.03**	**66.07**	**73.18**
	MTJSR	39.84	46.55	54.64	61.93
open-set 400	Eigen-PEP	**42.33**	**49.57**	**61.23**	**69.62**
	MTJSR	37.51	42.91	48.41	53.91
open-set 800	Eigen-PEP	**35.90**	**44.27**	**54.60**	**61.07**
	MTJSR	33.50	37.71	42.41	46.03
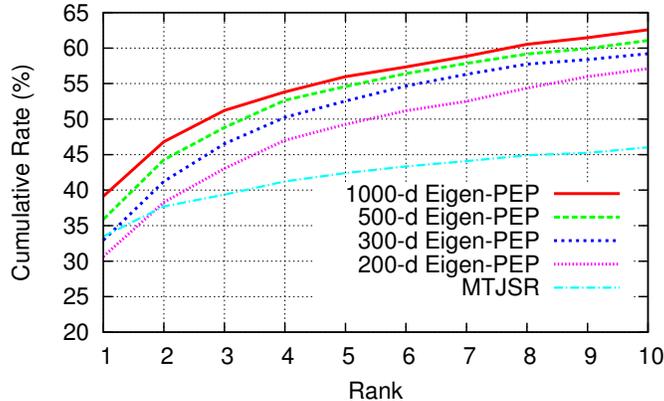
Fig. 7. Performance of different dimensional Eigen-PEPs over Celebrity-1000 dataset.

4.3   Image Face Verification

Although we propose the compact PEP representation for video-based face recognition, it naturally applies to the image-based setting by processing the image as a one-frame video. Furthermore, we can actually generate a two-frame video by horizontally flipping the face image to better leverage the facial symmetry in the Eigen-PEP representation. The Labeled Faces in the Wild (LFW) [19] dataset is designed to address the unconstrained image-based face verification problem. This challenging dataset contains more than 13,000 images from 5,749 people.

We follow the *image-restricted, no outside data* protocol of LFW [27] using the faces roughly aligned with the funneling method [28]. Besides that we do not leverage any external data for strong face alignment, feature extraction or recognition model training.

Similarly, we extract SIFT descriptors in 8×8 sliding window with 2-pixel spacing in the center cropped 150×150 images. The PEP model is of 1024 components and the PCA reduces the dimensionality to 100.

As shown in Table 3 and Figure 8, our method outperforms the state-of-the-art algorithms on LFW. We also evaluate the performance of combining the joint Bayesian classifier with the PEP representation. Since it is not practical to apply joint Bayesian classifier over the high-dimensional PEP representation directly due to the large size of covariance matrices, we apply PCA to reduce the dimensionality of PEP representation to be 100 as well. To be fair, the PCA is trained separately over training PEP representations.

In a single frame case, our method is equivalent to applying the joint Bayesian classifier for the PEP representation after PCA. Compared with the results from APEM by Li *et al.* [14], which is essentially the PEP representation with a kernel SVM on the absolute difference of the PEP representations for verification with an additional step of Bayesian adaptation, it clearly shows the advantage of adopting the joint Bayesian classifier. We believe that taking the absolute difference of two PEP representations resulted in loss of important discriminative information.

Table 3. Performance comparison on the LFW, under *image-restricted, no outside data* protocol

Algorithm	Accuracy $\pm$ Error(%)
V1/MKL [29]	79.35 ± 0.55
Simonyan et al. [30]	87.47 ± 1.49
APEM (Fusion) [14]	84.08 ± 1.20
1-frame Eigen-PEP	86.27 ± 1.06
2-frame PEP representation	87.37 ± 0.66
2-frame Eigen-PEP	$\mathbf{88.47 \pm 0.91}$
2-frame Eigen-PEP (fusion)	$\mathbf{88.97 \pm 1.32}$

We also compare with the 2-frame PEP representation setting, in which a single PEP representation is built for the two images. Similarly, PCA is applied to the PEP representation and the joint Bayesian classifier is adopted for classification. As observed, the Eigen-PEP consistently outperforms the PEP representation in all the cases.

Following similar process in Li et al. [14], by fusing the additional result using Local Binary Pattern (LBP) [31] descriptors with a linear SVM, we observe further improvement on LFW.
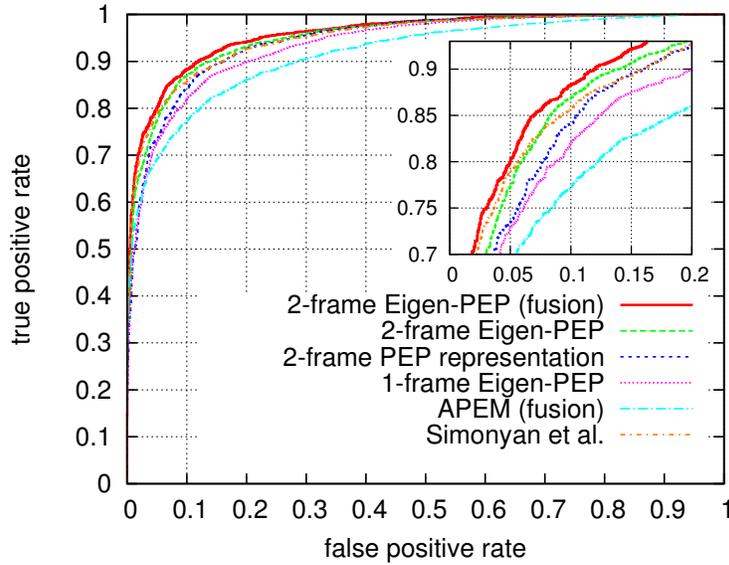
5   Conclusion

In this paper, we propose the Eigen-PEP video face representation. We combine the Eigen-PEP with the joint Bayesian classifier for video face recognition. The Eigen-PEP naturally integrates information from all video frames and is flexible to dynamical modification. The small footprint of the proposed Eigen-PEP makes the overall video face recognition framework to be scalable and be
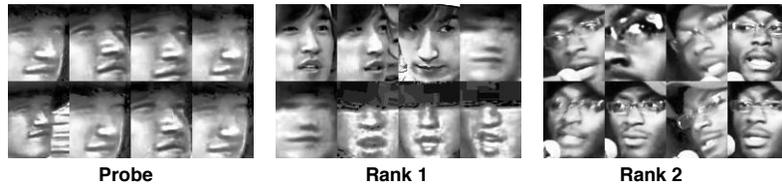
suitable for large-scale video face identification. Extensive experiments are conducted over three challenging real-world face recognition datasets to evaluate the proposed method in video face verification, video face identification and image face verification. The proposed method outperforms the existing state-of-the-art algorithms under all three tasks.

Acknowledgement

Fig. 8. Performance comparison on the LFW, under *image-restricted, no outside data* protocol

Fig. 9. Qualitative result on Celebrity-1000: shows a successful query and the top 2 candidates ranked by our system; 8 frames of the probe video are shown; 8 frames of the gallery subjects are selected from 8 video sequences chosen randomly.

References

1. Wolf, L., Hassner, T., Maoz, I.: Face recognition in unconstrained videos with matched background similarity. In: CVPR. (2011)
2. Chen, D., Cao, X., Wen, F., Sun, J.: Blessing of dimensionality: High dimensional feature and its efficient compression for face verification. In: CVPR. (2013)
3. Cao, X., Wipf, D., Wen, F., Duan, G.: A practical transfer learning algorithm for face verification. In: ICCV. (2013)
4. Liao, S., Jain, A., Li, S.: Partial face recognition: Alignment-free approach. T-PAMI (2013)
5. Barkan, O., Weill, Y., Wolf, L., Aronowitz., H.: Fast high dimensional vector multiplication based face recognition. In: ICCV. (2013)
6. Lei, Z., Pietikainen, M., Li., S.Z.: Learning discriminant face descriptor. In: T-PAMI. (2014)
7. Cao, Q., Ying, Y., Li, P.: Similarity metric learning for face recognition. In: ICCV. (2013)
8. Yuan, X.T., Liu, X., Yan, S.: Visual classification with multitask joint sparse representation. Image Processing, IEEE Transactions on (2012)
9. Chen, Y.C., Patel, V., Shekhar, S., Chellappa, R., Phillips, P.: Video-based face recognition via joint sparse representation. In: Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on. (2013)
10. Zhou, S., Krueger, V., Chellappa, R.: Probabilistic recognition of human faces from video. Comput. Vis. Image Underst. (2003)
11. Lee, K.C., Ho, J., Yang, M.H., Kriegman, D.: Video-based face recognition using probabilistic appearance manifolds. In: Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on. (2003)
12. Zhang, Y., Martnez, A.M.: A weighted probabilistic approach to face recognition from multiple images and video sequences. Image and Vision Computing (2006)
13. Zhao, W., Chellappa, R., Rosenfeld, A., Phillips, P.J.: Face Recognition: A Literature Survey. ACM Computing Surveys (2003)
14. Li, H., Hua, G., Lin, Z., Brandt, J., Yang, J.: Probabilistic elastic matching for pose variant face verification. In: CVPR. (2013)
15. Li, H., Hua, G., Lin, Z., Brandt, J., Yang, J.: Probabilistic elastic part model for unsupervised face detector adaptation. In: ICCV. (2013)
16. Chen, D., Cao, X., Wang, L., Wen, F., Sun, J.: Bayesian face revisited: A joint formulation. In: ECCV. (2012)
17. Beveridge, J.R., Zhang, H., Flynn, P.J., Lee, Y., Liong, V.E., Lu, J., de Assis Angeloni, M., de Freitas Pereira, T., Li, H., Hua, G., Struc, V., Krizaj, J., Phillips, P.J.: The IJCB 2014 pasc video face and person recognition competition. In: IJCB. (2014)
18. Liu, L., Zhang, L., Liu, H., Lao, S., Yan, S.: Towards large-population face identification in unconstrained videos. CSVT (2013)
19. Huang, G.B., Mattar, M., Berg, T., Learned-Miller, E.: Labeled Faces in the Wild: A Database forStudying Face Recognition in Unconstrained Environments. In: Faces in Real-Life Images Workshop in ECCV. (2008)
20. Parkhi, O.M., Simonyan, K., Vedaldi, A., Zisserman, A.: A compact and discriminative face track descriptor. In: CVPR. (2014)
21. Taigman, Y., Yang, M., Ranzato, M., Wolf, L.: DeepFace: Closing the Gap to Human-Level Performance in Face Verification. In: CVPR. (2014)

22. Wolf, L., Levy, N.: The svm-minus similarity score for video face recognition. In: CVPR. (2013)
23. Cui, Z., Li, W., Xu, D., Shan, S., Chen, X.: Fusing robust face region descriptors via multiple metric learning for face recognition in the wild. In: CVPR. (2013)
24. Mendez-Vazquez, H., Martinez-Diaz, Y., Chai, Z.: Volume structured ordinal features with background similarity measure for video face recognition. In: ICB. (2013)
25. Hu, J., Lu, J., Tan, Y.P.: Discriminative deep metric learning for face verification in the wild. In: CVPR. (2014)
26. Moon, H., Phillips, P.J.: Computational and performance aspects of pca-based facerecognition algorithms. Perception (2001)
27. Huang, G.B., Learned-Miller, E.: Labeled faces in the wild: Updates and new reporting procedures. Technical Report UM-CS-2014-003, UMass Amherst (2014)
28. Huang, G., Jain, V., Learned-Miller, E.: Unsupervised joint alignment of complex images. In: ICCV. (2007)
29. Pinto, N., DiCarlo, J.J., Cox, D.D.: How far can you get with a modern face recognition test set using only simple features? In: CVPR. (2009)
30. Simonyan, K., Parkhi, O.M., Vedaldi, A., Zisserman, A.: Fisher Vector Faces in the Wild. In: BMVC. (2013)
31. Ahonen, T., Hadid, A., Pietikainen, M.: Face recognition with local binary patterns. In: ECCV. (2004)