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 and Center for Automation Research, UMIACS
University of Maryland
College Park, USA

In recent years, with videos playing an increasingly important role in our everyday lives, video-based face recognition (VFR) has begun to attract considerable research interest. In this paper, we attempt to improve the performance of VFR based on the concept of intra-personal/extra-personal face variations. The concept was first proposed by Moghaddam et al. in [2] and has achieved great success in still-image based face recognition. Specifically, the intrapersonal subspace Ω_{In} is defined as the subspace constructed from within-class sample differences $\{\Delta_{In}\}$. It accounts for appearance variations of the same subject that arise from factors like pose, lighting, expression etc. Similarly, the extra-personal subspace Ω_{Ex}, which characterizes appearance variations caused by intrinsic identity differences, is constructed using the between-class sample differences $\{\Delta_{Ex}\}$. To apply this concept to the VFR problem, our solution is based on two aspects: To handle pose variations, we learn a Structural-SVM-based detector that simultaneously localizes the face fiducial points and estimates face pose. To model other face variations, we exploit the strengths of sparse codings by constructing intra-personal/extra-personal dictionaries. An overview of the proposed approach is shown in Figure 1.

For face normalization, we learn a mixture of fiducial point detectors which is used for geometric alignment. Each component of the mixture corresponds to a specific face pose. We localize the fiducial points L and estimate the face pose m jointly 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:

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

. In (1), $(I_n,\mathbf{z}_n)$ is an image-label pair in the training database and $\mathcal{Z}$ is the viable label configuration set. ξ_n is the slack variable . $\Delta(\mathbf{z},\mathbf{z}_n)$ is the loss function of the 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 $S(m)$. Compared with Zhu and Ramanan's recent Deformable Parts Model (DPM)-based face and feature detector [3], our objective function explicitly impose constraints on the margin between correct and wrong landmark predictions. Moreover, in our case the margin is re-scaled by a loss function $\Delta(\mathbf{z},\mathbf{z}_n)$ which penalizes the negative training samples according to their misalignment errors. As a result, although our method is not designed to produce face detection output in addition to feature point locations, it has higher accuracy in localizing fiducial points.

Based on the estimated pose, the localized faces in a video are then aligned to pose-specific common reference coordinate frames. They are further clustered using a non-parametric Bayesian model to remove temporal redundancy. The resulting model has infinite number of Gaussian mixtures controlled by a Dirichlet process $DP(\beta,H)$ [1], where β is the concentration parameter and H is the base probability measure. The mixture weights are generated from the Griffiths-Engen-McClosky (GEM) process. By using the Dirichlet process mixture model, new clusters can be generated when more frames are observed, and there is no need to know the number of clusters a priori.

In recent years, sparse coding has gained popularity in the field of image classification. In general, a dictionary $\mathbf{D} = [D_1, D_2, ..., D_K]$, where $D_k \in R^d$, can be learned unsupervisedly from training samples $\mathbf{X} = \{\mathbf{x}_i, i = 1,2,...,N\} \in R^{d\times N}$ (In our case, the training samples are intra/extra-personal difference of feature vectors which are extracted from faces of the same pose.) by solving the following constrained optimization problem:

$$\min_{\mathbf{D},\alpha} \sum_{i=1}^{N} \frac{1}{2}\|\mathbf{x}_i - \mathbf{D}\alpha_i\|_2^2 + \lambda\|\alpha_i\|_1 \quad (2)$$

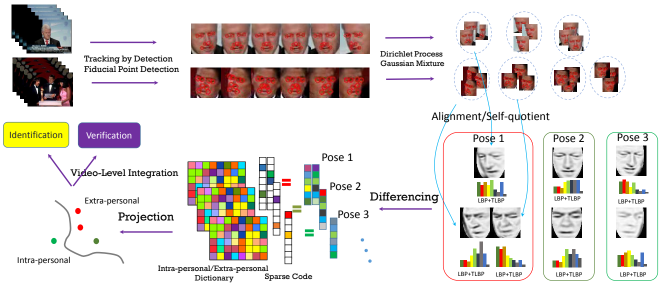

Figure 1: Processing pipeline of the proposed video-based face recognition algorithm.

However, to serve the purpose of classification better, we follow the Label-Consistent K-SVD (LC-KSVD) algorithm to jointly learn a generative shared dictionary and a discriminative projection matrix. Although the shared dictionary is composed of two sub-dictionaries corresponding to intrapersonal 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. As a result, the final optimization problem has the following form:

$$\min_{\mathbf{D},\mathbf{A}} \|\mathbf{X} - \mathbf{DA}\|_2^2 + \mu\|\mathbf{Q} - \mathbf{BA}\|_2^2 + \sigma\|\mathbf{F} - \mathbf{WA}\|_2^2 + \lambda\sum_i\|\alpha_i\|_1 \quad (3)$$

. In (3), 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. The matrix $\mathbf{W} \in R^{2\times d}$ encodes the discriminative information of the sparse codes $\mathbf{A}$ and is learned along with the shared dictionary. The linear transformation $\mathbf{B} \in R^{K\times d}$ encourages the samples from the same class to be reconstructed using similar atoms. 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. At test time, for each probe video, we extract feature vectors from the centers of clusters formed using the non-parametric Bayesian method introduced above, and take differences between them and the feature vectors similarly extracted from clusters in the gallery videos. Recognition results are then obtained using the learned intra/extra-personal dictionaries and the discriminant matrix $\mathbf{W}$.

One advantage of the proposed algorithm is its scalability. Traditionally, it requires a large amount of training data to effectively characterize a subject. More often than not, we have 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. In contrast, because the intra/extra-personal face variations are generic, our algorithm is flexible enough to learn a dictionary using either the training set from the same database or that of an entirely different set of subjects (i.e. cross-database dictionary). We demonstrate through experiments that the proposed approach achieved state-of-arts performance in both modes. Moreover, the proposed framework naturally supports the face verification protocol in addition to the recognition one.

[1] 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.

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

[3] 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.