DYNAMIC PROBABILISTIC LINEAR DISCRIMINANT ANALYSIS FOR VIDEO CLASSIFICATION

Alessandro Fabris[1], *Mihalis A. Nicolaou*[2,1], *Irene Kotsia*[3,4] *and Stefanos Zafeiriou*[1,5]

[1]Deparment of Computing, Imperial College London, UK
[2]Deparment of Computing, Goldsmiths, University of London, UK
[3]Middlesex University London, [4]International Hellenic University
[5]Center for Machine Vision and Signal Analysis, University of Oulu, Finland

ABSTRACT

Component Analysis (CA) comprises of statistical techniques that decompose signals into appropriate latent components, relevant to a task-at-hand (e.g., clustering, segmentation, classification). Recently, an explosion of research in CA has been witnessed, with several novel probabilistic models proposed (e.g., Probabilistic Principal CA, Probabilistic Linear Discriminant Analysis (PLDA), Probabilistic Canonical Correlation Analysis). PLDA is a popular generative probabilistic CA method, that incorporates knowledge regarding *class-labels* and furthermore introduces class-specific and sample-specific latent spaces. While PLDA has been shown to outperform several state-of-the-art methods, it is nevertheless a static model; any feature-level temporal dependencies that arise in the data are ignored. As has been repeatedly shown, appropriate modelling of temporal dynamics is crucial for the analysis of temporal data (e.g., videos). In this light, we propose the first, to the best of our knowledge, probabilistic LDA formulation that models dynamics, the so-called Dynamic-PLDA (DPLDA). DPLDA is a generative model suitable for *video* classification and is able to jointly model the label information (e.g., face identity, consistent over videos of the same subject), as well as dynamic variations of each individual video. Experiments on video classification tasks such as face and facial expression recognition show the efficacy of the proposed method.

Index Terms**—** Probabilistic Linear Discriminant Analysis, Face Recognition, Component Analysis

1. INTRODUCTION

Component analysis techniques can be grouped based on their probabilistic or deterministic nature [1, 2, 3]. Examples of well-known deterministic techniques include Principal CA (PCA) [4], Linear Discriminant Analysis (LDA) [5, 6] and Canonical Correlation Analysis (CCA) [7], which are now customarily used in many computer vision applications. Probabilistic techniques that gained popularity include Probabilistic PCA [8, 9, 10], Probabilistic LDA [3, 11, 12, 13, 14, 15] and Probabilistic CCA (PCCA) [16, 17, 18]. Probabilistic formulations of component analysis techniques are intuitively appealing, as they (a) explicitly model observation noise, (b) facilitate the application of Bayesian methods, including application of Bayesian nonparametric methodologies [19] for learning and inference, (c) offer the potential to build composite models via the use of mixture models [20], (d) allow the presence of missing values [21] and (d) can be used as general density models [10].

One of the first attempts to formulate a probabilistic generative model that incorporates information regarding labels (e.g., facial identity) was made in [15, 22] proposing the so-called Probabilistic Linear Discriminant Analysis (PLDA). PLDA differs from PPCA in the sense that it models the data generation as a process that combines two components (a) a component which depends only on the class-label but not the particular image (i.e., it describes between-class variation) and (b) a component which is different for every image (i.e., it represents within-class variations and noise). As shown in [22, 23], PLDA significantly outperforms many component analysis techniques including deterministic LDA in face recognition and verification, as well as speaker verification. While some alternative PLDA models have been proposed, based on e.g., Mixtures of PPCA, class modelling based on continuous latent variables and Markov Random fields [3, 11, 12, 13, 14, 15], all models are static, and therefore do not capture temporal dependencies in the data at-hand, inherently falling short in terms of capturing information in case of classifying videos or temporally enriched data in general.

In this paper, we propose the first dynamic PLDA model that captures both the *identity* or *class* of data sequences, while modelling temporal dynamics behind individual time-series variations that may otherwise distort the true identity or class of the subject. Summarizing, the contributions of this work are: (i) we introduce a generative probabilistic model that exploits both the discriminating class-label information, as well as temporal dynamics, (ii) we show how to efficiently learn the model parameters and perform inference, and (iii) we apply the proposed model in various video classification tasks, such as face and facial expression recognition on videos captured in unconstrained conditions (”in-the-wild”). As shown, the proposed method performs equally well or better to state-of-the-art challenging databases, *without* being trained on vast amounts of annotated data as in other works [24, 25] .

2. PROBABILISTIC LDA

In this section, we review the PLDA model proposed in [15, 22] as it is arguably the most popular PLDA flavour and is mostly relevant to our proposed DPLDA. PLDA assumes data are generated based on two different subspaces; one that depends on the class and one that depends on the sample. That is, assuming that we have a total of I classes and each class i containing a total of J_i samples, then the j-th image of the i-th class is defined as

$$\begin{aligned} \mathbf{x}_{ij} &= \boldsymbol{\mu} + \mathbf{F}\mathbf{h}_i + \mathbf{G}\mathbf{w}_{ij} + \boldsymbol{\epsilon}_{ij}, \\ \boldsymbol{\epsilon}_{ij} &\sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}), \mathbf{h}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \mathbf{w}_{ij} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \end{aligned} \tag{1}$$

Put simply, $\mathbf{x}_{ij}$ is a vector of pixel intensities representing the image itself. According to the generative model of (1), it depends on $\mathbf{h}_i$, which is the identity latent variable specific to the i-th class, and on $\mathbf{w}_{ij}$ which is the latent variable associated to the setting in which the image was taken. Both $\mathbf{h}_i$ and $\mathbf{w}_{ij}$ are latent random variables. Intuitively in case of face recognition, $\mathbf{h}_i$ captures the facial features of a person that consistently determine his/her appearance, whereas $\mathbf{w}_{ij}$ represents incidental conditions such as pose, illumination and expression that influenced the picture at the moment it was taken. $\mathbf{F}$ is a factor matrix whose columns span the between-individual (shared) subspace. Each class is assumed to have a unique position in said subspace, which sets it apart from everyone else and is represented by the hidden variable $\mathbf{h}_i$. Analogously $\mathbf{G}$ is a matrix whose columns span the within-individual (private) subspace. $\mathbf{w}_{ij}$ represents the position of image $\mathbf{x}_{ij}$ therein and it is responsible for the differences in photos of the same individual. Finally, vector $\boldsymbol{\mu}$ is simply the mean of all images and ϵ_{ij} is a stochastic noise. Optimization is performed with EM, while the model is exploited for inferences about identity, including verification and identification by using the maximum a-posteriori (MAP) criterion [22].

3. DYNAMICAL PLDA

While PLDA has been used for face recognition in still images, one can still employ static models on videos as a set-to-set (probe set to gallery set) matching that compares all frames of a probe video against all frames of a gallery video. The most likely identity can then be identified by e.g., majority voting. Nevertheless, this approach ignores the temporal information that goes with the video. It is reasonable to expect that two consecutive frames will look fairly similar; it’s just as reasonable to impose proximity for the within-individual latent variables in two consecutive frames of the same video. Such property relies on a “video as image sequence” representation, that specifically takes into account and models temporal dependencies and dynamics. Motivated by the latter, in this section we present Dynamic PLDA (DPLDA). Let $\mathbf{x}_{ij}^t$ denote the observation on the t-th frame from the j-th video of the i-th individual and $\mathbf{w}_{ij}^t$ the respective private latent variable, PLDA is defined as

$$\begin{aligned} \mathbf{x}_{ij}^t &= \boldsymbol{\mu} + \mathbf{F}\mathbf{h}_i + \mathbf{G}\mathbf{w}_{ij}^t + \boldsymbol{\epsilon}_{ij}^t, \\ \mathbf{w}_{ij}^t &= \mathbf{A}_{ij}\mathbf{w}_{ij}^{t-1} + \mathbf{v}_{ij}^t, \\ \mathbf{w}_{ij}^1 &\sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \mathbf{h}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \\ \boldsymbol{\epsilon}_{ij}^t &\sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}), \mathbf{v}_{ij}^t \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \end{aligned} \tag{2}$$

where $\mathbf{A}_{ij}$ is the transition matrix, $\mathbf{v}_{ij}^t$ the process noise for the latent variable $\mathbf{w}_{ij}^t$, and ϵ_{ij}^t the observation noise affecting frame $\mathbf{x}_{ij}^t$. We also assume that there are I individuals in the gallery, each of which appearing in J_i videos of T_i frames each. The idea underlying the proposed model (2), is that the evolution of the frames in a video depicting the face of a person, can be ascribed to the evolution of an underlying latent variable comprising *pose, illumination* and *expression* variations. In addition to this, a constant latent variable related to identity influences the aspect of each frame, while remaining the same across all videos of one person. Explicitly modelling the dynamics of the private variable $\mathbf{w}_{ij}^t$ is likely to bring about an improvement to the discriminatory power of PLDA models when applied to videos. Intuitively, matrices $\mathbf{A}_{ij}$ will have stable dominant eigenvalues, very close to 1, which will force variations for $\mathbf{w}_{ij}^t$ to be smooth in time. This constraint is clearly absent in PLDA, and allows to better differentiate the influence of the private latent variables $\mathbf{w}_{ij}^t$, $\mathbf{w}_{ij}^{t+1}$ from the influence of the public latent variable $\mathbf{h}_i$ within consecutive frames. We optimize the parameter set $\theta = (\boldsymbol{\mu}, \mathbf{F}, \mathbf{G}, \boldsymbol{\Sigma}, \mathbf{A}_{11,\ldots,IJ})$ with respect to the expectation of the joint log-likelihood $\ln P(\mathbf{X}, \mathbf{Z}|\theta)$ wrt. posterior,

$$\begin{aligned} &\mathbb{E}\left[\ln p(\mathbf{X}, \mathbf{Z}|\theta)\right] = \\ &-\textstyle\sum_{ijt}\left\{\frac{1}{2}\ln|\boldsymbol{\Sigma}| + \frac{1}{2}(\mathbf{x}_{ij}^t - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}_{ij}^t - \boldsymbol{\mu})\right. \\ &\left.+\tfrac{1}{2}\mathbb{E}\left[\mathbf{z}_{ij}^{tT}\mathbf{B}^T\boldsymbol{\Sigma}^{-1}\mathbf{B}\mathbf{z}_{ij}^t\right] - \mathbb{E}\left[\mathbf{z}_{ij}^{tT}\mathbf{B}^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}_{ij}^t - \boldsymbol{\mu})\right]\right\} \\ &-\textstyle\sum_{ij}\sum_{t=2}^T\left\{\frac{1}{2}\mathbb{E}\left[\mathbf{w}_{ij}^{t-1\;T}\mathbf{A}_{ij}^T\mathbf{A}_{ij}\mathbf{w}_{ij}^{t-1}\right] - \mathbb{E}\left[\mathbf{w}_{ij}^{t-1\;T}\mathbf{A}_{ij}^T\mathbf{w}_{ij}^t\right]\right\}. \end{aligned} \tag{3}$$

where we ignore terms independent of θ, while $\mathbf{B}$ and $\mathbf{z}_{ij}^t$ are defined as $\mathbf{B} = [\mathbf{F}\ \mathbf{G}]$, $\mathbf{z}_{ij}^{tT} = [\mathbf{h}_i^T\mathbf{w}_{ij}^{tT}]$. By subsequently taking the derivatives with respect to θ and setting to zero, we arrive at the update equations,

$$\begin{aligned} \boldsymbol{\mu} &= \tfrac{1}{IJT}\textstyle\sum_{ijt}\mathbf{x}_{ij}^t \\ \mathbf{B} &= \left(\textstyle\sum_{ijt}(\mathbf{x}_{ij}^t - \boldsymbol{\mu})\mathbb{E}[\mathbf{z}_{ij}^{tT}]\right)\left(\textstyle\sum_{ijt}\mathbb{E}[\mathbf{z}_{ij}^t\mathbf{z}_{ij}^{tT}]\right)^{-1} \\ \boldsymbol{\Sigma} &= \tfrac{1}{IJT}\textstyle\sum_{ijt} tr\Big\{(\mathbf{x}_{ij}^t - \boldsymbol{\mu})(\mathbf{x}_{ij}^t - \boldsymbol{\mu})^T - 2\mathbf{B}\mathbb{E}[\mathbf{z}_{ij}^T] \\ &\quad (\mathbf{x}_{ij}^t - \boldsymbol{\mu})^T + \mathbf{B}\mathbb{E}[\mathbf{z}_{ij}^t\mathbf{z}_{ij}^{tT}]\mathbf{B}^T\Big\} \\ \mathbf{A}_{ij} &= \left(\textstyle\sum_{t=2}^T\mathbb{E}\left[\mathbf{w}_{ij}^t(\mathbf{w}_{ij}^{t-1})^T\right]\right)\left(\textstyle\sum_{t=2}^T\mathbb{E}\left[\mathbf{w}_{ij}^{t-1}(\mathbf{w}_{ij}^{t-1})^T\right]\right)^{-1}. \end{aligned} \tag{4}$$

In order to complete the EM for DPLA, we need to estimate the first and second order moments for the latent variables at hand. The main variation of DPLDA with respect to a typical LDS is the presence of the latent variable $\mathbf{h}_i$, shared across *all* J_i videos corresponding to subject i. The most intuitive way to tackle this is to formulate the augmented system

$$\begin{bmatrix}\mathbf{x}_{i1}^t\\ \mathbf{x}_{i2}^t\\ \vdots\\ \mathbf{x}_{ij}^t\end{bmatrix} = \begin{bmatrix}\boldsymbol{\mu}\\ \boldsymbol{\mu}\\ \vdots\\ \boldsymbol{\mu}\end{bmatrix} + \begin{bmatrix}\mathbf{F} & \mathbf{G} & \mathbf{0} & \cdots & \mathbf{0}\\ \mathbf{F} & \mathbf{0} & \mathbf{G} & \cdots & \mathbf{0}\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ \mathbf{F} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{G}\end{bmatrix} \begin{bmatrix}\mathbf{h}_i\\ \mathbf{w}_{i1}^t\\ \vdots\\ \mathbf{w}_{iJ}^t\end{bmatrix} + \begin{bmatrix}\epsilon_{i1}^t\\ \epsilon_{i2}^t\\ \vdots\\ \epsilon_{iJ}^t\end{bmatrix}, \tag{5}$$

$$\begin{bmatrix}\mathbf{h}_i\\ \mathbf{w}_{i1}^t\\ \vdots\\ \mathbf{w}_{iJ}^t\end{bmatrix} = \begin{bmatrix}I & & & & \\ & \mathbf{A}_{i1} & & & \\ & & \mathbf{A}_{i2} & & \\ & & & \ddots & \\ & & & & \mathbf{A}_{1J}\end{bmatrix} \begin{bmatrix}\mathbf{h}_i\\ \mathbf{w}_{i1}^{t-1}\\ \vdots\\ \mathbf{w}_{iJ}^{t-1}\end{bmatrix} + \begin{bmatrix}\mathbf{0}\\ \mathbf{v}_{i1}^t\\ \vdots\\ \mathbf{v}_{iJ}^t\end{bmatrix}, \tag{6}$$

which can be re-written in a more compact form as

$$\mathbf{x}_i^t = \bar{\mathbf{C}}\mathbf{z}_i^t + \bar{\boldsymbol{\mu}} + \epsilon_i, \quad \mathbf{z}_i^t = \bar{\mathbf{A}}_i\mathbf{z}_i^{t-1} + \mathbf{v}_i^t \tag{7}$$

where $\epsilon_i^t \sim \mathcal{N}(0, \bar{\mathbf{\Sigma}})$ with $\bar{\mathbf{\Sigma}}$ being a block diagonal matrix, $\bar{\mathbf{\Sigma}} = \text{blkdiag}(\mathbf{\Sigma}, \ldots, \mathbf{\Sigma})$, and $v_i^t \sim \mathcal{N}(0, \mathbf{\Gamma})$ where $\mathbf{\Gamma} = \text{blkdiag}(\mathbf{0}, \mathbf{I}, \ldots, \mathbf{I})$. By resorting to the augmented system, we can readily apply the Rauch-Tung-Striebel (RTS) smoother [26] thus obtaining the sufficient statistics for the latent variables $\mathbf{h}_i^T$ and $\mathbf{w}_{ij}^{tT}$ since $\mathbf{z}_{ij}^{tT} = [\mathbf{h}_i^T \mathbf{w}_{ij}^{tT}]$. Finally, note that Kalman filtering can also be used to estimate the likelihood of a specific sequence of frames sharing the same identity. This allows to carry out inference tasks with simplicity while exploiting the quantity $\mathbf{c}_i^t = P(\mathbf{x}_i^t | \mathbf{x}_i^1, \ldots, \mathbf{x}_i^{t-1})$ which can be computed through Kalman filtering, multiplied across all values of t to obtain the sequence likelihood.

4. EXPERIMENTS

We evaluate the proposed PLDA by performing experiments on various tasks, such as face recognition and verification using a privately collected dataset, along with the widely used Youtube face database [27] (Sec. 4.1). Furthermore, experiments for facial expression recognition are performed on the FERA database [28] (Sec. 4.2).

4.1. Face Verification and Identification Experiments

In video-based face recognition the database that has been arguably used the most is the YouTube face database [27]. It consists of 3,425 videos of 1,595 different subjects, all of which downloaded from Youtube. An average of 2.15 videos are available per subject (ranging from 1 to 6) with a mean duration of 181 frames. All videos have been tracked using a publicly available implementation of the facial landmark tracker [29]. Using the tracked landmarks the faces have been frontalised using [30]. The images were rescaled in 60 × 60 and Image Gradient orientation (IGO) features were extracted [31]. These features were used in all compared methods.

Method	Accuracy ± SE
LDA	0.723 ± 0.54
PLDA	0.830 ± 0.91
DPLDA	0.845 ± 0.65

Table 1. YTF face verification experiment accuracy

A benchmark protocol is defined for verification tasks. In more detail, 5000 pairs of videos were selected, half of them depicting the same subject, the other half belonging to different ones. The pairs are further divided into 10 splits, onto which verification has to be performed separately, exploiting the information from the other splits. The restricted protocol only allows access to this information, whereas the unrestricted protocol allows to incorporate information about the identity of the subjects during the training procedure. Since we're testing class-based methods derived from LDA, we resort to the latter protocol. For each split we selected all the people with 4 or more videos and exploit them for training. Typically the training set for each split consists of roughly 200 different identities, with 4-6 videos each.

For LDA, we computed the distance between all the frames of the first video in a pair from all the frames in the second video of the same pair. Their average is regarded as a video-to-video distance. When considering a specific split, we incorporate the same/not-same information from the remaining 9 splits to learn a linear distance-based classifier, which we then apply to the distances in the current split, determining the final confirm/reject choice. For PLDA we adopt a similar fusion metric, applying it to both the sum of the likelihoods of two frames taken separately − one from video 1, one from video 2 − and to the joint likelihood under the hypothesis that they share the same identity. We take their difference as a sufficient statistic for which we learn a confirm/reject threshold, based on the remaining 9 splits, and apply it to the current one. Finally, DPLDA was applied in its most basic version, with $\mathbf{A}_{ij} = \mathbf{A}$, i.e. estimating a single state evolution matrix supposed valid for every video. Thresholding is performed analogously to PLDA and no score fusion is necessary since the algorithm acts directly on whole videos instead of single frames. Separate results are computed for each split, and finally the 10 verification rates are averaged. The mean values are reported in Table 1.

Note that both DPLDA and PLDA perform significantly better than deterministic LDA. Also, despite the simplification $\mathbf{A}_{ij} = \mathbf{A}$, DPLDA outperforms PLDA by more than 1%. We believe that this is the case because the videos are of short duration, hence they contain a small amount of dynamical information. Furthermore, by inspecting the results from the YouTube database our method produces comparable results to the state-of-the-art, such as the method [32] which achieves an average accuracy of 84.8%, the deep learning method in [33] which achieves an average accuracy of 82.3% and [34] which achieves an average accuracy of 81.3%. Of course, our method cannot be directly compared with the deep convolutional methodology proposed in [25] which achieves an average accuracy of 91.4%, since, we did not have access to millions of annotated training data.

4.1.1. *Identification*

We assembled an “in-the-wild” database consisting of 250 videos (50 subjects, 5 videos each). The videos, featuring famous people not present in the YTF database, were downloaded from Youtube at 24fps in medium quality having around 1,000 frames each. For each video we manually checked the identity of the person depicted and the quality of the whole sequence, in order to guarantee the presence of significant temporal information to exploit while avoiding still-image slide-shows. Exploiting the pipeline described above, we test PLDA and DPLDA for closed-set and open-set identification on this database. A five cross validation experiment was carried out by picking one of the 5 videos for each person to incorporate in the probe and utilizing the remaining 4 for training. As in all PLDA methods the size of shared subspace was fixed to the number of people (i.e., 50). Fixing the size of the shared subspace to 50 we let the size of the private subspace vary from 1 to 50. For the closed set recognition PLDA achieves around 83% average accuracy, compared to 86% achieved by DPLDA.

For open set recognition, we keep the same setting, simply adding a variable number of external videos (distractors). For convenience we picked them from the YFT database, with which our database has no overlap whatsoever. As a performance measurement we pick the generalized identification accuracy, by simply contemplating an additional external class for which no training data is available. We test for different levels of “openness” by letting the percentage of external videos in the probe vary from 17 % (which was used as a validation set to derive the threshold) to 66 % of the number of videos. For every probe video, we compute the conditional likelihood of it sharing its identity with all the individuals in the gallery, storing the maximum value, along with the marginal likelihood of the probe video alone. The probe video is labelled as external if the difference between conditional and marginal likelihood is smaller than a specific threshold (set to 0.4 as was learned in the validation set), otherwise it is given the label that maximizes the conditional likelihood. The results are summarized in Table 2.

In case of 50 % distractors, DPLDA achieves around 79% while PLDA 68%. Finally, in the case of 66 % distractors, the performance of DPLDA decreases to 76%, while the performance of PLDA decreases significantly, to 58%. Hence, the incorporation of dynamic information plays a crucial role in dealing with distractors in open set face recognition.

Method	50 % distractors	66 % distractors
PLDA	68%	58%
DPLDA	79%	76%

Table 2. Open set identification experiment accuracy with varying number of distractors

Method	Average Classification
PLDA	74%
DPLDA	81%

Table 3. Average classification accuracy for facial expression recognition on FERA

4.2. Facial Expression Recognition

The final experiment we conducted was a facial expression recognition experiment using the FERA database [28]. The FERA database provides 155 labelled videos of 10 actors displaying five emotional states anger (32 videos), fear (31 videos), joy (30 videos), sadness (31 videos), and relief (31 videos). We tracked all videos using the same algorithms used in the face recognition experiments. Dense SIFT features [35] were extracted around the 49 facial landmarks that reside within the facial region and the dimensionality of the features was reduced to 20 around its landmark using Principal Component Analysis (PCA) leading to a feature vector of 980 dimensions to represent each frame. A leave-one-subject-out facial expression recognition experiment was carried out and the average classification rate was measured. PLDA achieved 74% average classification while the DPLDA 81%. Even though the result is not directly comparable with the results in FERA [28], it is worth mentioning that the best performing methodology presented in the competition did not achieve an average classification of more than 75.2%.

5. CONCLUSIONS

In this paper, we proposed the first, to the best of our knowledge, probabilistic latent variable model designed to model both (i) class-label information, and (ii) temporal dynamics, deeming it suitable for video classification. In particular, the proposed Dynamic Probabilistic Linear Discriminant Analysis (DPLDA) decomposes the observed signal into two latent parts, one based on static label information and one that models temporal dynamics and is video-dependent. The performance of DPLDA was demonstrated via experiments on tasks such as face and facial expression recognition from videos.

6. ACKNOWLEDGEMENTS

This work was partially funded by the FiDiPro program of Tekes (project number: 1849/31/2015), as well as by the European Community Horizon 2020 [H2020/2014-2020] under grant agreement no. 688520 (TeSLA).

7. REFERENCES

[1] Sam Roweis and Zoubin Ghahramani, "A unifying review of linear gaussian models," *Neural computation*, vol. 11, no. 2, pp. 305–345, 1999.

[2] Fernando De la Torre, "A least-squares framework for component analysis," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 34, no. 6, pp. 1041–1055, 2012.

[3] Mihalis A. Nicolaou, Stefanos Zafeiriou, and Maja Pantic, "A unified framework for probabilistic component analysis," in *European Conf. Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD'14)*, Nancy, France, 2014.

[4] Ian Jolliffe, *Principal component analysis*, Wiley Online Library, 2002.

[5] Keinosuke Fukunaga, *Introduction to statistical pattern recognition*, Academic press, 2013.

[6] Daniel L Swets and John Juyang Weng, "Using discriminant eigenfeatures for image retrieval," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, , no. 8, pp. 831–836, 1996.

[7] David R Hardoon, Sandor Szedmak, and John Shawe-Taylor, "Canonical correlation analysis: An overview with application to learning methods," *Neural computation*, vol. 16, no. 12, pp. 2639–2664, 2004.

[8] Baback Moghaddam and Alex Pentland, "Probabilistic visual learning for object representation," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 19, no. 7, pp. 696–710, 1997.

[9] Sam Roweis, "Em algorithms for pca and spca," *Advances in neural information processing systems*, pp. 626–632, 1998.

[10] Michael E Tipping and Christopher M Bishop, "Probabilistic principal component analysis," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 61, no. 3, pp. 611–622, 1999.

[11] Moh Edi Wibowo, Dian Tjondronegoro, Ligang Zhang, and Ivan Himawan, "Heteroscedastic probabilistic linear discriminant analysis for manifold learning in video-based face recognition," in *Applications of Computer Vision (WACV), 2013 IEEE Workshop on*. IEEE, 2013, pp. 46–52.

[12] Yu Zhang and Dit-Yan Yeung, "Heteroscedastic probabilistic linear discriminant analysis with semi-supervised extension," in *Machine Learning and Knowledge Discovery in Databases*, pp. 602–616. Springer, 2009.

[13] Shipeng Yu, Kai Yu, Volker Tresp, Hans-Peter Kriegel, and Mingrui Wu, "Supervised probabilistic principal component analysis," in *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2006, pp. 464–473.

[14] Sergey Ioffe, "Probabilistic linear discriminant analysis," in *Computer Vision–ECCV 2006*, pp. 531–542. Springer, 2006.

[15] Simon JD Prince and James H Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*. IEEE, 2007, pp. 1–8.

[16] Arto Klami, Seppo Virtanen, and Samuel Kaski, "Bayesian canonical correlation analysis," *The Journal of Machine Learning Research*, vol. 14, no. 1, pp. 965–1003, 2013.

[17] Francis R Bach and Michael I Jordan, "A probabilistic interpretation of canonical correlation analysis," 2005.

[18] Mihalis A. Nicolaou, Vladimir Pavlovic, and Maja Pantic, "Dynamic Probabilistic CCA for Analysis of Affective Behaviour and Fusion of Continuous Annotations," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2014.

[19] Neil Lawrence, "Probabilistic non-linear principal component analysis with gaussian process latent variable models," *The Journal of Machine Learning Research*, vol. 6, pp. 1783–1816, 2005.

[20] Michael E Tipping and Chris M Bishop, "Mixtures of probabilistic principal component analyzers," *Neural computation*, vol. 11, no. 2, pp. 443–482, 1999.

[21] Cédric Archambeau, Nicolas Delannay, and Michel Verleysen, "Mixtures of robust probabilistic principal component analyzers," *Neurocomputing*, vol. 71, no. 7, pp. 1274–1282, 2008.

[22] Peng Li, Yun Fu, Umar Mohammed, James H Elder, and Simon JD Prince, "Probabilistic models for inference about identity," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 34, no. 1, pp. 144–157, 2012.

[23] Patrick Kenny, Pierre Ouellet, Najim Dehak, Vishwa Gupta, and Pierre Dumouchel, "A study of interspeaker variability in speaker verification," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 16, no. 5, pp. 980–988, 2008.

[24] Florian Schroff, Dmitry Kalenichenko, and James Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 815–823.

[25] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lars Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*. IEEE, 2014, pp. 1701–1708.

[26] Christopher M Bishop, *Pattern recognition and machine learning*, springer, 2006.

[27] Lior Wolf, Tal Hassner, and Itay Maoz, "Face recognition in unconstrained videos with matched background similarity," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 529–534.

[28] Michel F Valstar, Marc Mehu, Bihan Jiang, Maja Pantic, and Klaus Scherer, "Meta-analysis of the first facial expression recognition challenge," *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, vol. 42, no. 4, pp. 966–979, 2012.

[29] Vahdat Kazemi and Josephine Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*. IEEE, 2014, pp. 1867–1874.

[30] Tal Hassner, Shai Harel, Eran Paz, and Roee Enbar, "Effective face frontalization in unconstrained images," *arXiv preprint arXiv:1411.7964*, 2014.

[31] Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic, "Subspace learning from image gradient orientations," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 34, no. 12, pp. 2454–2466, 2012.

[32] Haoxiang Li, Gang Hua, Xiaohui Shen, Zhe Lin, and Jonathan Brandt, "Eigen-pep for video face recognition," in *Computer Vision–ACCV 2014*, pp. 17–33. Springer, 2015.

[33] Junlin Hu, Jiwen Lu, and Yap-Peng Tan, "Discriminative deep metric learning for face verification in the wild," in *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*. IEEE, 2014, pp. 1875–1882.

[34] Junlin Hu, Jiwen Lu, Junsong Yuan, and Yap-Peng Tan, "Large margin multi-metric learning for face and kinship verification in the wild," in *Computer Vision–ACCV 2014*, pp. 252–267. Springer, 2015.

[35] Andrea Vedaldi and Brian Fulkerson, "Vlfeat: An open and portable library of computer vision algorithms," in *Proceedings of the international conference on Multimedia*. ACM, 2010, pp. 1469–1472.