Accurate and Efficient Face Recognition from Video

Ognjen Arandjelović
http://mi.eng.cam.ac.uk/~oa214

Trinity College, University of Cambridge
CB2 1TQ, UK

Computer-based face recognition continues to be a problem area of active research and outstanding practical challenges. Recently, an increasingly popular trend in the field has been to use multiple image input. These may be acquired as a video sequence or collected over time, for example after each successful authentication. The appeal of using more than a single image for recognition is rooted in greater information content available, usually in the form of multiple poses, which allows for the creation of appearance models of higher accuracy and consequently more robust matching. However, new research difficulties are introduced as well. Of most fundamental nature is the question of how multiple appearances of a person can be used optimally to form a unified and coherent representation of the person's appearance. The dramatic increase in the amount of collectable also data poses challenges of efficiency: it is impractical (and, in principle unnecessary) to retain all available data and burdensome to explicitly use all of it every time a match is required. Consequently, the scope of research challenges is broadened to invariant recognition with additional efficiency requirements.

In this paper a novel method for face recognition from image sets is introduced, based on the concept of generic illumination-shape invariant first proposed in [1]. This work is motivated by the outstandingly successful recognition performance of the original algorithm on the one hand and, as shown here, its efficiency shortcomings on the other. The approach introduced inherits the high discriminability of the underlying framework, while accomplishing a large decrease in the associated storage and computational demands.

Re-illumination Overhead. The optimization problem used to find set-to-set pose correspondences – a task of pivotal importance in the original algorithm – performs matching by evaluating geodesic distances between all appearance images. This makes it both computationally expensive (as analyzed in detail in the paper) and imposes the requirement of having the original data available each time a novel sequence is matched.

Personal Appearance Model Here it is shown that accurate re-illumination of face sets can be achieved by maintaining two *linked* mixture models for each training image set. The focus is placed on accurately capturing the distribution of each person's pose signatures. This distribution is then related to the image space by inferring the distortion of the local, linear approximations to the appearance manifold when mapped into the pose signature space. Given a set of face appearances $I = \{i_1, \ldots, i_n\}$ (where $i_q \in \mathbb{R}^D$), the aforementioned signature and appearance space mixtures are computed through the following sequence of steps:

1: Appearance images in I are mapped into the signature space, in the same manner as in the original algorithm:

$$I \xrightarrow{\text{pose sig.}} S(I) : \; S(I) = \{ \, s_q \mid s_q = S(i_q) \, \}. \tag{1}$$

2: The computed set of pose signatures is then used to estimate an approximation to the underlying probability density function, in the form of a Q-component mixture of Probabilistic PCA:

$$\hat{p}(s) = \sum_{q=1}^{Q} \left[\alpha_q \cdot \hat{G}(s \; ; \; \hat{\mu}_q, \; \hat{\Sigma}_q) \right], \tag{2}$$

where $\hat{G}(s \; ; \; \hat{\mu}_q, \; \hat{\Sigma}_q)$ is a multivariate Gaussian function in $\mathbb{R}^D$, with the mean $\hat{\mu}_q$ and the covariance matrix $\hat{\Sigma}_q$. The maximal dimensionality of the principal subspace captures the inherently low-dimensional structure of the personal face manifold and is a free parameter of the algorithm.

3: The signature space mixture is then implicitly re-projected into the image space, forming also a Q-component mixture:

$$\hat{p}(s) \xrightarrow{\text{image space}} p(i) : \; p(i) = \sum_{q=1}^{Q} \left[\alpha_q \cdot G(s \; ; \; \mu_q, \; \Sigma_q) \right]. \tag{3}$$

The unknown parameters of mixture components are computed from the known correspondences between the image and signature space, that is all i_r and s_r, and the component-wise likelihoods in the signature space $\hat{G}(s_r \; ; \; \hat{\mu}_q, \; \hat{\Sigma}_q)$. Thus, the means are:

$$\mu_q = \sum_{r=1}^{n} i_r \; \hat{G}(s_r \; ; \; \hat{\mu}_q, \; \hat{\Sigma}_q), \tag{4}$$

while the covariance matrices Σ_q are computed by fitting a Probabilistic PCA model $\Sigma_q = \mathbf{P}_q \; \Lambda_q \; \mathbf{P}_q{}^T + \rho \; \mathbf{C}_q \; \mathbf{C}_q{}^T$ to the set of intermediate estimates Σ'_q:

$$\Sigma'_q = \sum_{r=1}^{n} (i_r - \mu_q) \; (i_r - \mu_q)^T \; \hat{G}(s_r \; ; \; \hat{\mu}_q, \; \hat{\Sigma}_q). \tag{5}$$

4: In general the principal directions of Σ_q and $\hat{\Sigma}_q$ do not correspond. To infer the correspondence between the directions in the regions of appearance and pose-signature spaces dominated by respectively Σ_q and $\hat{\Sigma}_q$, the optimal linear transformation $\mathbf{R}_q$ is found which relates the projections of pose-signatures of training faces in the pose-signature space and the projections of original images in the appearance space:

$$\mathbf{R}_q = \arg \min_{\mathbf{R}} \left\{ w^T \left[\mathbf{R} \, \hat{\mathbf{P}}_q \, (\Delta_q \mathbf{S}) - \mathbf{P}_q \, (\Delta_q \mathbf{I}) \right]^T \left[\mathbf{R} \, \hat{\mathbf{P}}_q \, (\Delta_q \mathbf{S}) - \mathbf{P}_q \, (\Delta_q \mathbf{I}) \right] w \right\},$$

where

$$\Delta_q S = \left[s_1 - \hat{\mu}_q \mid \ldots \mid s_n - \hat{\mu}_q \right] \quad \Delta_q I = \left[i_1 - \mu_q \mid \ldots \mid i_n - \mu_q \right],$$

and w is a vector of weights, scaling the contribution of each face according to its Mahalanobis proximity to $\hat{\mu}_q$:

$$w(r) = (s_r - \hat{\mu}_q) \; \hat{\Sigma}_q^{-1} \; (s_r - \hat{\mu}_q), \; r = 1, \ldots, n. \tag{6}$$

Minimization in (6) can performed by a straightforward but cluttered differentiation of the quadratic form.

Matching Image Sets and Personal Appearance Models The appearance model described in the previous section was constructed so as to allow efficient re-illumination and matching of a novel set of face images. Specifically, a novel face i_r is used to compute a synthetically re-illuminated face i_r^* by independently re-illuminating it with each pair of mixture components $G(s_r \; ; \; \hat{\mu}_q, \; \hat{\Sigma}_q)$ and $\hat{G}(s_r \; ; \; \hat{\mu}_q, \; \hat{\Sigma}_q)$, and choosing the best pair under the learnt shape-illumination invariant. Thus, $i_r^* = i_{r,q^*}$ where:

$$q^* = \arg\max_q \mathscr{P}(\log i_{r,q} - \log i_r),$$
$$i_{r,q} = \mathbf{P}_q \; \mathbf{R}_q \; \hat{\mathbf{P}}_q^T \; (s_r - \hat{\mu}_q) + \mu_q.$$

The postulated shape-illumination effects are then extracted as in the original algorithm and used to compute the overall likelihood of the model [1].

Summary of results On a large data set, the proposed algorithm achieved a recognition accuracy comparable to that of the original method, with a dramatic improvement in storage requirements and matching speed, a 1600-image query set requiring only 6.7 s, which is an over 700-fold speed-up.

[1] O. Arandjelović and R. Cipolla. Face recognition from video using the generic shape-illumination manifold. *In Proc. European Conference on Computer Vision (ECCV)*, 4:27–40, May 2006.