Video-Based Face Recognition Using the Intra/Extra-Personal Difference Dictionary

Ming Du
mingdu@umd.edu
Rama Chellappa
rama@umiacs.umd.edu

Department of Electrical and Computer Engineering
University of Maryland
College Park, USA

Abstract

Face recognition in unconstrained videos is challenging due to large variations in pose, illumination, expression etc. We address the problem from two different aspects: To handle pose variations, we learn a Structural-SVM based detector which can simultaneously localize face fiducial points and estimate the face pose. By adopting a different optimization criterion from existing algorithms, we are able to improve localization accuracy. To model other face variations, we use intra-personal/extra-personal dictionaries. The proposed framework is advantageous in terms of both accuracy and scalability. We demonstrate through experiments that our algorithm achieves state-of-arts performance on challenging public databases, even when the training data come from a different database.

1 Introduction

We are witnessing a growing interest in video-based face recognition (VFR) research in recent years. Part of this is driven by the increasing demand for processing digital video contents over the Internet. It is reported that over 14,000 hours of new videos are uploaded to Youtube every day. From a technical perspective, the attraction of videos comes from the fact that they contain extra spatial-temporal information that can be exploited to improve recognition performance. Moreover, videos arise naturally in many applications like surveillance. It is expected that VFR can play an important role in cases where the still image-based algorithms do not return satisfactory results.

In this paper, we attempt to improve the performance of VFR in the following two aspects:

Face Localization and Normalization As the first steps in almost every VFR algorithm, this is where we try to bridge the gap between unconstrained and constrained videos in terms of source data quality. Recent advances in object detection technology have stimulated new research on "tracking by detection" approaches and facial feature detectors. The former is more robust against drift errors in comparison with traditional trackers. The latter enables us to perform accurate face alignment when large pose variations are present. However, tracking and aligning faces "in the wild" is still a highly challenging task.

Scalability and Generalization The majority of existing VFR algorithms are devoted to discovering features which are closely correlated with identity. However, it requires a large amount of training data to effectively characterize a subject. More often than not, we have

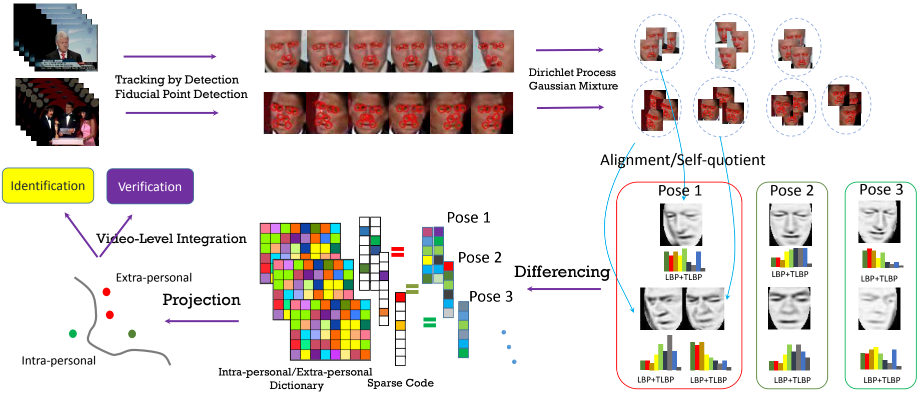

Figure 1: Processing pipeline of the proposed video-based face recognition algorithm.

insufficient training samples to account for all possible variations for each subject. As a result, decision boundaries of the classifiers are often highly dependent on the training data and are prone to change every time we add new subjects to the database. Such a strategy is inherently inflexible and unscalable. Moreover, for practical uses, while it is desirable that a VFR algorithm be capable of working across databases, most existing approaches have difficulty in addressing this issue.

Our fully automatic VFR algorithm works as follows: Faces are first localized from videos using a tracking by detection approach. A fiducial point detector is then applied to each tracked face. The detector is a structural SVM (SSVM) learned by optimizing an objective function that emphasizes on improved localization accuracy. It provides both coordinates of feature points and the quantized face pose. Based on the estimated pose, the localized faces are then aligned to pose-specific common reference coordinate frames. They are further clustered using a non-parametric Bayesian model to remove temporal redundancy. We construct pose-specific dictionaries as our classifiers. However, in our work, the discriminative dictionaries do not directly assign an identity label to each test sample. Rather, it attempts to distinguish the intra-personal face appearance variations from the extra-personal ones. Such dictionaries are generic in nature and are capable of working across data domains. An overview of the proposed approach is given in Figure 1.

Our contributions are three-fold: First, we develop a novel VFR algorithm based on discriminative dictionary learning and the concept of intra-personal variations. As a result, the algorithm can achieve good performance in terms of accuracy, generalizability and scalability at the same time. Second, we propose an end-to-end solution to the real-world VFR problem. It allows us to reliably localize and recognize face videos “in the wild”. Third, we demonstrate through comprehensive experiments that the proposed algorithm outperforms state-of-arts methods on multiple public VFR databases.

2 Related Works

Video-based face recognition can be viewed as a special case of a broader category: face recognition based on image set. In practice, the two terms are often used interchangeably when the image sets are sampled from videos. Various representations of image sets have been explored, including linear subspaces [30], dictionaries [5, 22], manifolds [10, 15, 25, 26], probability distributions [1], dynamical models [19] etc.

In recent years, sparse coding [4] has gained popularity in the field of image classification. Wright et al. [28] successfully applied their Sparse Representation-based Classification (SRC) framework to the still image-based face recognition problem. The K-SVD algorithm [23], an iterative method to learn over-complete dictionaries, is one of the most widely used dictionary learning approaches. However, as it focuses on the reconstruction error using sparse codes, K-SVD is not well suited for classification tasks. Many discriminative dictionary learning algorithms which include classification error terms in the objective function have been proposed. Jiang et al. [13] presented a discriminative dictionary learning framework by enforcing label consistency constraints in addition to sparsity and reconstruction error terms. The projection matrix used for classification is learned along with the dictionary. In [31], the additional constraints include the discriminative fidelity terms and the discriminative coefficient term based on Fisher's discriminant.

In [20], Moghaddam et al. first proposed the Bayesian face recognition algorithm. The intrapersonal subspace is defined as the subspace constructed from within-class sample differences. Similarly, the extrapersonal subspace is constructed using the between-class sample differences. At test time, the difference between a probe and a galley image is projected onto the two subspaces and a Bayesian classifier is applied to obtain the recognition result. The metric learning approach proposed in [9] attempted to learn a symmetric positive definite matrix, which can be used to calculate the Mahalanobis distance for a pair of images. This is closely related to the method based on Gaussian densities proposed for Bayesian face recognition.

Face fiducial points detection has been shown to be critical for solving the unconstrained face recognition problem. Various detectors [2, 8, 32] have been proposed to utilize both spatial relationship and appearance information to localize the feature points. Zhu and Ramanan [32] extended the Deformable Parts Model (DPM) for face detection, pose estimation and feature localization. The model is also a mixture of tree-structure sub-models, each of which corresponds to a pose prototype. It is trained using the max margin criterion and hence can be globally optimized. The facial feature detector used in our work shares some similarities with this work in that we also enforce max margin constraints to train a mixture of pose-specific models. However, as we show in Section 3, while their objective function is designed to guarantee the capabilities of detecting both the whole face and the facial features, ours is tuned to improve the accuracy in fiducial points localization.

3 Face Localization and Alignment

Our face localization module falls in the "tracking-by-detection" paradigm. We apply a Viola-Jones face detector to each frame of a video. Then we evaluate the image likelihood of each face candidate as:

$$L(\mathbf{x}_{i,t}|I_t) = \ln \mathcal{N}(\mathbf{x}_{i,t}|\mathbf{x}_{t-1}, \mathbf{\Sigma}) + \lambda \ln p(\mathbf{x}_{i,t}|W_{t-1}) \tag{1}$$

, where $\mathbf{x}_{i,t}$ is the bounding box's coordinates of the i-th face candidate found in frame I_t, and W is a WSL appearance model [12] updated at each frame. Apparently, the two terms penalize location inconsistency and appearance inconsistency respectively. The parameter λ which determines the relative weights of the two terms is usually set empirically. The candidate with the largest likelihood is added to the face track and updates the appearance model. If no detection responses have likelihood values above a set threshold or no faces are detected at all in the current frame, a particle filter will be initiated [1]. It performs face tracking until the detector starts to find a valid face again. The particle filter also uses the likelihood model as defined in (1). This simple strategy proved to be very effective in our experiments.

To detect face fiducial points from the localized face, we train an SSVM. Its coefficients control the relative weights of feature functions which are computed based on a mixture of pictorial structure models $\{T_m, m = 1,2,...,M\}$. Each component of the mixture accounts for the configuration of fiducial points for a specific range of face poses. Here, we divide the poses according to the yaw angle of face and the boundaries are set as $\{-45^\circ, -30^\circ, -15^\circ, 15^\circ, 30^\circ, 45^\circ\}$. We opt for multiple pose-specific models rather than a single shared model for two reasons. First, face fiducial points could have totally different configurations across poses. For example, when a face is in profile pose, half of the fiducial points will be occluded. Even for those feature points which are visible in all the poses, the pose-specific model can enforce constrains on the state space. Second, such a mixture model will allow us to estimate the face pose as a byproduct. Pose information is required when we construct the intra/extra-personal dictionaries at the next stage. Note that for the purpose of face alignment, usually a set of sparse features is sufficient. Following [24], we pick eye corners, mouth corners, nose corners and nose tip as points of interest. Intuitively, the number of feature points in each model varies due to occlusion.

The structure of our face fiducial point model is similar to that of the mixture of pictorial model defined in [32]. For a fiducial point configuration $z = \{L, m\}$, where $L = \{l^i\} = \{(x^i, y^i)\}$ are the image coordinates and m is index of the mixture component that the fiducial points are associated with, we define its score function as:

$$f(I,\mathbf{z}) = \mathbf{w}^T\boldsymbol{\Phi}(I,\mathbf{z}) = \mathbf{w}_m^T\phi_m(I,L) = \sum_{i\in V_m}\mathbf{q}_m^{iT}\psi_m(I,l^i) + \sum_{ij\in E_m} a_m^{ij}dx^2 + b_m^{ij}dx + c_m^{ij}dy^2 + d_m^{ij}dy \tag{2}$$

, where $\mathbf{w}^T = [\mathbf{w}_1^T, \mathbf{w}_2^T, ..., \mathbf{w}_M^T]$, $\boldsymbol{\Phi}(I,\mathbf{z})^T = [0,...,0,\phi_m(I,L),0,...0]$. In (2), V_m and E_m are the nodes and edges of the m-th pictorial model in the mixture, respectively. $\psi_m(I,l^i)$ is a local visual descriptor extracted at the neighbourhood of l^i. In our case, the CENTRIST descriptor [29] is used. For every pair of fiducial points connected by an edge, the pairwise term in (2) captures their spatial relationship. As defined in [32], dx and dy are the displacements of fiducial point i w.r.t. fiducial point j in x and y directions. The sparse augmented feature function $\phi_m(I,L)$ only activates the mixture component whose index is encoded in $\mathbf{z}$. We can jointly localize the fiducial points and estimate the face pose by maximizing the potential function: $\mathbf{z}^* = \{L^*, m^*\} = \underset{L,m}{\operatorname{argmax}}\,\mathbf{w}_m^T\phi_m(I,L)$.

To learn the parameter $\mathbf{w}$, we solve the following margin re-scaling structure SVM problem:

$$\min_{\mathbf{w}}\frac{1}{2}\|\mathbf{w}\|^2 + C\sum_n \max_{\mathbf{z}\in\mathcal{Z}}[\Delta(\mathbf{z},\mathbf{z}_n) + \mathbf{w}^T\boldsymbol{\Phi}(I_n,\mathbf{z})] - \mathbf{w}^T\boldsymbol{\Phi}(I_n,\mathbf{z}_n) \tag{3}$$

. In (3), $(I_n, \mathbf{z}_n)$ is an image-label pair in the training database and $\mathcal{Z}$ is the viable label configuration set. As in the single-output SVM case, each training sample is assigned with a slack variable ξ_n to relax the constraints. $\Delta(\mathbf{z},\mathbf{z}_n)$ is the loss function of a output $\mathbf{z}$ when measured against the ground-truth label $\mathbf{z}_n$. Suppose there are S fiducial points in total and the subset of indexes of those fiducial points visible for the m-th pictorial model is $S(m)$. The

[1] In most of the videos we experimented with, at least one face is present in each frame.

loss function is defined as follows:

$$\Delta(\mathbf{z},\mathbf{z}_n)=\sum_{s=1}^{S}\|\delta_s\|_2,\ \delta_s=\begin{cases}L_s-L_{n,s} & \text{if } s\in S(m)\cap S(m_n)\\ L_s & \text{if } s\in S(m)\setminus S(m_n)\\ c & \text{if } s\in S(m_n)\setminus S(m))\end{cases} \tag{4}$$

. We assign a constant c in the third case because if a false positive feature point shows up in prediction, it should be penalized uniformly, irrespective of its coordinates.

In comparison, the optimization function used in [32] is:

$$\min_{\mathbf{w},\,\xi_n\geq 0}\frac{1}{2}\|\mathbf{w}\|^2+C\sum_n\xi_n$$
$$s.t.\ \forall n,\forall\, I_n\in neg,\mathbf{z}\in\mathbf{z}\,\mathbf{w}^T\boldsymbol{\Phi}(I_n,\mathbf{z})\leq-1+\xi_n\ \forall I_n\in pos,\mathbf{w}^T\boldsymbol{\Phi}(I_n,\mathbf{z}_n)\geq 1-\xi_n,\ \forall k,\mathbf{w}_k\leq 0 \tag{5}$$

, where *pos* contains the positive training images with a face and *neg* contains the negative ones with only background. Apparently, the constraints in (5) focus on the margin between face and non-face images. In contrast, we use a different definition about positive and negative training samples: Every training image in our case has a face in it. The positive samples are the ground-truth fiducial point configurations of the faces and the negative samples are just any configurations other than the ground-truth ones. Therefore, our objective function explicitly imposes constraints on the margin between correct and wrong landmark predictions. Moreover, while [32] treats all the fiducial point configuration equally for a negative training image, in our case the margin is re-scaled by a loss function $\Delta(\mathbf{z},\mathbf{z}_n)$ which penalizes the negative samples according to their misalignment errors. In summary, our method is not designed to detect face and facial feature points simultaneously as in [32]. Instead, it aims for higher accuracy in localizing the landmarks from a previously detected face.

We employ the subgradient algorithm to learn the parameter $\mathbf{w}$. At test time, we follow a two-step procedure to solve the inference problem defined in the objective function. First, we solve for the best L for each individual model in the mixture. Although the cardinality of the entire configuration space is extremely large (in the order of 10^{18}), we only need to be concerned with a very small portion of it at run-time, thanks to the models' tree structure. Dynamic programming (more specifically in this case, the Max-sum inference algorithm) can be applied at this step to select the best configuration efficiently. Then we compare across models to choose the optimal solution. The result of model selection also gives a rough estimate of face pose. We detect fiducial points on every face localized by the detector or tracker and it takes about 30 ms on a workstation equipped with an Intel Core i5 3.3GHz CPU. A linear conformal image transformation calculated from point correspondences is then applied to align faces to a canonical frame. Note that there are M such canonical frames, each of which is associated with a model from the mixture.

4 Intra-personal/Extra-personal Difference Dictionary

4.1 Sparse Coding

We now discuss the problem of modeling intrapersonal face appearance differences using sparse coding. Since video can be viewed as a special case of an image set, we will first discuss general image/frame-based recognition using the intrapersonal dictionary and leave the video case to Section 4.2. Let $\mathbf{X}=\{\mathbf{x}_i,i=1,2,...,N\}\in R^{d\times N}$ be the set of vectorized intrapersonal difference training images/frames. The intrapersonal dictionary $\mathbf{D}=[D_1,D_2,...,D_K]$, where $D_k \in R^d$, is learned by solving the following constrained optimization problem:

$$\min_{D,\alpha} \sum_{i=1}^{N} \frac{1}{2} \|\mathbf{x}_i - \mathbf{D}\alpha_i\|_2^2 \; s.t. \; \forall i, \; \|\alpha_i\|_0 < \varepsilon \tag{6}$$

. In other words, the goal is to minimize the L_2 reconstruction error and guarantee the reconstruction coefficient vector to be sparse at the same time. Each D_k is called an atom of the dictionary, and α_i is called a sparse code. The sparsity parameter ε is usually empirically chosen from the range 10 to 40. One of the most frequently used dictionary learning methods is the K-SVD algorithm [23]. It is an iterative procedure with two alternating optimization steps: First fix the dictionary to solve for the sparse code, and then fix the sparse code to update the dictionary.

4.2 Label-Consistent Dictionary Learning for VFR

It has been argued that separating dictionary learning from classier design may lead to suboptimal solutions for the final classification task. In view of this, we follow the Label-Consistent K-SVD (LC-KSVD) algorithm [13] to jointly learn a generative shared dictionary and a discriminative projection matrix. Although the shared dictionary is composed of two sub-dictionaries corresponding to intra-personal and extra-personal differences respectively, the sparse code of any input difference vector is computed by using the complete set of atoms in the dictionary. This is different from the class-specific dictionaries in Section 4.1. On the other hand, a matrix $\mathbf{W} \in R^{2\times d}$ that encodes the discriminative information of the sparse codes is learned along with the shared dictionary. For the sparse codes $\mathbf{A} = [\alpha_1, \alpha_2, ..., \alpha_N]$ resulting from s set of intra-personal and extra-personal difference vectors, the projection $\mathbf{WA}$ is supposed to form two well-separated clusters. Aside from that, the LC-KSVD also looks for a linear transformation $\mathbf{B} \in R^{K\times d}$ which encourages the samples from the same class to be reconstructed using similar atoms, i.e. the entries in the sub-dictionary of that class. This constraint can be written in the form: $\mathbf{BX} = \mathbf{Q}$, where $\mathbf{Q} \in R^{K\times N}$ has a block diagonal form: The c-th block contains entry $Q_{ij}, i \in \mathbf{v}_c, j \in \mathbf{h}_c$, where v_c are the indices of atoms from class c (i.e. intra-personal or extra-personal) and $\mathbf{h}_c$ are the indices of training instances from class c. All the non-zero entries in $\mathbf{Q}$ are assigned with unit value. To summarize, the final optimization problem has the following form:

$$\min_{\mathbf{D},\mathbf{A}} \|\mathbf{X} - \mathbf{DA}\|_2^2 + \mu\|\mathbf{Q} - \mathbf{BA}\|_2^2 + \sigma\|\mathbf{F} - \mathbf{WA}\|_2^2 + \lambda \sum_i \|\alpha_i\|_1 \tag{7}$$

, where the columns of $\mathbf{F} \in R^{2\times N}$ are labels of the training instances in $\mathbf{X}$, represented using the 1-of-K coding scheme. (7) can be converted to a typical K-SVD objective function and solved using the same procedure.

According to a large body of empirical research, pose variations often cause within-class variance to exceed between-class variance in face recognition. Predictably, they present a great challenge to the intrapersonal/extrapersonal difference dictionary learning. Therefore, we choose to separate pose from other nuisance factors which case variations in the intrapersonal/extrapersonal domain. To this end, we first group the aligned training images according to face pose that has been estimated along with the fiducial points in 3. The difference images are then calculated within each pose group and are used to learn pose-specific shared dictionaries $\{\mathbf{D}^m\}$, where m corresponds to the mixture index in Section 3. Naturally, to predict the class label (i.e. same-person or different-person) of a difference image with pose m at test time, only the dictionary $\mathbf{D}^m$ is relevant and will be activated in calculations. Therefore, we drop the mixture/pose superscript to avoid cluttered notation and keep the dependency on pose implicit.

In our work, calculating sparse codes for every frame pair is not only computationally expensive, but also unnecessary due to the significant temporal redundancy present in video signals. The redundancy can be removed by finding representative frames, which was often accomplished using the K-means algorithm. However, it is still an open problem to adaptively determine K at run-time, and it is obvious that a pre-determined K would be unsatisfactory considering the large variations of video contents. In view of that, we choose to fit a non-parametric Bayesian model to each video. The resulting model has infinite number of Gaussian mixtures controlled by a Dirichlet process $DP(\beta,H)$ [17], where β is the concentration parameter and H is the base probability measure. The mixture weights $\{\pi_k, k = 1,2,...,\infty\}$ are generated from the Griffiths-Engen-McClosky (GEM) process [21], i.e.:

$$\pi_k = \rho_k \prod_{l=1}^{k-1}(1-\rho_l) \quad \rho_k \sim Beta(1,\beta) \tag{8}$$

. The mean and covariance parameters $\{\theta_k\}$ of the mixtures are sampled from H. Given a video V, we assume that each frame $\{I_f, f = 1,...,F\}$ is assumed to be generated by first drawing a component label z_f from a Multinoulli distribution with parameter $\{\pi_k, k = 1,2,...,\infty\}$ and then sample from a Gaussian distribution with parameter $\{\theta_k\}$. We adopt the variational inference approach to fit the model due to its efficiency. The posterior distribution $P(z_f|V)$ is used for clustering. By using the Dirichlet process mixture model, new clusters can be generated when more frames are observed, and there is no need to know number of clusters a priori.

After fitting the model, a video $\mathbf{V}$ with K clusters can be characterized by the set of cluster centers. We further exrtact feature vectors $\{\mathbf{v}^k, k = 1,2,...,K\}$ from these representative images. Both training and test videos go through this process. For the training videos, the intrapersonal features $\{\mathbf{x}_{In} = \mathbf{v}_i^m - \mathbf{v}_j^n, ID(\mathbf{V}_i) = ID(\mathbf{V}_j)\}$ and the extrapersonal ones $\{\mathbf{x}_{Ex} = \mathbf{v}_i^m - \mathbf{v}_j^n, ID(\mathbf{V}_i) \neq ID(\mathbf{V}_j)\}$ are employed to learn the dictionary $\mathbf{D}$ and the projection matrix W. At the test stage, we iterate over every probe-gallery video pair $\{\mathbf{V}_p, \mathbf{V}_g\}$ and calculate feature difference vectors $\{\mathbf{x}_{p,g}^{m,n} = \mathbf{v}_p^m - \mathbf{v}_g^n\}$ from the representative cluster centers. We then solve for the sparse representation of $\mathbf{x}_{p,g}^{m,n}$: $\alpha_{p,g}^{m,n} = \underset{\alpha}{\operatorname{argmin}} \sum_{i=1}^N \frac{1}{2}\|\mathbf{x}_{p,g}^{m,n} - \mathbf{D}\alpha\|_2^2 + \lambda\|\alpha\|_1$. As mentioned earlier, there is an implicit pose index in the equations above. That is, we only calculate feature vector differences for face images with the same pose, and activate the dictionary of the corresponding pose to compute sparse codes.

For video-based recognition, we have: $ID(\mathbf{V}_p) = \underset{g}{\operatorname{argmax}}\, s(p,g)$, where

$$s(p,g) = \sum_{m=1}^{M}\sum_{n=1}^{N} \mathbf{1}(\mathbf{t}_1\mathbf{W}\alpha_{p,g}^{m,n} > \mathbf{t}_0\mathbf{W}\alpha_{p,g}^{m,n})/MN \tag{9}$$

. Here, we use $\mathbf{t}_0 = [0,1]^T$ and $\mathbf{t}_1 = [1,0]^T$ to denote the 1-of-K coding label for intrapersonal and extra-personal class, respectively. $\mathbf{1}(\cdot)$ is the indicator function. One of the attractive features of the proposed algorithm is that it naturally fits in with the verification protocol. In a hard decision scheme, for each video pair $\{\mathbf{V}_p, \mathbf{V}_g\}$, we apply majority voting on top of the binary "same person/different person" results of the frame pairs. This will yield a single operating point on the ROC curve. Alternatively, we may adopt a soft decision rule. The entry of the similarity matrix is the same as the $s(p,g)$ defined in (9).

Table 1: Comparison of Video-Based Face Recognition Results on the Youtube Celebrity Video and the Honda/UCSD database

Method	Youtube Celebrity	Honda/UCSD
MSM[30]	61.1	92.5
MMD[26]	62.9	97.1
MDA[25]	65.3	**100.0**
CHISD[3]	66.3	90.5
SANP[10]	68.4	93.6
COV + PLS[27]	70.1	**100.0**
MA[7]	74.6	99.0
MSSRC[22]	80.8	-
Proposed-I(Same-Database)	**81.9**	97.4
Proposed-II(Cross-Database)	78.6	97.4

5 Experiments

5.1 Facial Feature Localization

We trained our facial feature detector and evaluated its performance on a subset of the Annotated Facial Landmarks in the Wild (AFLW) database [16]. The database contains about 25,000 face images downloaded from Flickr, each manually annotated with up to 21 fiducial points. There are 5872 and 2000 face images in the selected training set and the test set, respectively. They are mutually exclusive. We cropped the face region using the response of a Viola-Jones face detector and normalize it to 60×60. The training data were partitioned into groups according to pose. Although filter responses were computed for M mixture components at test time, we trained $\frac{M-1}{2}$ of them by utilizing the symmetric property of a human face and mirroring the left-posed face images. Within each group of data, we collected statistics of L to determine the configuration space $\mathcal{Z}$. The reference algorithms used for comparison were the DPM-based one proposed in [32] and the one based on the Haar feature + Gaussian mixture tree [24]. The localization error was measured by the average distance (in pixels) between the predicted fiducial points and the ground truth ones, and normalized by inter-ocular distance. As shown in Figure 2, the proposed facial feature localization algorithm outperforms the two reference algorithms. However, the DPM detector is able to provide face detection output that is not supported by our method. It has also been observed that for large poses, the advantage of the proposed approach in localization accuracy is more evident.

5.2 Video-Based Face Recognition

Youtube Celebrity Video Database: This database [14] has been widely adopted for evaluating the video-based face recognition algorithms. The database contains 1910 Youtube video clips of 47 subjects. Most of the videos were extracted from news TV or movies, and hence exhibit large pose and illumination variations. The low resolution of the videos also poses a challenge to face recognition. In other words, this database aims to test the performance of VFR algorithms under uncontrolled settings. We follow the protocol in [7, 22, 25], i.e., randomly choosing 3 clips per subject as galleries and 6 per subject as probes.

Honda UCSD Database: This database [18] consists of 59 videos of 20 subjects. The videos are divided into a training set which contains one video per subject and a testing set

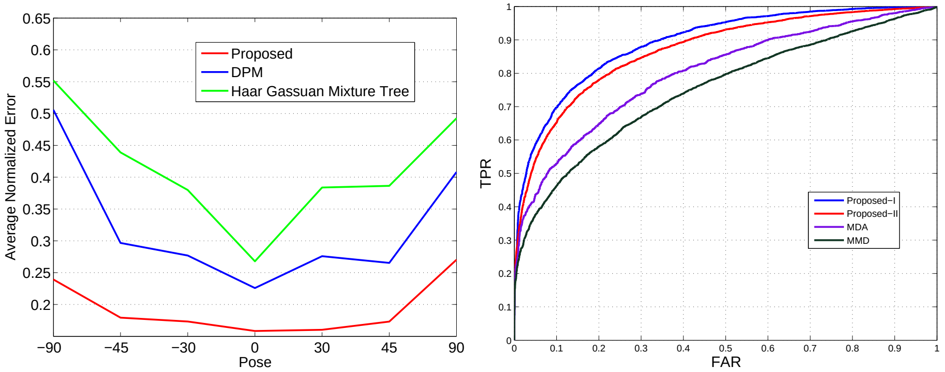

Figure 2: Face fiducial point detection results on the AFLW database (left) and the face verification results on the Youtube Celebrity Video database (right).

Table 2: Comparison of Video-Based Face Recognition Results on the Buffy database

Method	**Buffy Database**
LDML[6]	85.9
MSSRC[22]	86.3
Proposed-I(Same-Database)	**88.3**
Proposed-II(Cross-Database)	85.2

which contains 1 to 4 videos per subject. Each video sequence is recorded in an indoor environment at 15 frames/second and lasts at least 15 frames. Faces in the database undergo significant head motions and expression variations.

Buffy Database: This dataset consists of 639 face tracks from the TV series “Buffy the Vampire Slayer”. We removed the face tracks whose id is labeled as unknown characters, leaving a subset of 483 face tracks for 8 main characters. Following [6], they are separated into a training set of size 227 and a test set of size 256.

For the Youtube Celebrity Video database and the Honda UCSD database, we applied the tracking-by-detection method as described in Section 3 to localize the face. For the Buffy dataset, we used the face tracks provided by the ground truth directly. We then simultaneously detected facial fiducial points and estimated the face pose using the proposed structural-SVM detector. The result was then employed to align the face region to a canonical frame pre-specified for the corresponding pose. We calculated the self-quotient image to normalize the illumination. Pose-specific masks were imposed to suppress the background pixels. LBP and TP-LBP features were extracted and concatenated to form the feature vector. PCA was applied to reduce the dimension of feature vector to 400.

We trained our shared dictionary under two different settings. In the first one, the dictionary was learned from each database’s own training set. We call this the same-database dictionary mode. Alternatively, because the intra-personal/extra-personal face variations are generic, we can learn a dictionary using training data of an entirely different set of subjects. We call this second case the cross-database dictionary mode. The number of intra-personal or extra-personal feature vector pairs that can be used for training is in $O(NK^2)$ and $O(N^2K^2)$ respectively, where N is the number of subjects and K is the average number of clusters discovered by the Dirichlet process Gaussian mixture model from the videos of the same subject. The potential number is huge for a large database like the Youtube

Celebrity Video dataset, especially when we are concerned with the extra-personal pairs. This is also true if we are to learn a dictionary from an external database. On the other hand, the number of intra-personal pairs generated from a small training set, such as that of the Honda/UCSD database, might be insufficient for learning a dictionary. In the former case, we pruned candidate pairs by keeping only around 4000 samples in each of the intra-personal and extra-personal training set. We attempted to distribute the samples as evenly as possible and avoided only using samples from a small subset of videos. In the latter case, we augmented the pool of intra-personal pairs with samples from external data. To train the dictionary in the cross-database mode, we used the LFW database[11] which has 5749 people, among which 1680 subjects have two or more images. We expect that the different variations covered by the database can lead to a dictionary with good generalization property.

We compare the proposed methods with several existing VFR algorithms on the three databases. It is apparent from a careful study of reported experimental results that not all the algorithms are compared on all the databases. On the Youtube Celebrity Video database and the Honda/UCSD database, the compared existing algorithms include: Mutual Subspace Method (MSM)[30], Manifold-Manifold Distance (MMD)[26], Manifold Discriminant Analysis (MDA)[25], Convex Hull based Image Set Distance (CHISD)[3], Sparse Approximated Nearest Point (SANP)[10], Covariance Partial Least Square (COV + PLS)[27], Manifold Alignment (MA)[7] and Mean Sequence Sparse Representation-based Classification (MSSRC) [22]. On the Buffy database, we compare with Logistic Discriminant-based Metric Learning (LDML) [6] and MSSRC. The results are presented in Tables 1 and 2. As shown in the tables, in all three databases, both of the same-database and the cross-database dictionary modes of the proposed algorithm achieve comparable results w. r. t. the state-of-the-art. On the most challenging Youtube Celebrity Video database, our method produces slightly better results than the one most recently reported in [22] and outperforms the other algorithms by a large margin. The relative lower classification rate on the Honda/UCSD database may be due to insufficient training samples. A noticeable fact is that using the cross-database dictionary learned from the external database usually leads to a degraded performance. This is consistent with our intuition that cross-domain learning is in general a more difficult problem. But the cross-domain dictionary is advantageous in terms of scalability and flexibility, as the training difference vectors are complementary to each other and can be shared. Finally, the proposed framework naturally supports the face verification protocol. Therefore, we also investigate the performance of our algorithm in the verification mode that is described in Section 4. The result on the Youtube Celebrity Video database is plotted in the form of ROC curves in Figure 2. We compare with the MMD and MDA because their outputs are distances, from which the ROC curves can be conveniently generated.

6 Conclusion

We introduced a novel framework for video-based face recognition. It is based on the generic concept of intra-personal/extra-personal variations, and hence leads to greater scalability. We exploited the strengths of sparse coding in classification and learned a discriminative dictionary from these variations. In addition, we presented a facial feature detection method for accurate face alignment in unconstrained videos. Our scheme is flexible enough to work in both identification and verification modes. It can also be trained and tested on different databases. We conducted experiments on three public databases and demonstrated the performance of the proposed approach through comparison with existing algorithm.

References

[1] O. Arandjelovic and R. Cipolla. Face recognition from face motion manifolds using robust kernel resistor-average distance. In *Proceedings of the 2004 Conference on Computer Vision and Pattern Recognition Workshop*, volume 5, pages 88–93, June 2004.

[2] P. N. Belhumeur, D. W. Jacobs, D. Kriegman, and N. Kumar. Localizing parts of faces using a consensus of exemplars. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 545–552, June 2011.

[3] H. Cevikalp and B. Triggs. Face recognition based on image sets. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2567–2573, June 2010.

[4] S. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM Review*, 43(1):129–159, January 2001.

[5] Y.-C. Chen, V. M. Patel, and R. Chellappa. Dictionary-based face recognition from video. In *European Conference on Computer Vision*, pages 766–779, October 2012.

[6] R. G. Cinbis, J. Verbeek, and C. Schmid. Unsupervised metric learning for face identification in TV video. In *IEEE International Conference on Computer Vision*, pages 1559–1566, November 2011.

[7] Z. Cui, S. Shan, H. Zhang, S. Lao, and X. Chen. Image sets alignment for video-based face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2626–2633, June 2012.

[8] M. Everingham, J. Sivic, and A. Zisserman. "Hello! my name is... Buffy" – automatic naming of characters in TV video. In *British Machine Vision Conference*, volume 3, pages 899–908, September 2006.

[9] M. Guillaumin, J. Verbeek, and C. Schmid. Is that you? Metric learning approaches for face identification. In *IEEE International Conference on Computer Vision*, pages 498–505, September 2009.

[10] Y. Hu, A. S. Mian, and R. Owens. Sparse approximated nearest points for image set classification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 121–128, June 2011.

[11] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: a database for dtudying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007.

[12] A.D. Jepson, D.J. Fleet, and El-Maraghi. T.F. Robust online appearance model for visual tracking. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1:415–422, 2001.

[13] Z. Jiang, Z. Lin, and L. S. Davis. Label consistent K-SVD: learning a discriminative dictionary for recognition. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 35(11):2651–2664, November 2013.

[14] M.Y. Kim, S. Kumar, V. Pavlovic, and H.A. Rowley. Face tracking and recognition with visual constraints in real-world videos. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, June 2008.

[15] T. K. Kim, J. Kittler, and R. Cipolla. Discriminative learning and recognition of image set classes using canonical correlations. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 29(6):1005–1018, June 2007.

[16] M. Koestinger, P. Wohlhart, P. M. Roth, and H. Bischof. Annotated facial landmarks in the wild: a large-scale, real-world database for facial landmark localization. In *IEEE International Workshop on Benchmarking Facial Image Analysis Technologies*, pages 2144–2151, November 2011.

[17] K. Kurihara, M. Welling, and Teh Y. W. Collapsed variational Dirichlet process mixture models. In *International Joint Conference on Artificial Intelligence*, pages 2796–2801, January 2007.

[18] K. C. Lee, J. Ho, M. H. Yang, and D. Kriegman. Video-based face recognition using probabilistic appearance manifolds. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages 313–320, June 2003.

[19] X. Liu and T. Chen. Video-based face recognition using adaptive hidden Markov models. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, 2003.

[20] B. Moghaddam. Principal manifolds and probabilistic subspaces for visual recognition. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 24(6):780–788, 2002.

[21] K. P. Murphy. *Machine Learning: A Probabilistic Perspective*. The MIT Press, 2012.

[22] E. G. Ortiz, A. Wright, and M. Shah. Face recognition in movie trailers via mean sequence sparse representation-based classification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3531–3538, June 2013.

[23] R. Rubinstein, A. M. Bruckstein, and M. Elad. Dictionaries for sparse representation modeling. *Proceedings of the IEEE*, 98(6):1045–1057, June 2010.

[24] J. Sivic, M. Everingham, and A. Zisserman. "Who are you?" – learning person specific classifiers from video. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1145–1152, June 2009.

[25] R. Wang and X. Chen. Manifold discriminant analysis. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 429–436, June 2009.

[26] R. Wang, S. Shan, X. Chen, and G. Wen. Manifold-manifold distance with application to face recognition based on image set. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, June 2008.

[27] R. Wang, H. Guo, L. Davis, and Q. Dai. Covariance discriminative learning: a natural and efficient approach to image set classification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2496–2503, June 2012.

[28] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma. Robust face recognition via sparse representation. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 31 (2):210–227, 2009.

[29] J. Wu and J. M. Rehg. CENTRIST: A visual descriptor for scene categorization. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 33(8):1489–1501, 2011.

[30] O. Yamaguchi, K. Fukui, and K. Maeda. Face recognition using temporal image sequence. In *IEEE International Conference on Automatic Face and Gesture Recognition*, pages 318–323, April 1998.

[31] M. Yang, D. Zhang, X. Feng, and D. Zhang. Fisher discrimination dictionary learning for sparse representation. In *IEEE International Conference on Computer Vision*, pages 543–550, Nov 2011.

[32] X. Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2879–2886, June 2012.