Face Set Classification using Maximally Probable Mutual Modes

Ognjen Arandjelović Roberto Cipolla
Department of Engineering, University of Cambridge, Cambridge, UK CB2 1PZ
{oa214,cipolla}@eng.cam.ac.uk

Abstract

In this paper we consider face recognition from sets of face images and, in particular, recognition invariance to illumination. The main contribution is an algorithm based on the novel concept of Maximally Probable Mutual Modes (MMPM). Specifically: (i) we discuss and derive a local manifold illumination invariant and (ii) show how the invariant naturally leads to a formulation of "common modes" of two face appearance distributions. Recognition is then performed by finding the most probable mode, which is shown to be an eigenvalue problem. The effectiveness of the proposed method is demonstrated empirically on a challenging database containing the total of 700 video sequences of 100 individuals.

1. Introduction

In this work we are interested in illumination invariance for automatic face recognition (AFR), and, in particular, the case when both training and novel data to be matched are image *sets or sequences*. Invariance to changing lighting is perhaps the most significant practical challenge for AFR. The illumination setup in which recognition is performed is in most cases impractical to control, its physics difficult to accurately model and face appearance differences due to changing illumination are often larger than differences between individuals [1]. Additionally, the nature of most real-world AFR application is such that prompt, often real-time system response is needed, demanding appropriately efficient matching algorithms.

In this paper we propose a novel framework for face set matching under varying illumination. By considering face appearance we identify an illumination invariant in the local topology of the corresponding manifold. We next propose a manifold representation based on the invariant and derive an efficient matching method robust to pose variation.

Previous work – AFR across illumination Two most influential approaches to achieving robustness to changing lighting conditions are the illumination cones of Belhumeur *et al.* [3, 7] and the 3D morphable model of Blanz and Vetter [4]. In [3] the authors showed that the set of images of a convex, Lambertian object, illuminated by an arbitrary number of point light sources at infinity, forms a convex polyhedral cone in the image space with dimension equal to the number of distinct surface normals. In [7], Georghiades et al. successfully used this result for AFR by reilluminating images of frontal faces. In the 3D morphable model method, parameters of a complex generative model which includes the pose, shape and albedo of a face are recovered in an analysis-by-synthesis fashion.

Both illumination cones and the 3D morphable model have significant shortcomings for practical AFR use. The former approach assumes very accurately registered face images, illuminated from seven to nine different well-posed directions for each head pose. This is difficult to achieve in practical imaging conditions (see §4). On the other hand, the 3D morphable model requires high resolution [5], struggles with non-Lambertian effects and multiple light sources, has convergence problems in the presence of background clutter and partial occlusion (glasses, facial hair), and is very computationally demanding.

2. Manifold Illumination Invariants

Let us start by formalizing our recognition framework. Let $\mathbf{x}$ be an image of a face and $\mathbf{x} \in \mathbb{R}^D$, where D is the number of pixels in the image and $\mathbb{R}^D$ the corresponding image space. Then $\mathbf{f}(\mathbf{x}, \boldsymbol{\Theta})$ is an image of the same face after the rotation with parameter $\boldsymbol{\Theta} \in \mathbb{R}^3$ (yaw, pitch and roll). Function $\mathbf{f}$ is the generative function of the corresponding *face motion manifold*, obtained by varying Θ.

As a slight digression, note that strictly speaking, $\mathbf{f}$ should be *person-specific*. Due to occlusion of parts of the face, $\mathbf{f}$ cannot produce plausible images of rotated faces simply from a single image $\mathbf{x}$. However, in our work, the range of head rotations is sufficiently restricted that under the standard assumption of face symmetry [2], $\mathbf{f}$ can be considered generic.

Rotation affected appearance changes Now, consider the appearance change of a face due to small rotation $\Delta\boldsymbol{\Theta}$:

$$\Delta\mathbf{x} = \mathbf{f}(\mathbf{x}, \Delta\boldsymbol{\Theta}) - \mathbf{x}. \quad (1)$$

For small rotations, geodesic neighbourhood of $\mathbf{x}$ is linear and using Taylor's theorem we get:

$$\mathbf{f}(\mathbf{x}, \Delta\boldsymbol{\Theta}) - \mathbf{x} \approx \mathbf{f}(\mathbf{x}, \mathbf{0}) + \nabla\mathbf{f}|_{(\mathbf{x},0)} \cdot \Delta\boldsymbol{\Theta} - \mathbf{x}. \quad (2)$$

where $\nabla \mathbf{f}|_{(\mathbf{x},\boldsymbol{\Theta})}$ is the Jacobian matrix evaluated at $(\mathbf{x}, \boldsymbol{\Theta})$. Noting that $\mathbf{f}(\mathbf{x}, \mathbf{0}) = \mathbf{x}$ and writing $\mathbf{x} = \mathbf{x}_L + \mathbf{x}_H$:

$$\begin{aligned}\Delta\mathbf{x} \approx & \nabla \mathbf{f}|_{\mathbf{x},\mathbf{0}} \cdot \Delta\boldsymbol{\Theta} = \nabla \mathbf{f}|_{\mathbf{x}_L,\mathbf{0}} \cdot \Delta\boldsymbol{\Theta} + \\ & \nabla \mathbf{f}|_{\mathbf{x}_H,\mathbf{0}} \cdot \Delta\boldsymbol{\Theta} \end{aligned} \quad (3)$$

But $\mathbf{x}_L$ is by definition slowly spatially varying and therefore:

$$\left\| \nabla \mathbf{f}|_{\mathbf{x}_L,\mathbf{0}} \cdot \Delta\boldsymbol{\Theta} \right\| \ll \left\| \nabla \mathbf{f}|_{\mathbf{x}_H,\mathbf{0}} \cdot \Delta\boldsymbol{\Theta} \right\|, \quad (4)$$

and

$$\Delta\mathbf{x}/\Delta\boldsymbol{\Theta} \approx \nabla \mathbf{f}|_{\mathbf{x}_H,\mathbf{0}}. \quad (5)$$

It can be seen that $\Delta\mathbf{x}$ is a function of the person-specific $\mathbf{x}_H$ but not the illumination affected $\mathbf{x}_L$. Hence, the directions (in $\mathbb{R}^D$) of face appearance changes due to small head rotations form a *local manifold invariant* with respect to illumination variation.

Invariant representation. To effectively exploit the derived invariant, we (i) cluster manifold samples (faces) using k-means algorithm and (ii) represent each manifold as a collection of Gaussian distributions $p_i(\mathbf{x}) \sim \mathcal{N}(\mathbf{0}, \mathbf{C}_i)$, each corresponding to a single cluster c_i:

$$\mathbf{C}_i = \frac{1}{|c_i| - 1} \sum_{j \in c_i} (\mathbf{x}_j - \bar{\mathbf{x}})^T (\mathbf{x}_j - \bar{\mathbf{x}}) / \|\mathbf{x}_j - \bar{\mathbf{x}}\|^2. \quad (6)$$

The normalization of length is done in order to learn only the directions of appearance changes, not the extent which is dependant on the magnitude of pose changes.

3. Comparing Appearance PDFs

In §2 we identified a face manifold illumination invariant and addressed the issue of appropriate representation of face appearance variation, see (6). Now we turn our attention to the problem of comparing two such representations, while achieving robustness to pose changes and noise.

We pair-wise compare all clusters of two manifolds and choose the maximal of cluster similarities as the overall manifold similarity. To compare two Gaussian clusters, we propose to find the most probable mode of mutual variation between the two, disregarding the remainder of (potentially very differing) variation. Our method bears most resemblance to the canonical correlation-based pattern recognition algorithms so we next briefly summarize these.

3.1. Canonical Correlations

Canonical correlations $1 \geq \rho_1 \geq \ldots \geq \rho_d \geq 0$ between two D-dimensional linear subspaces U_1 and U_2 are uniquely defined as the maximal correlations between any two vectors of the subspaces:

$$\rho_i = \max_{\mathbf{u}_i \in U_1} \max_{\mathbf{v}_i \in U_2} \mathbf{u}_i^T \mathbf{v}_i \quad (7)$$

(a) (b)

Figure 1: *First two principal vectors for a comparison of two linear subspaces corresponding to the same (a) and different individuals (b). In the former case, the most similar modes of variation, represented by the principal vectors, are very much alike* in spite of different illumination conditions *used in data acquisition.*

subject to:

$$\mathbf{u}_i^T \mathbf{u}_i = \mathbf{v}_i^T \mathbf{v}_i = 1,\ \mathbf{u}_i^T \mathbf{u}_j = \mathbf{v}_i^T \mathbf{v}_j = 0,\ j = 1, ..., i-1. \quad (8)$$

We will refer to $\mathbf{u}_i$ and $\mathbf{v}_i$ as the i-th pair of *principal vectors*. Intuitively, the first pair of principal vectors corresponds to the most similar modes of variation of two linear subspaces; every next pair to the most similar modes orthogonal to all previous ones. This concept is illustrated in Fig. 1 on the example of sets of face appearance images.

In practice, the orthonormal basis vectors of U_i are typically found by Principal Component Analysis (PCA): only the top d components are retained for each class by assuming that class data is intrinsically low-dimensional [6]. Novel data is then classified according to the first canonical correlation ρ_1, used as a similarity score between classes.

3.2. Maximally Probably Mutual Modes

Unlike in the case when dealing with subspaces, in general both of the compared distributions can generate *any* point in the D-dimensional embedding space. Hence, the concept of the most-correlated patterns from the two classes is not meaningful in this context. Instead, we are looking for a mode – i.e. a linear direction in the pattern space – along both distributions corresponding to the two classes are most likely to "generate" observations.

We define the mutual probability $p_m(\mathbf{x})$ to be the product of two densities at $\mathbf{x}$:

$$p_m(\mathbf{x}) = p_1(\mathbf{x}) p_2(\mathbf{x}). \quad (9)$$

Generalizing this, the mutual probability $S_\mathbf{v}$ of an entire linear mode $\mathbf{v}$ is then:

$$S_\mathbf{v} = \int_{-\infty}^{+\infty} p_1(x\mathbf{v}) p_2(x\mathbf{v}) dx. \quad (10)$$

Substituting $\frac{1}{(2\pi)^{D/2} |\mathbf{C}_i|^{1/2}} \exp\left[-\frac{1}{2} \mathbf{x}^T \mathbf{C}_i^{-1} \mathbf{x}\right]$ for

$p_i(\mathbf{x})$, we obtain:

$$S_{\mathbf{v}} = \int_{-\infty}^{+\infty} \frac{1}{(2\pi)^{D/2}|\mathbf{C}_1|^{1/2}} \exp\left[-\frac{1}{2}x\mathbf{v}^T\mathbf{C}_1^{-1}\mathbf{v}x\right] \frac{1}{(2\pi)^{D/2}|\mathbf{C}_2|^{1/2}} \exp\left[-\frac{1}{2}x\mathbf{v}^T\mathbf{C}_2^{-1}\mathbf{v}x\right] dx = \quad (11)$$

$$\frac{1}{(2\pi)^{D}|\mathbf{C}_1\mathbf{C}_2|^{1/2}} \int_{-\infty}^{+\infty} \exp\left[-\frac{1}{2}x^2\mathbf{v}^T\left(\mathbf{C}_1^{-1}+\mathbf{C}_2^{-1}\right)\mathbf{v}\right] dx. \quad (12)$$

Noting that the integral is now over a 1D Gaussian distribution (up to a constant):

$$S_{\mathbf{v}} = \frac{1}{(2\pi)^{D}|\mathbf{C}_1\mathbf{C}_2|^{1/2}}(2\pi)^{1/2}\left[\mathbf{v}^T\left(\mathbf{C}_1^{-1}+\mathbf{C}_2^{-1}\right)\mathbf{v}\right]^{-1/2} = \quad (13)$$

$$(2\pi)^{1/2-D}|\mathbf{C}_1\mathbf{C}_2|^{-1/2}\left[\mathbf{v}^T\left(\mathbf{C}_1^{-1}+\mathbf{C}_2^{-1}\right)\mathbf{v}\right]^{-1/2} \quad (14)$$

The expression above favours directions in which both densities have large variances, i.e. in which Signal-to-Noise ratio is the highest, as one would intuitively expect.

The mode that maximizes the mutual probability $S_{\mathbf{v}}$ can be found by considering eigenvalue decomposition of $\mathbf{C}_1^{-1}+\mathbf{C}_2^{-1}$. Writing:

$$\mathbf{C}_1^{-1}+\mathbf{C}_2^{-1} = \sum_{i=1}^{D}\lambda_i\mathbf{u}_i\mathbf{u}_i^T, \quad (15)$$

where $0 \leq \lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_D$ and

$$\begin{aligned}\mathbf{u}_i \cdot \mathbf{u}_j &= 0, i \neq j \\ \mathbf{u}_i \cdot \mathbf{u}_i &= 1.\end{aligned} \quad (16)$$

and since $\{\mathbf{u}_i\}$ span $\mathbb{R}^D$:

$$\mathbf{v} = \sum_{i=1}^{D}\alpha_i\mathbf{u}_i, \quad (17)$$

it is then easy to shown that the maximal value of (14) is:

$$\max_{\mathbf{v}} S_{\mathbf{v}} = (2\pi)^{1/2-D}|\mathbf{C}_1\mathbf{C}_2|^{-1/2}\lambda_D^{-1/2}. \quad (18)$$

This defines the class similarity score ν. It is achieved for $\alpha_{1,\ldots,D-1} = 0$, $\alpha_D = 1$ or $\mathbf{v} = \mathbf{u}_D$, i.e. the direction of the eigenvector corresponding to the smallest eigenvalue of $\mathbf{C}_1^{-1}+\mathbf{C}_2^{-1}$. A visualization of the most probable mode between two face sets of Fig. 2 is shown in Fig. 3.

3.3. Numerical and Implementation Issues

The expression for the similarity score $\nu = \max_{\mathbf{v}} S_{\mathbf{v}}$ in (18) involves the computation of $|\mathbf{C}_1\mathbf{C}_2|^{-1/2}$. This is problematic as $\mathbf{C}_1\mathbf{C}_2$ may be a singular, or a nearly singular, matrix (e.g. because the number of face images if much lower than the image space dimensionality D).

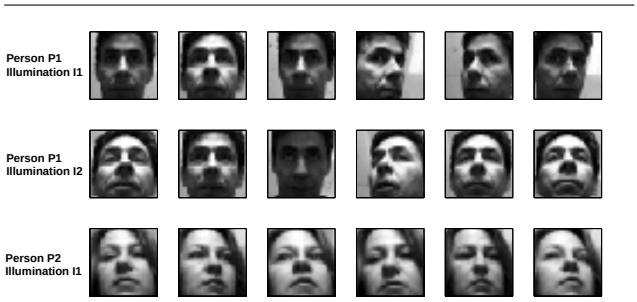
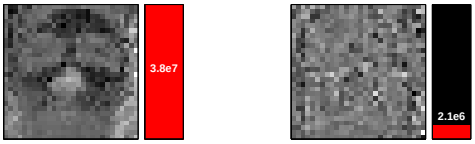


Figure 2: *Examples of detected faces from the database used for method evaluation. A wide range of illumination and pose changes is present.*



(a) [P1, I1] – [P1, I2] (b) [P1, I1] – [P2, I1]

Figure 3: *The maximally probable mutual mode, shown as image, when two compared faces sets belong to the (a) same and (b) different individuals (also see Fig. 2).*

We solve this problem by assuming that the dimensionality of the principal linear subspaces corresponding to $\mathbf{C}_1$ and $\mathbf{C}_2$ is of dimensionality $M \ll D$, and that data is perturbed by isotropic Gaussian noise. If $\lambda_1^{(i)} \leq \lambda_2^{(i)} \leq \cdots \leq \lambda_D^{(i)}$ are the eigenvalues of $\mathbf{C}_i$:

$$\forall j > M.\ \lambda_{D-j}^{(1)} = \lambda_{D-j}^{(2)}. \quad (19)$$

Then, writing

$$|\mathbf{C}_i| = \prod_{j=1}^{D}\lambda_j^{(i)}, \quad (20)$$

we get:

$$\nu = (2\pi)^{1/2-D}|\mathbf{C}_1\mathbf{C}_2|^{-1/2}\lambda_D^{-1/2} = \quad (21)$$

$$= const \times \left(\lambda_D \prod_{i=D-M+1}^{D}\lambda_i^{(1)}\lambda_i^{(2)}\right)^{-1/2}. \quad (22)$$

4. Experimental evaluation

Methods in this paper were evaluated on a database of video sequences acquired in our laboratory (from here on referred to as *FaceDB100*). This database contains 100 individuals of varying age and ethnicity, and equally represented genders. For each person in the database we collected 7

Table 1: *Recognition performance statistics (%).*

	MPMM	FaceIt	MSM	KLD
average	92.0	64.1	58.3	17.0
std	7.8	9.2	24.3	8.8

video sequences of the person in arbitrary motion (significant translation, yaw and pitch, negligible roll). Each sequence corresponds to a different illumination setting, acquired for 10s at 10fps and 320×240 pixel resolution (face size ≈ 40 to 80 pixels)[1].

To establish baseline performance, we compared our recognition algorithm to:

- State-of-the-art commercial system FaceIt® by Identix [8] (the best performing software in the recent Face Recognition Vendor Test [10]),
- *Mutual Subspace Method* (MSM) [6], used in a state-of-the-art commercial system FacePass® [13],
- KL divergence-based algorithm of Shakhnarovich *et al.* (KLD) [11],
- Majority vote across all pairs of frames using *Eigenfaces* of Turk and Pentland [14].

4.1. Results

A summary of the experimental results is shown in Table 4.1. Firstly, note that the proposed algorithm significantly outperforms other methods. A high average of 92% was obtained under extreme lighting changes.

The difficulty of the recognition task is witnessed by the performance of the KLD method which can be considered a proxy for gauging the difficulty of the task, seeing that it is expected to perform well if imaging conditions are not greatly different between training and test [11]. It is also important to observe that the standard deviation of our algorithm's performance across different training and test illuminations is lower than that of other methods, showing less dependency on the exact imaging conditions used for acquisition.

Finally, the Receiver-Operator Characteristics (ROC) curve for the MPMM method is shown in Fig. 4.

5 Summary and Conclusions

This paper introduced a novel face recognition algorithm for matching face sets, based on a local manifold invariant.

[1] See `http://removedforreview` for more information on this database and examples of video sequences.

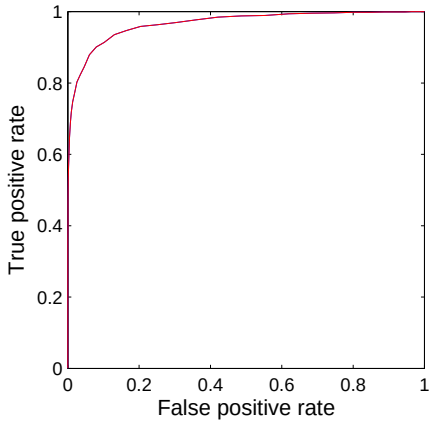


Figure 4: *Receiver-Operator Characteristic of MPMM.*

In an empirical evaluation on a large dataset of face motion sequences it outperformed state-of-the-art commercial software and set matching methods in the literature.

The main direction for future work is to match more complex manifold models (e.g. based on redundant locally linear patches), which we hope would more effectively exploit the identified manifold invariant. Another possible improvement to the method would be to augment each data set with synthetic samples obtained by random perturbations (see [9, 12]).

References

[1] Y Adini, Y. Moses, and S. Ullman. Face recognition: The problem of compensating for changes in illumination direction. *PAMI*, 19(7), 1997.

[2] O. Arandjelović and A. Zisserman. Automatic face recognition for film character retrieval in feature-length films. *CVPR*, 1, 2005.

[3] P. N. Belhumeur and D. J. Kriegman. What is the set of images of an object under all possible lighting conditions? *CVPR*, 1996.

[4] V. Blanz and T. Vetter. A morphable model for the synthesis of 3D faces. *SIGGRAPH*, 1999.

[5] M. Everingham and A. Zisserman. Automated person identification in video. *CIVR*, 2004.

[6] K. Fukui and O. Yamaguchi. Face recognition using multi-viewpoint patterns for robot vision. *Int. Symp. of Robotics Research*, 2003.

[7] A. S. Georghiades, D. J. Kriegman, and P. N Belhumeur. Illumination cones for recognition under variable lighting: Faces. *CVPR*, 1998.

[8] Identix. Faceit. `http://www.FaceIt.com/`.

[9] A. M. Martinez. Recognizing imprecisely localized, partially occluded and expression variant faces from a single sample per class. *PAMI*, 24(6), 2002.

[10] P. J. Phillips, P. Grother, R. J. Micheals, D. M. Blackburn, E. Tabassi, and J. M. Bone. FRVT 2002: Overview and summary. *Technical report, National Institute of Justice*, March 2003.

[11] G. Shakhnarovich, J. W. Fisher, and T. Darrel. Face recognition from long-term observations. *ECCV*, 3, 2002.

[12] K. K. Sung and T. Poggio. Example-based learning for view-based human face detection. *PAMI*, 20(1), 1998.

[13] Toshiba. Facepass. `www.toshiba.co.jp/mmlab/tech/w31e.htm`.

[14] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1), 1991.