Probabilistic Matching of Image Sets for Video-Based Face Recognition

Probabilistic Matching of Image Sets for Video-Based Face Recognition

Moh Edi Wibowo, Dian Tjondronegoro, Vinod Chandran
Faculty of Science and Engineering
Queensland University of Technology
Brisbane, Australia
moh.wibowo@student.qut.edu.au, dian@qut.edu.au, v.chandran@qut.edu.au

Abstract—**We address the problem of face recognition on video by employing the recently proposed probabilistic linear discriminant analysis (PLDA). The PLDA has been shown to be robust against pose and expression in image-based face recognition. In this research, the method is extended and applied to video where image set to image set matching is performed. We investigate two approaches of computing similarities between image sets using the PLDA: the closest pair approach and the holistic sets approach. To better model face appearances in video, we also propose the heteroscedastic version of the PLDA which learns the within-class covariance of each individual separately. Our experiments on the VidTIMIT and Honda datasets show that the combination of the heteroscedastic PLDA and the closest pair approach achieves the best performance.**

Keywords-video-based face recognition; image set matching; heteroscedastic probabilistic linear discriminant analysis

I. INTRODUCTION

Automatic face recognition has long been studied as part of the efforts toward adopting human capabilities to computers. Recognition on still images has been the main focus of the study. It has achieved an impressive progress to an extent that some automatic systems can outperform humans in recognizing unfamiliar faces of frontal pose across changes of illumination [1]. Nevertheless, under generic situations where the pose, illumination, occlusion, expression, time delay, and resolution may vary together, fully automatic systems are still far less robust than humans.

More recently, researchers have started to use video in face recognition tasks. Video inherently contains more information than images. It is now also widely available following the rapid advancement of multimedia technologies. There are two advantages of using video in face recognition: multiple observations and temporal continuity [2]. Video contains many frames each of which can be considered as one observation. With many observations from such frames, improvements might be obtained in the recognition stage. Techniques such as score fusion [3] and image set matching [4] have proven to be useful in disambiguating the decision choices. The availability of multiple observations may also mitigate the effects of non-optimal viewing conditions as well as inaccurate localization and feature extraction. Video frames are subject to temporal continuity meaning that object positions and appearances within adjacent frames do not differ much. This property has enabled the use of tracking, temporal features, dynamic models and recovery of 3D shapes.

In this paper, we employ video as image sets and apply the probabilistic linear discriminant analysis (PLDA) [5] through the derivation of some set to set similarity measures. The PLDA has demonstrated good performance and robustness to pose [6] and expression in still image recognition. Within its framework, similarity between two data points is computed as the likelihood of an underlying data generation model. Interestingly, this model can be straightforwardly followed to formulate similarities (matching scores) between sets. We investigate two approaches of computing such similarities: the closest pair approach and the holistic sets approach. The objective is to find out which approach is better than the other. We also investigate whether constructing individual-specific within-class covariances for the PLDA (heteroscedastic model) improves performance or not. This is motivated by the fact that unlike in the traditional PLDA, there are many examples for each individual provided by the video. To facilitate a fully automatic recognition system, we also develop a front-end based on multi-view ASMs/AAMs which serves to automatically detect face regions, localizing feature points, registering the faces, and identifying the poses. Ideally, we would consider different kinds of variations such as of pose, illumination, resolution, occlusion, expression, etc. However, we limit our discussion to pose variation since it is the most commonly encountered one in video.

The rest of the paper is organized as follows. Section 2 summarizes some literatures related to this work. Section 3 briefly describes the PLDA models. Section 4 presents the proposed system and some of its important parts. Evaluation of the system is discussed in Section 5. Conclusion is outlined in section 6.

II. RELATED WORKS

In this section, we focus our discussion on recognition methods which treat video as image sets with temporal aspects ignored. These methods utilize different representations of sets and different similarity measures between them. Yamaguchi et al. [7] propose linear subspaces as the representations and

measure the similarities using canonical angles between subspaces. Following this work, Tat-Jun et al. [4] apply an incremental SVD to compute linear subspaces on-line and use chordal distances as the distance metrics. Subspace-to-subspace distances have also been analyzed in the framework of Grassmannian manifolds [8] i.e. manifolds where the data points are subspaces. Another simple approximation of image set is the affine/convex hull recently proposed by Cevikalp and Triggs [9]. This affine hull approximates the region occupied by images of an individual in the input space. Recognition is performed by finding the "nearest-point distance" between the hulls which is equal to synthesizing the closest pair of examples. Lately, Hu et al. [10] improve the computation of between-hull distances by enforcing sparseness to the coefficients of the affine combinations. Subspace or affine hull representations normally do not attempt to model nonlinearity of the face manifolds. This notion of manifold states that appearances of a person's face form a highly nonlinear surface yet continuous and smooth with intrinsically lower dimension than the input space. We believe that modelling this nonlinearity might improve recognition.

To approximate face manifolds, image clustering and piece-wise linear models have been widely adopted. Hadid and Pietikainen [11] choose an exemplar from each cluster (which is the centre of the cluster) after applying the locally linear embedding, K-means, and self organizing map (SOM) to video frames. These exemplars are used to train PCA/LDA classifiers which perform majority/probabilistic voting to classify the probe video. Krueger and Zhou [12] select representative exemplars from training video using an on-line version of radial basis function. These exemplars facilitate the computation of observation likelihoods in the proposed simultaneous tracking and recognition framework. Lee et al. [13] apply K-means clustering to video frames and treat the obtained clusters as components of the manifold. Similarly, hierarchical clustering has been employed by Fan and Yeung [14] to discover such local structures. Components of the manifolds are eventually represented as linear subspaces. Lee et al. [13] compute L2–Hausdorff distances between video frames and face manifolds probabilistically in the recognition stage. This recognition method though seems to rely only on the last frame while the previous frames are more for pose-tracking purpose. Fan and Yeung [14] measure manifold similarities by computing canonical angles distances between manifold components and apply majority voting.

Aside from distance-based clustering, several approaches propose to use "semantic clustering". Arandjelovic and Cipolla [15] define three pose clusters on video frames (frontal face, face left, and face right) whose illumination is normalized using gamma intensity correction (GIC). Matching scores between manifolds are computed based on cluster centre distances and a Bayesian likelihood fusion. Li et al. [16] construct an identity surface of an individual by partitioning the pose space and construct a plane in each partition using face images of the partition. The dissimilarity of a novel video is obtained by projecting the frames to the surface and summing up the projection distances. Our proposed approach is somehow similar to [15] in the sense that we construct semantic clusters based on pose. In our method however, these clusters are matched using the PLDA which considers not only the cluster centres but all images within them (the distributions). We believe that comparison between semantic manifold components is more natural and meaningful than comparison between ones found by distance-based clustering. The cost of this approach is high level information such as poses of the faces needs to be provided or estimated beforehand.

III. PROBABILISTIC LINEAR DISCRIMINANT ANALYSIS

A. *Face Representation*

Within this work, faces of an individual obtained from a video sequence are represented as a group of image/feature sets. Every set/subgroup corresponds to a particular "discrete" pose such as frontal or left-profile, etc. Each member of the subgroup is actually a segmented and registered face of the individual or some features derived from it. We represent these members as vectors and restrict them to have equal lengths. Figure 1 illustrates the group $\mathbf{X}_i$ of an individual i which consists of K subgroups $\mathbf{X}_{i1}$, $\mathbf{X}_{i2}$, … , $\mathbf{X}_{iK}$ corresponding to K poses. All vectors within the subgroups are of $D\times1$ dimension.

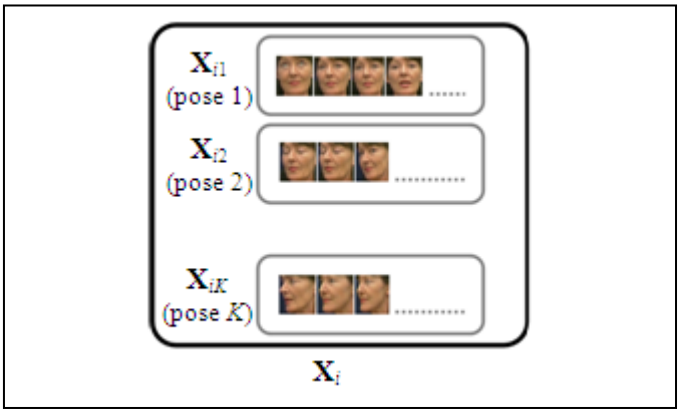


Figure 1. A representation $\mathbf{X}_i$ of a video sequence containing faces of an individual ***i*** spreading over K poses.

B. *PLDA Models*

If we denote the j-th observation (face image, feature vector) of the i-th individual (with no pose variation) as $\mathbf{x}_{ij}$, the data generation process of the PLDA [5] can be expressed as

$$\mathbf{x}_{ij} = \mu + \mathbf{F}\mathbf{h}_i + \mathbf{G}\mathbf{w}_{ij} + \varepsilon_{ij} \,. \qquad (1)$$

Each observed data point $\mathbf{x}_{ij}$ is assumed to be generated from $\mathbf{h}_i$ and $\mathbf{w}_{ij}$ which are points in latent spaces. We call the space of $\mathbf{h}_i$ as the between-individual space and the space of $\mathbf{w}_{ij}$ as the within-individual space. As indicated by the subscripts, observations from the same individual share the same value of $\mathbf{h}$ but have their own values of $\mathbf{w}$. Hence, the term $\mathbf{h}$ is also called the latent identity variable (LIV) since it is unique for each individual. The vectors $\mathbf{h}_i$ and $\mathbf{w}_{ij}$ should have smaller

lengths than the vector $\mathbf{x}_{ij}$. They are mapped to the observation space via linear transformations **F** and **G** respectively and the addition of the observation mean μ and the residual noise ε_{ij}. Note that **x**, **h**, **w**, and ε are random variables with multivariate Gaussian distributions and (1) can be described in terms of conditional probabilities:

$$P(\mathbf{x}_{ij}|\mathbf{h}_i, \mathbf{w}_{ij}, \theta) = g_{\mathbf{x}}[\mu + \mathbf{F}\mathbf{h}_i + \mathbf{G}\mathbf{w}_{ij}, \Sigma] , \quad (2)$$

$$P(\mathbf{h}_i) = g_{\mathbf{h}}[\mathbf{0}, \mathbf{I}] , \quad (3)$$

$$P(\mathbf{w}_{ij}) = g_{\mathbf{w}}[\mathbf{0}, \mathbf{I}] . \quad (4)$$

Here, $\theta = (\mu, \mathbf{F}, \mathbf{G}, \Sigma)$ is the model parameters and Σ is the diagonal covariance matrix of the residual noise ε.

There are two phases in using the above models. The first one is the *training* (*offline*) *phase* where we learn the parameters θ using the training data $\mathbf{x}_{ij}$. The second one is the *recognition* (*online*) *phase* where we use the trained models to infer the identities of the probe data. Prince et al. [5] have developed an EM algorithm for the training of the PLDA models. They also propose a Bayesian model comparison approach to compute matching scores between gallery and probe images. If there are *M* individuals in the gallery each of whom has one example image, the matching score between a gallery image $\mathbf{x}_m$ and a probe image $\mathbf{x}_p$ is defined as

$$S(\mathbf{x}_m, \mathbf{x}_p) = P(\mathbf{x}_{1 \ldots M}, \mathbf{x}_p|\mathbf{M}_m) = P(\mathbf{x}_m, \mathbf{x}_p) \prod_{i=1 \ldots M,\ i \neq m} P(\mathbf{x}_i), \quad (5)$$

$$P(\mathbf{x}_m, \mathbf{x}_p) = \iiint P(\mathbf{x}_m, \mathbf{x}_p, \mathbf{h}_m, \mathbf{w}_m, \mathbf{w}_p)\, d\mathbf{h}_m d\mathbf{w}_m d\mathbf{w}_p , \quad (6)$$

$$P(\mathbf{x}_i) = \iint P(\mathbf{x}_i, \mathbf{h}_i, \mathbf{w}_i)\, d\mathbf{h}_i d\mathbf{w}_i . \quad (7)$$

$\mathbf{M}_m$ represents the situation that $\mathbf{x}_m$ and $\mathbf{x}_p$ are generated from the same LIV while the other gallery images are generated from their own LIVs. $P(\mathbf{x}_m, \mathbf{x}_p)$ and $P(\mathbf{x}_i)$ are the likelihoods of such generations. The matching score in (5) can be simplified into

$$S(\mathbf{x}_m, \mathbf{x}_p) = P(\mathbf{x}_m, \mathbf{x}_p)(P(\mathbf{x}_m))^{-1}. \quad (8)$$

To evaluate $P(\mathbf{x}_m, \mathbf{x}_p)$ and $P(\mathbf{x}_i)$, we can rewrite the generative equations as

$$\begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \end{bmatrix} = \begin{bmatrix} \mu \\ \mu \\ \vdots \end{bmatrix} + \begin{bmatrix} \mathbf{F} & \mathbf{G} & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{F} & \mathbf{0} & \mathbf{G} & \ldots & \mathbf{0} \\ \vdots & \vdots & \vdots & & \vdots \\ \mathbf{F} & \mathbf{0} & \mathbf{0} & \ldots & \mathbf{G} \end{bmatrix} \begin{bmatrix} \mathbf{h} \\ \mathbf{w}_1 \\ \mathbf{w}_2 \\ \vdots \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \end{bmatrix} , \quad (9)$$

$$\mathbf{x}' = \mu' + \mathbf{A}\mathbf{y} + \varepsilon', \quad (10)$$

and compute $g_{\mathbf{x}'}[\mu', \mathbf{A}\mathbf{A}^T + \Sigma']$ where Σ' is the diagonal covariance matrix of ε'. Note that (9) and (10) have generic forms which can be used to obtain $P(\mathbf{x}_m, \mathbf{x}_p)$ as well as $P(\mathbf{x}_i)$ Equations (5) – (10) also naturally generalize to the case of image set matching $S(\mathbf{X}_m, \mathbf{X}_p)$. The inferred identity can then be obtained as $\mathbf{argmax}_{i=1 \ldots M}\, S(\mathbf{x}_i, \mathbf{x}_p)$ or $\mathbf{argmax}_{i=1 \ldots M}\, S(\mathbf{X}_i, \mathbf{X}_p)$.

IV. THE PROPOSED SYSTEM

A. Recognition System Framework

The framework of the proposed system is shown in Figure 2. It has three processing modules: front-end, learning module, and matching module. The front-end serves to localize face regions in video frames, extract the features, identify the poses, and group the features based on the poses into several subgroups. By this front-end, a video sequence will be "transformed" into its face representation. The learning module builds a number of PLDA models from training data in the offline stage. These trained models will be used by the matching module in the online stage to compute matching scores between groups of feature sets of different individuals. Note that the front-end is utilized in both the offline and online stages.

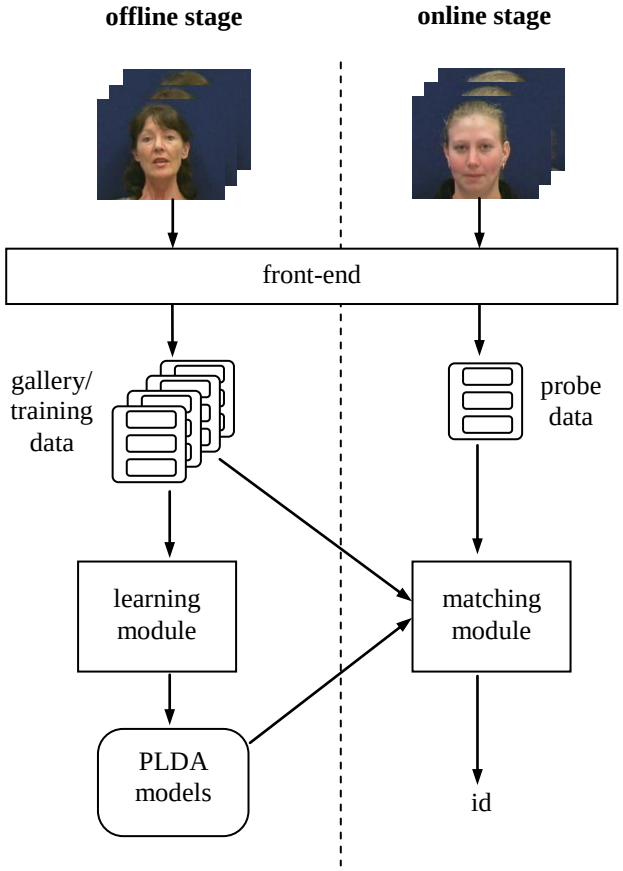


Figure 2. The proposed face recognition system.

B. Front-End

The first tasks to be done by the front-end are localizing faces and identifying the poses. To localize faces in video, the most straightforward approach is applying a face detector to every frame. Another approach is detecting the faces followed by tracking or tracking the faces with some guidance from face detectors in every frame. To estimate head pose, various methods have been proposed including the appearance-based matching, manifold embedding, nonlinear regression, detection by array, deformable model fitting, geometric estimation, etc

[17]. Nevertheless, no one prominent solution has been devised for different kinds of situations.

To partially accomplish the above two tasks, we develop a simple method employing Viola-Jones face detectors and multi-view ASMs/AAMs. Five "discrete" poses are defined for the system: frontal, left/right half profile, and left/right profile. We construct five ASMs/AAMs which share a number of feature points and correspond to those five poses (Figure 3(a)). The method then comprises the following steps

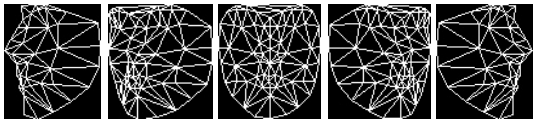
(i) Detect a face from a video frame using a frontal/profile Viola-Jones face detector.

(ii) Fit the five ASMs (frontal, left/right half profile, and left/right profile) and choose one which best fits the detected face as the valid model.

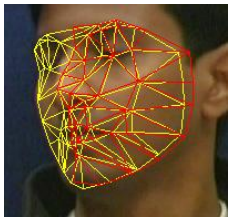
(iii) If the valid model is profile, check also the fitting result of the adjacent half profile model. If the profile model has its nose tip inside the area (with a margin ω) of the half profile model (Figure 3(b)), take the half profile model as the valid one instead.

(iv) Deduce the face's pose as the pose of the valid ASM and use the obtained feature points to perform registration.

To choose the best fitting ASM, we need a measure to evaluate the fitting quality. One possible solution is using the proportion of outlier pixels in the fitted region [18]. A pixel $\mathbf{p}$ is considered as an outlier if the reconstruction error $|\boldsymbol{I}(\mathbf{p}) - \boldsymbol{A}(\mathbf{p})|$ is greater than a predefined threshold σ. Here, $\boldsymbol{I}(\mathbf{p})$ and $\boldsymbol{A}(\mathbf{p})$ is the pixel value obtained from the input image and the AAM-synthesized texture respectively. The best fitting ASM/AAM is the one which produces the smallest proportion of outlier pixels.



(a)



(b)

Figure 3. (a) Five ASM/AAMs employed by the front-end; (b) fitting results of the profile and half profile models.

The proposed method works well for the database used in the experiments. The detectors successfully find frontal to profile faces while at the same time ignore faces of other "difficult" poses. One concern regarding the method is it is unable to reject poor fitting results. Therefore, the method assumes that at least one model fits well to the input image. The detected faces are registered through a rotation, scaling, and cropping based on the centre point of the mouth and the middle point between the two eyes. The last tasks performed by the front-end are extracting features from the cropped faces and bundle them into a group of feature sets. The group only contains three feature sets corresponding to the frontal pose, right half profile, and right profile. Faces of the left half profile and left profile are flipped horizontally.

C. *Learning Module*

The learning module learns some PLDA models from the collected training data. In addition to the original PLDA, we also develop the heteroscedastic PLDA. With this new model, each individual has its own within-class covariance and the model parameters change into $\theta = (\mu, \mathbf{F}, \mathbf{G}_{1 \ldots M}, \Sigma)$ where M is the number of individuals. We derive the training algorithm of this model in a similar way as the original PLDA and come up with the following procedure

(i) **Compute** $\mu = (1/N)\Sigma_{ij}\mathbf{x}_{ij}$ and **initialize** $\mathbf{F}$, $\mathbf{G}_{1 \ldots M}$, Σ.

(ii) **Repeat until** converged:

E-step: Compute $E(\mathbf{z}_{ij}|\mathbf{x}_{i\bullet})$ and $E(\mathbf{z}_{ij}\mathbf{z}_{ij}^T|\mathbf{x}_{i\bullet})$ for each data point $\mathbf{x}_{ij}$, given the current θ.

M-step: Compute the new $\mathbf{F}$, $\mathbf{G}_{1 \ldots M}$, and Σ using the values obtained from the E-step.

N is the number of data points and $\mathbf{z}_{ij} = [\mathbf{h}_i^T \mathbf{w}_{ij}^T]^T$. Since data points of the same identity are generated from the same $\mathbf{h}$, the expectation values of the E-step are computed simultaneously for $\mathbf{x}_{i\bullet}$ i.e. all data points of a particular individual i. Following the approach in [5], we first arrange the generative equations of these points as in (9). This time, however, we replace $\mathbf{G}$ in (9) with $\mathbf{G}_i$ if the person being considered is person i. From the resulted equations we then calculate

$$E(\mathbf{y}|\mathbf{x}_{i\bullet}) = (\mathbf{I} + \mathbf{A}^T\Sigma'^{-1}\mathbf{A})^{-1}\mathbf{A}^T\Sigma'^{-1}(\mathbf{x}' - \mu'), \quad (11)$$

$$E(\mathbf{y}\mathbf{y}^T|\mathbf{x}_{i\bullet}) = (\mathbf{I} + \mathbf{A}^T\Sigma'^{-1}\mathbf{A})^{-1} + E(\mathbf{y}|\mathbf{x}_{i\bullet})E(\mathbf{y}|\mathbf{x}_{i\bullet})^T, \quad (12)$$

where $\mathbf{y}$, $\mathbf{x}'$, μ', $\mathbf{A}$, and Σ' are the terms of (10). $E(\mathbf{z}_{ij}|\mathbf{x}_{i\bullet})$ and $E(\mathbf{z}_{ij}\mathbf{z}_{ij}^T|\mathbf{x}_{i\bullet})$ can be extracted from $E(\mathbf{y}|\mathbf{x}_{i\bullet})$ and $E(\mathbf{y}\mathbf{y}^T|\mathbf{x}_{i\bullet})$ respectively.

To update $\mathbf{F}$, $\mathbf{G}_{1 \ldots M}$, and Σ, the **M**-step of the heteroscedastic models executes the following update rules

$$\mathbf{F}_{new} = \Sigma_{ij}\{(\mathbf{x}_{ij} - \mu)E(\mathbf{h}_i|\mathbf{x}_{i\bullet})^T - \mathbf{G}_i\, E(\mathbf{w}_{ij}\mathbf{h}_i^T|\mathbf{x}_{i\bullet})\}\times (\Sigma_{ij}E(\mathbf{h}_i\mathbf{h}_i^T|\mathbf{x}_{i\bullet}))^{-1}, \quad (13)$$

$$\mathbf{G}_{i\ new} = \Sigma_j\{(\mathbf{x}_{ij} - \mu)E(\mathbf{w}_{ij}|\mathbf{x}_{i\bullet})^T - \mathbf{F}_{new}\, E(\mathbf{h}_i\mathbf{w}_{ij}^T|\mathbf{x}_{i\bullet})\}\times (\Sigma_j E(\mathbf{w}_{ij}\mathbf{w}_{ij}^T|\mathbf{x}_{i\bullet}))^{-1}, \quad \text{for } i = 1 \ldots M, \quad (14)$$

$$\Sigma_{new} = (1/N)\mathrm{diag}\{\Sigma_{ij}\{(\mathbf{x}_{ij} - \mu)(\mathbf{x}_{ij} - \mu)^T - [\mathbf{F}_{new}\ \mathbf{G}_{i\ new}]\, E([\mathbf{h}_i^T \mathbf{w}_{ij}^T]^T|\mathbf{x}_{i\bullet})\,(\mathbf{x}_{ij} - \mu)^T\}\}. \quad (15)$$

All expectation terms at the above rules are extracted from $E(\mathbf{z}_{ij}|\mathbf{x}_{i\bullet})$ and $E(\mathbf{z}_{ij}\mathbf{z}_{ij}^T|\mathbf{x}_{i\bullet})$ which are calculated at the E-step.

Take a note that we will construct separate models for different poses. Model for pose k is trained using image sets $\mathbf{X}_{1k}, \mathbf{X}_{2k}, \ldots, \mathbf{X}_{Mk}$.

D. Matching Module

In this module, the probe data are compared to all gallery data and the identity is inferred. In order to match two groups of feature sets $\mathbf{X}_m$ (gallery) and $\mathbf{X}_p$ (probe), matching scores of the corresponding sets $\mathbf{X}_{mk}$ and $\mathbf{X}_{pk}$ are computed first and then fused across all k e.g. using the product rule. We consider two approaches to compute those matching scores. The first one is the closest pair approach which can be described as

$$S(\mathbf{X}_{mk}, \mathbf{X}_{pk}) = \mathbf{max}_{ij}\, S(\mathbf{x}^i_{mk}, \mathbf{x}^j_{pk}) = \mathbf{max}_{ij}\, P(\mathbf{x}^i_{mk}, \mathbf{x}^j_{pk})(P(\mathbf{x}^i_{mk}))^{-1}. \quad (16)$$

The terms $\mathbf{x}^i_{mk}$ and $\mathbf{x}^j_{pk}$ represent a data point in $\mathbf{X}_{mk}$ and $\mathbf{X}_{pk}$ respectively. The second one is the holistic sets approach which can be written as

$$S(\mathbf{X}_{mk}, \mathbf{X}_{pk}) = \{P(\mathbf{X}_{mk}, \mathbf{X}_{pk})(P(\mathbf{X}_{mk}))^{-1}\}^{-1/|\mathbf{X}_{pk}|}. \quad (17)$$

$P(\mathbf{x}^i_{mk}, \mathbf{x}^j_{pk})$, $P(\mathbf{x}^i_{mk})$, $P(\mathbf{X}_{mk}, \mathbf{X}_{pk})$, and $P(\mathbf{X}_{mk})$ are computed according to what explained in Section 3. Both the traditional and the heteroscedastic PLDA can be used to evaluate (16) and (17). In the case of heteroscedastic PLDA, we use $\mathbf{G}_m$ instead of $\mathbf{G}$ in the likelihood evaluation. Note that matching scores in (16) and (17) are insensitive to the number of points in each set. For (17), this has been made possible by the inclusion of the fractional power $-1/|\mathbf{X}_{pk}|$ which resembles the geometric mean. The fused score over multiple poses can be written as $S(\mathbf{X}_m, \mathbf{X}_p) = \Pi_k\, S(\mathbf{X}_{mk}, \mathbf{X}_{pk})$. Other possibilities to make use data across poses are employing tied PLDA [6] or decision level fusions.

V. Experiments

We evaluate the proposed method on the VidTIMIT [19] and Honda [13] datasets. The VidTIMIT dataset contains 43 individuals who are asked to perform extended head rotation sequences starting from the centre to the right, left, back to centre, up, down, and finally return to centre. For each person, there are three such sequences recorded in three sessions with resolution of 512×384. The Honda dataset contains 60 video sequences involving 20 individuals with 640×480 pixel resolution. Each video contains variation of pose and sometimes variation of expression as well as occlusion.

For the VidTIMIT dataset, we conduct the experiments with cross validation. In each test pass, we use video from one of the sessions as the training/gallery data and video from the other sessions are the probe data. For the Honda dataset, 20 videos of different individuals are used as the training/gallery data and the remaining 40s are used as the test data. Each video is processed by the front-end. On every frame, the front-end localizes the face and pre-processes it into a 41×41 registered face. LBP codes are then computed on all pixels and concatenated into a single feature vector. We train several PLDA models i.e. models for frontal, half profile, and profile pose and vary the number of basis vectors in $\mathbf{F}$ and $\mathbf{G}$. During the recognition, we test four similarity measures from the possible combinations of the closest pair vs holistic sets approaches and the PLDA vs heteroscedastic PLDA models.

Figure 4 and 5 show classification rates on the VidTIMIT and Honda datasets respectively with regard to the four similarity measures. It can be seen that the heteroscedastic PLDA outperforms the traditional PLDA with a large margin especially when the number of basis vectors is small. This confirms that given enough data to learn, modeling within-class covariances separately for each individual improves performance. As the number of basis vectors increases, the traditional PLDA also achieves better performance. Interestingly, the heteroscedastic PLDA is able to achieve high classification rates i.e. nearly 100% in the VidTIMIT dataset and 85% –100% in the Honda dataset even with a few basis vectors. Similarity measures based on closest pair approach also consistently outperform the holistic sets based one. It is clearly noticeable if we look at the traditional PLDA cases or at the results on the VidTIMIT dataset. The experimental results also show that combination of the heteroscedastic PLDA and the closest pair approach has been the best among the four similarity measures. If we consider recognition on individual

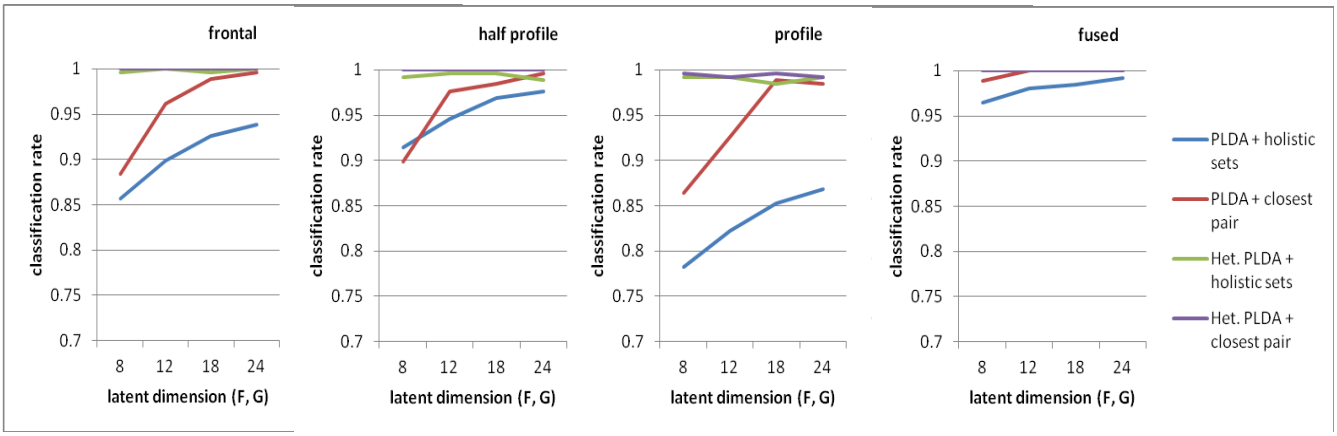


Figure 4. Classification rates on the VidTIMIT dataset.

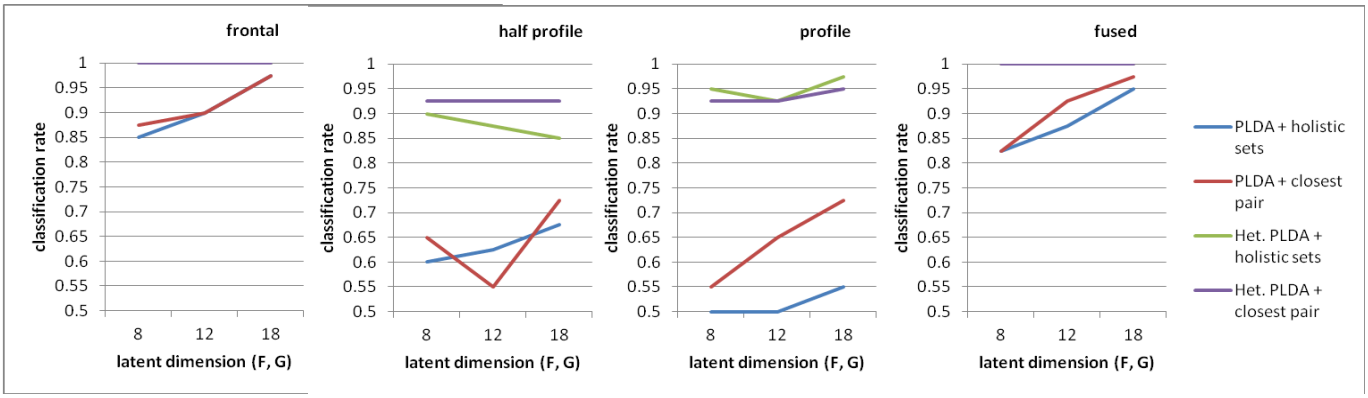


Figure 5. Classification rates on the Honda dataset.

pose, it seems that there is no significant correlation between pose and classification performance. The lower classification rates on the profile pose of the VidTIMIT dataset and on the half profile and profile poses of the Honda dataset are most likely due to the lower number of face images in those poses. When matching scores from different poses are fused, the classification rates get consistently improved. This indicates that information from different poses might complement each other in face recognition tasks. We also try to compare the proposed method with the mutual subspace method (MSM) of [7]. Repeating the same experiments on the Honda dataset, the MSM manages to achieve 97.5% classification rates when 98% of the total eigen values is included in the subspace basis.

VI. Conclusion

A video based face recognition method which employs the probabilistic linear discriminant analysis is proposed. The method evaluates four different similarity measures of image sets which are the possible combinations of closest pair vs holistic sets approaches and PLDA vs heteroscedastic PLDA models. The heteroscedastic PLDA is adapted from the traditional PLDA by modeling the individual-specific within-class covariances. Experiments on the VidTIMIT and Honda datasets show that the heteroscedastic PLDA is able to model face appearances in video better than the traditional one. Combination of the heteroscedastic PLDA and the closest pair approach is also shown to the best among the four similarity measures.

References

[1] P. J. Phillips, W. T. Scruggs, A. J. O'Toole, P. J. Flynn, K. W. Bowyer, C. L. Schott, and M. Sharpe, "FRVT 2006 and ICE 2006 large-scale experimental results," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, pp. 831-846, 2010.

[2] S. Zhou and R. Chellappa, "Face recognition from video: The essential guide to video processing 2 ed.", Boston: Academic Press, pp. 653-688, 2009.

[3] P. Unsang, A. K. Jain, and A. Ross, "Face recognition in video: Adaptive fusion of multiple matchers," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2007, CVPR '07, 2007, pp. 1-8.

[4] C. Tat-Jun, U. James, K. Schindler, and D. Suter, "Face recognition from video by matching image sets," in *Digital Image Computing: Techniques and Applications*, 2005, DICTA '05, pp. 28-28.

[5] S. J. D. Prince and J. H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *the 11th IEEE International Conference on Computer Vision*, 2007, ICCV 2007, pp. 1-8.

[6] L. Peng, F. Yun, U. Mohammed, J. H. Elder, and S. J. D. Prince, "Probabilistic models for inference about identity," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, pp. 144-157, 2012.

[7] O. Yamaguchi, K. Fukui, and K.-I. Maeda, "Face recognition using temporal image sequence," in *International Symposium of Robotics Research*, pp. 318–323, 1998.

[8] T. Wang and P. Shi, "Kernel Grassmannian distances and discriminant analysis for face recognition from image sets," *Pattern Recognition Letters*, vol. 30, pp. 1161-1165, 2009.

[9] H. Cevikalp and B. Triggs, "Face recognition based on image sets," in *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*) 2010, 2010, pp. 2567-2573.

[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* (*CVPR*) 2011, 2011, pp. 121-128.

[11] A. Hadid and M. Pietikainen, "From still image to videobased face recognition: An experimental analysis," in *International Conference on Automatic Face and Gesture Recognition*, 2004.

[12] V. Krueger and S. Zhou, "Exemplar-based face recognition from video," *in European Conference on Computer Vision* (*ECCV*) 2002, 2002, pp. 361-365.

[13] K. Lee, J. Ho, Y. Ming-Hsuan, and D. Kriegman, "Video-based face recognition using probabilistic appearance manifolds," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 2003, 2003, pp. I-313-I-320 vol.1.

[14] W. Fan and D.-Y. Yeung, "Locally linear models on face appearance manifolds with application to dual-subspace based classification," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2006.

[15] O. Arandjelovic and R. Cipolla, "A pose-wise linear illumination manifold model for face recognition using video," *Computer Vision and Image Understanding*, vol. 113, pp. 113-125, 2009.

[16] Y. Li, S. Gong, and H. Liddell, "Video-based online face recognition using identity surfaces," in *IEEE ICCV Workshop on Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems* 2001, 2001, pp. 40-46.

[17] E. Murphy-Chutorian and M. M. Trivedi, "Head pose estimation in computer vision: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, pp. 607-626, 2009.

[18] J. Sung and D. Kim, "Adaptive active appearance model with incremental learning," *Pattern Recognition Letters*, vol. 30, pp. 359-367, 2009.

[19] C. Sanderson and B.C. Lovell, "Multi-region probabilistic histograms for robust and scalable identity inference," *Lecture Notes in Computer Science* (*LNCS*), vol. 5558, pp. 199-208, 2009.