Automatic Face Recognition for Film Character Retrieval in Feature-Length Films

Ognjen Arandjelović Andrew Zisserman
Engineering Department, University of Oxford, UK
E-mail: oa214@cam.ac.uk,az@robots.ox.ac.uk

Abstract

The objective of this work is to recognize all the frontal faces of a character in the closed world of a movie or situation comedy, given a small number of query faces. This is challenging because faces in a feature-length film are relatively uncontrolled with a wide variability of scale, pose, illumination, and expressions, and also may be partially occluded. We develop a recognition method based on a cascade of processing steps that normalize for the effects of the changing imaging environment. In particular there are three areas of novelty: (i) we suppress the background surrounding the face, enabling the maximum area of the face to be retained for recognition rather than a subset; (ii) we include a pose refinement step to optimize the registration between the test image and face exemplar; and (iii) we use robust distance to a sub-space to allow for partial occlusion and expression change. The method is applied and evaluated on several feature length films. It is demonstrated that high recall rates (over 92%) can be achieved whilst maintaining good precision (over 93%).

1. Introduction

The problem of automatic face recognition (AFR) concerns matching a detected (roughly localized) face against a database of known faces with associated identities. This task, although very intuitive to humans and despite the vast amounts of research behind it, still poses a significant challenge to computer methods, see [2, 19] for surveys. Much AFR research has concentrated on the user authentication paradigm. In contrast, we consider the content-based multimedia retrieval setup: our aim is to retrieve, and rank by confidence, film shots based on the presence of specific actors. A query to the system consists of the user choosing the person of interest in one or more keyframes. Possible applications include rapid DVD browsing or multimedia-oriented web search.

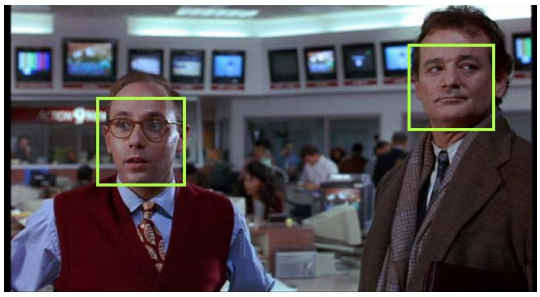


Figure 1. *Automatically detected faces in a typical frame from the feature-length film "Groundhog day". The background is cluttered, pose, expression and illumination very variable.*

We proceed from the *face detection* stage, assuming localized faces. Face detection technology is fairly mature and a number of reliable face detectors have been built, see [13, 16, 18]. We use a local implementation of the method of Schneiderman and Kanade [16] and consider a face to be correctly detected if both eyes and the mouth are visible, see Figure 1. In a typical feature-length film we obtain 2000-5000 face images which result from a cast of 10-20 primary and secondary characters.

Problem challenges. A number of factors other than identity influence the way a face appears in an image. Lighting conditions, and especially light angle, drastically change the appearance of a face [1]. Facial expressions, including closed or partially closed eyes, also complicate the problem, just as head pose does. Partial occlusions, be they artefacts in front of a face or resulting from hair style change, or growing a beard or moustache also cause problems. Films therefore provide an uncontrolled, realistic working environment for face recognition algorithms.

Method overview. Our approach consists of computing a numerical value, a distance, expressing the degree of belief that two face images belong to the same person. Low distance, ideally zero, signifies that images are of the same person, whilst a large one signifies that they are of different people.

0-7695-2372-2/05/$20.00 (c) 2005 IEEE

(a) (b) (c)

Figure 2. *The effects of imaging conditions – illumination (a), pose (b) and expression (c) – on the appearance of a face are dramatic and present the main difficulty to AFR.*

The method involves computing a series of transformations of the original image, each aimed at removing the effects of a particular extrinsic imaging factor. The end result is a *signature image* of a person, which depends mainly on the person's identity (and expression) and can be readily classified. This is summarized in Figure 3 and Algorithm 1.

1.1. Previous Work

Little work in the literature addresses AFR in a setup similar to ours. Fitzgibbon and Zisserman [11] investigated face clustering in feature films, though without explicitly using facial features for registration. Berg *et al.* [3] consider the problem of clustering detected frontal faces extracted from web news pages. In a similar manner to us, affine registration with an underlying SVM-based facial feature detector is used for face rectification. The classification is then performed in a Kernel PCA space using combined image and contextual text-based features. The problem we consider is more difficult in two respects: (i) the variation in imaging conditions in films is typically greater than in newspaper photographs, and (ii) we do not use any type of information other than visual cues (i.e. no text). The difference in the difficulty is apparent by comparing the examples in [3] with those used for evaluation in Section 3. For example, in [3] the face image size is restricted to be at least 86×86 pixels, whilst a significant number of faces we use are of lower resolution.

Everingham and Zisserman [9] consider AFR in situation comedies. However, rather than using facial feature detection a quasi-3D model of the head is used to correct for varying pose. Temporal information via shot tracking is exploited for enriching the training corpus. In contrast, we do not use any temporal information, and the use of local features (Section 2.1) allows us to compare two face images in spite of partial occlusions (Section 2.5).

Algorithm 1 Method overview

Input: novel image $\mathbf{I}$,
training signature image $\mathbf{S}_r$.
Output: distance $d(\mathbf{I}, \mathbf{S}_r)$.

1: **Facial feature localization**
$\{\mathbf{x}_i\} \leftarrow \mathbf{I}$

2: **Pose effects: registration by affine warping**
$\mathbf{I}_R = f\left(\mathbf{I}, \{\mathbf{x}_i\}, \mathbf{S}_r\right)$

3: **Background clutter: face outline detection**
$\mathbf{I}_F = \mathbf{I}_R . * mask(\mathbf{I}_R)$

4: **Illumination effects: band-pass filtering**
$\mathbf{S} = \mathbf{I}_F * \mathbf{B}$

5: **Pose effects: registration refinement**
$\mathbf{S}_f = \hat{f}(\mathbf{I}_F, \mathbf{S}_r)$

6: **Occlusion effects: robust distance measure**
$d(\mathbf{I}, \mathbf{S}_r) = \|\mathbf{S_r} - \mathbf{S_f}\|$

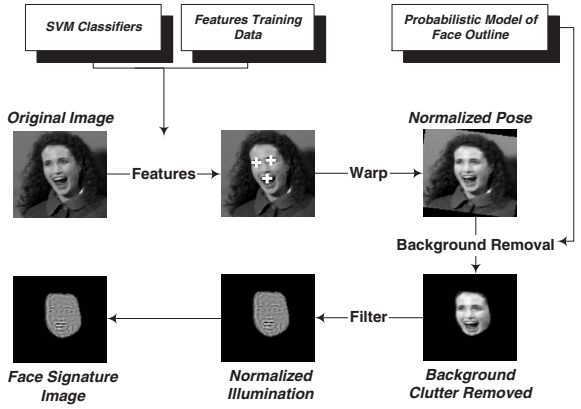


Figure 3. *Face representation: Each step in the cascade produces a result invariant to a specific extrinsic factor.*

2. Method Details

In the proposed framework, the first step in processing a face image is the normalization of the subject's pose – *registration*. After the face detection stage, faces are only roughly localized and aligned – more sophisticated registration methods are needed to correct for the effects of varying pose. One way of doing this is to "lock onto" the characteristic facial points. In our method, these facial points are the locations of the mouth and the eyes.

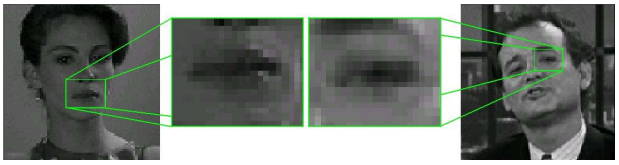


Figure 4. *The difficulties of facial feature detection: without context, distinguishing features in low resolution and bad illumination conditions is a hard task even for a human. Shown are a mouth and an eye that although easily recognized within the context of the whole image, are very similar in isolation.*

2.1. Facial Feature Detection

In the proposed algorithm Support Vector Machines[1] (SVMs) [6, 17] are used for facial feature detection. For a related approach see [3]; alternative methods include pictorial structures [10] or the method of Cristinacce *et al.* [7].

We represent each facial feature, i.e. the image patch surrounding it, by a feature vector. An SVM with a set of parameters (kernel type, its bandwidth and a regularization constant) is then trained on a part of the training data and its performance iteratively optimized on the remainder. The final detector is evaluated by a one-time run on unseen data.

2.1.1 Training

For training we use manually localized facial features in a set of 300 randomly chosen faces from the feature-length film "Groundhog day". Examples are extracted by taking rectangular image patches centred at feature locations (see Figures 4 and 5). We represent each patch $\mathbf{I} \in \mathbb{R}^{N \times M}$ with a feature vector $\mathbf{v} \in \mathbb{R}^{2N \times M}$ with appearance and gradient information (we used $N = 17$ and $M = 21$):

$$v_A(Ny + x) = I(x, y) \tag{1}$$

$$v_G(Ny + x) = |\nabla I(x, y)| \tag{2}$$

$$\mathbf{v} = \begin{pmatrix} \mathbf{v}_A \\ \mathbf{v}_G \end{pmatrix} \tag{3}$$

Local information. In the proposed method, implicit local information is included for increased robustness. This is done by complementing the image appearance vector $\mathbf{v}_A$ with the greyscale intensity gradient vector $\mathbf{v}_G$, as in (3).

Synthetic data. For robust classification, it is important that training data sets are representative of the whole spaces that are discriminated between. In uncontrolled imaging conditions, the appearance of facial features exhibits a lot of variation, requiring an appropriately large training corpus. This makes the approach with manual feature extraction impractical. In our method, a large portion of training data (1500 out of 1800 training examples) was synthetically generated. Seeing that the surface of the face is smooth and roughly fronto-parallel, its 3D motion produces locally affine-like effects in the image plane. Therefore, we synthesize training examples by applying random affine perturbations to the manually detected set.

[1] We used the LibSVM implementation freely available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm/`

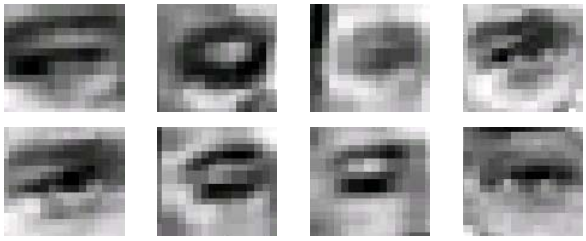


Figure 5. *A subset of data (1800 in total) used to train the eye detector. Notice the low resolution and the importance of the surrounding image context for precise localization.*

2.1.2 SVM-based Feature Detector

SVMs only provide classification decision for individual feature vectors, but no associated probabilistic information. Therefore, performing classification on all image patches produces as a result a binary image (a feature is either present or not in a particular location) from which only one feature location needs to be selected.

Our method is based on the observation that due to the robustness to noise of SVMs, the binary image output consists of connected components of positive classifications (we will refer to these as *clusters*), see Figure 6. We use a prior on feature locations to focus on the cluster of interest. Priors corresponding to the 3 features are assumed to be independent and Gaussian (2D, with full covariance matrices) and are learnt from the training corpus of 300 manually localized features described in Section 2.1.1. We then consider the total 'evidence' for a feature within each cluster:

$$\int_{\mathbf{x} \in \mathcal{S}} P(\mathbf{x}) d\mathbf{x} \tag{4}$$

where $\mathcal{S}$ is a cluster and $P(\mathbf{x})$ the Gaussian prior on the facial feature location. An unbiased feature location estimate with $\sigma \approx 1.5$ pixels was obtained by choosing the mean of the cluster with largest evidence as the final feature location, see Figures 6 and 7.

2.2. Registration

In the proposed method dense point correspondences are implicitly or explicitly used for background clutter removal, partial occlusion detection and signature image comparison (Sections 2.3- 2.5). To this end, images of faces are affine

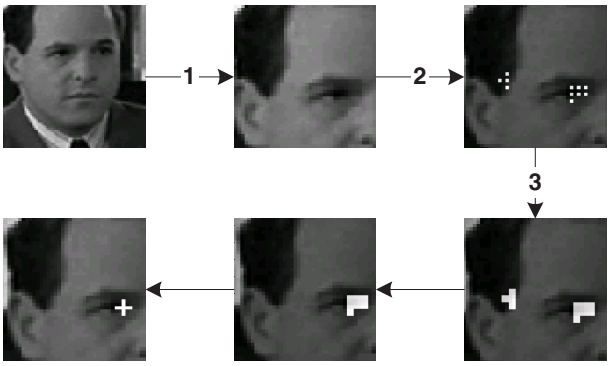


Figure 6. *Efficient SVM-based eye detection. 1: Prior on feature location restricts the search region. 2: Only $\sim 25\%$ of the locations are initially classified. 3: Morphological dilation is used to approximate the dense classification result from a sparse output.*

warped to have salient facial features aligned. The six transformation parameters are uniquely determined from three pairs of point correspondences – between detected facial features (the eyes and the mouth) and their canonical locations. In contrast to global appearance-based methods (e.g. [5, 8]) our approach is more robust to partial occlusion. It is summarized in Algorithm 2 with typical results shown in Figure 8.

Algorithm 2 Face Registration

Input: canonical facial feature locations $\mathbf{x_{can}}$,
face image $\mathbf{I}$,
facial feature locations $\mathbf{x_{in}}$.
Output: registered image $\mathbf{I}_{reg}$.

1: **Estimate the affine warp matrix**
$\mathbf{A} \leftarrow (x_{can}, x_{in})$

2: **Compute eigenvalues of A**
$\{\lambda_1, \lambda_2\} = eig(\mathbf{A})$

3: **Impose prior on shear and rescaling by A**
if $(|\mathbf{A}| \in [0.9, 1.1] \wedge \lambda_1/\lambda_2 \in [0.6, 1.3])$ **then**

4: **Warp the image**
$\mathbf{I}_{reg} = f(\mathbf{I}; \mathbf{A})$

5: **else**

6: **Face detector false +ve**
Report("$\mathbf{I}$ is not a face")

7: **endif**

2.3. Background Removal

The bounding box of a face, supplied by the face detector, typically contains significant background clutter. To realize a reliable comparison of two faces, segmentation to

Figure 7. *Automatically detected facial features: High accuracy is achieved in spite of wide variation in facial expression, pose, illumination and the presence of facial wear (glasses).*

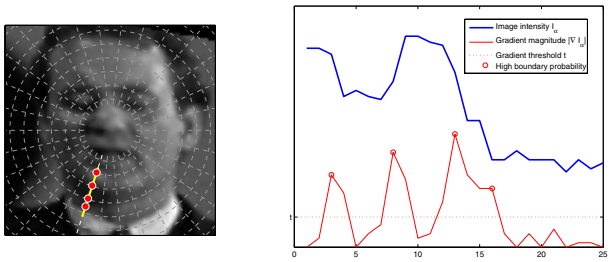
foreground (i.e. face) and background regions has to be performed. We show that the face outline can be robustly detected by combining a learnt prior on the face shape and a set of measurements of intensity discontinuity.

In detecting the face outline, we only consider points confined to a discrete mesh corresponding to angles equally spaced at $\Delta\alpha$ and radii at Δr, see Figure 9 (a). At each mesh point we measure the image intensity gradient in the radial direction – if its magnitude is locally maximal and greater than a threshold t, we assign it a constant high-probability and a constant low probability otherwise, see Figure 9 (a,b). Let $\mathbf{m}_i$ be a vector of probabilities corresponding to discrete radius values at angle $\alpha_i = i\Delta\alpha$, and r_i the boundary location at the same angle. We seek the maximum *a posteriori* estimate of the boundary radii:

$$\{r_i\} = \arg\max_{\{r_i\}} P(r_1, .., r_N | \mathbf{m}_1, .., \mathbf{m}_N) = \arg\max_{\{r_i\}} P(\mathbf{m}_1, .., \mathbf{m}_N | r_1, .., r_N) P(r_1, .., r_N) \quad (5)$$

We make the Naïve Bayes assumption for the first term in (5), whereas the second term we assume to be a first-order Markov chain. Formally:

Figure 8. *Original (top) and corresponding registered images (bottom). The eyes and the mouth in all registered images are at the same, canonical locations. Registration transformations are significant.*



(a) (b)

Figure 9. *(a) A discrete mesh in radial coordinates (only 10% of the points are shown for clarity) to which the boundary is confined. Also shown is a single measurement of image intensity in the radial direction and the detected high probability points. The plot of image intensity along this direction is shown in (b) along with the gradient magnitude used to select the high probability locations.*

$$P(\mathbf{m}_1, .., \mathbf{m}_N | r_1, .., r_N) = \prod_{i=1}^{N} P(\mathbf{m}_i | r_i) = \prod_{i=1}^{N} m_i(r_j) \tag{6}$$

$$P(r_1, .., r_N) = P(r_1) \prod_{i=2}^{N} P(r_i | r_{i-1}) \tag{7}$$

In our method model parameters (prior and conditional probabilities) are learnt from 500 manually delineated face outlines. The application of the model by maximizing (5) is efficiently realized using the Viterbi algorithm [12].

Feathering. Foreground/background segmentation produces a binary mask image $\mathbf{M}$. As well as masking the corresponding face image $\mathbf{I}_R$ (see Figure 10), we smoothly suppress image information around the boundary to achieve robustness to small errors in its localization:

$$\mathbf{M}_F = \mathbf{M} * \exp - \left(\frac{r(x,y)}{4} \right)^2 \tag{8}$$

$$I_F(x,y) = I_R(x,y) M_F(x,y) \tag{9}$$

Figure 10. *Original image, image with detected face outline, and the resulting image with the background masked.*

2.4. Compensating for Changes in Illumination

The last step in processing a face image to produce its signature is the removal of illumination effects. A crucial premise of our work is that the most significant modes of illumination changes are rather coarse – ambient light varies in intensity, while the dominant illumination source is either frontal, illuminating from the left, right, top or bottom (seldom). Noting that these produce mostly slowly varying, low spatial frequency variations [11], we normalize for their effects by band-pass filtering, see Figure 3:

$$\mathbf{S} = \mathbf{I}_F * \mathbf{G}_{\sigma=0.5} - \mathbf{I}_F * \mathbf{G}_{\sigma=8} \tag{10}$$

This defines the signature image $\mathbf{S}$.

2.5. Comparing Signature Images

In Sections 2.1–2.4 a cascade of transformations applied to face images was described, producing a signature image insensitive to illumination, pose and background clutter. We now show how the accuracy of facial feature alignment and the robustness to partial occlusion can be increased further when two signature images are compared.

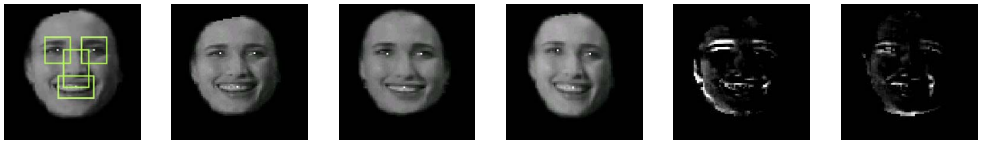
2.5.1 Improving Registration

In the registration method proposed in Section 2.2, the optimal warp parameters were estimated from 3 point correspondences in 2D. Therefore, the 6 degrees of freedom of the affine transformation were uniquely determined, making the estimate sensitive to noise. To increase the accuracy of registration, we propose a dense appearance-based affine correction to the already computed feature correspondence-based registration.

In our algorithm, the corresponding characteristic regions of two faces, see Figure 11 (a), are perturbed by small translations to find the optimal residual shift (giving the highest normalized cross-correlation score between the two overlapping regions). These new point correspondences overdetermine the residual affine transformation (estimated in the least L_2 error sense) that is applied to the image. The results are shown in Figure 11.

2.5.2 Distance

Single query image. Given two signature images in precise correspondence (see Section 2.5.1), $\mathbf{S}_1$ and $\mathbf{S}_2$, we



(a) (b) (c) (d) (e) (f)

Figure 11. *Pose refinement: (a) Salient regions of the face. (b)(c) Images aligned using features alone. (d) The salient regions shown in (a) are used to refine the pose of (b) so that it is more closely aligned with (c). The residual rotation between (b) and (c) is removed. This correction can be seen clearly in the difference images: (e) is* $|\mathbf{S}_c - \mathbf{S}_b|$*, and (f) is* $|\mathbf{S}_c - \mathbf{S}_d|$.

compute the following distance between them:

$$d_S(\mathbf{S}_1, \mathbf{S}_2) = \sum_x \sum_y h(S_1(x,y) - S_2(x,y)) \qquad (11)$$

where $h(\Delta S) = (\Delta S)^2$ if the probability of occlusion at (x, y) is low and a constant value k otherwise. This is effectively the L_2 norm with added outlier (e.g. occlusion) robustness, similar to [4]. We now describe how this threshold is determined.

Partial occlusions. Occlusions of imaged faces in films are common. Whilst some research has addressed detecting and removing specific artefacts only, such as glasses [14], here we give an alternative non-parametric approach, and use a simple appearance-based statistical method for occlusion detection. Given that the error contribution at (x, y) is $\varepsilon = \Delta S(x, y)$, we detect occlusion if the probability $P_s(\varepsilon)$ that ε is due to inter- or intra- personal differences is less than 0.05. Pixels are classified as occluded or not on an independent basis. $P_s(\varepsilon)$ is learnt in a non-parametric fashion from a face corpus with no occlusion.

The proposed approach achieved a reduction of 33% in the expected within-class signature image distance, while the effect on between-class distances was found to be statistically insignificant.

Multiple query images. The distance introduced in (11) gives the confidence measure that two signature images correspond to the same person. Often, however, more than a single image of a person is available as a query: these may be supplied by the user or can be automatically added to the the query corpus as the highest ranking matches of a single image-based retrieval. In either case we want to be able to quantify the confidence that the person in the novel image is the same as in the query *set*.

Seeing that the processing stages described so far greatly normalize for the effects of changing pose, illumination and background clutter, the dominant mode of variation across a query corpus of signature images $\{\mathbf{S}_i\}$ can be expected to be due to facial expression. We assume that the corresponding manifold of expression is linear, making the problem that of point to subspace matching [4]. Given a novel signature image $\mathbf{S}_N$ we compute a robust distance:

$$d_G(\{\mathbf{S}_i\}, \mathbf{S}_N) = d_S(\mathbf{F}\mathbf{F}^T\mathbf{S}_N, \mathbf{S}_N) \qquad (12)$$

where $\mathbf{F}$ is the projection matrix corresponding to the linear subspace that explains 95% of energy of $\{\mathbf{S}_i\}$.

3. Evaluation and Results

The proposed algorithm was evaluated on automatically detected faces from the situation comedy "Fawlty Towers" ("A touch of class" episode), and feature-length films "Groundhog day" and "Pretty woman"[2]. Detection was performed on every 10th frame, producing respectively 330, 5500, and 3000 detected faces (including incorrect detections). Face images (frame regions within bounding boxes determined by the face detector) were automatically resized to 80×80 pixels, see Figure 16 (a).

3.1. Evaluation methodology

Empirical evaluation consisted of querying the algorithm with each image in turn (or image set for multiple query images) and ranking the data in order of similarity to it. Two means of assessing the results were employed – using recall/precision curves and the *rank ordering score* ρ quantifying the goodness of a similarity-ranked ordering of data.

Rank ordering score. Given that data is recalled with a higher recall index corresponding to a lower confidence, the normalized sum of indexes corresponding to in-class faces is a meaningful measure of the recall accuracy. We call this the rank ordering score and compute it as follows:

$$\rho = 1 - \frac{S - m}{M} \qquad (13)$$

where S is the sum of indexes of retrieved in-class faces, and m and M, respectively, the minimal and maximal values S and $(S - m)$ can take. The score of $\rho = 1.0$ can be seen to correspond to orderings which correctly cluster all the data (all the in-class faces are recalled first), 0.0 to those that invert the classes (the in-class faces are recalled last), while 0.5 is the expected score of a random ordering. The *average normalized rank* [15] is equivalent to $1 - \rho$.

[2] Available at `http://www.robots.ox.ac.uk/~vgg/data/`

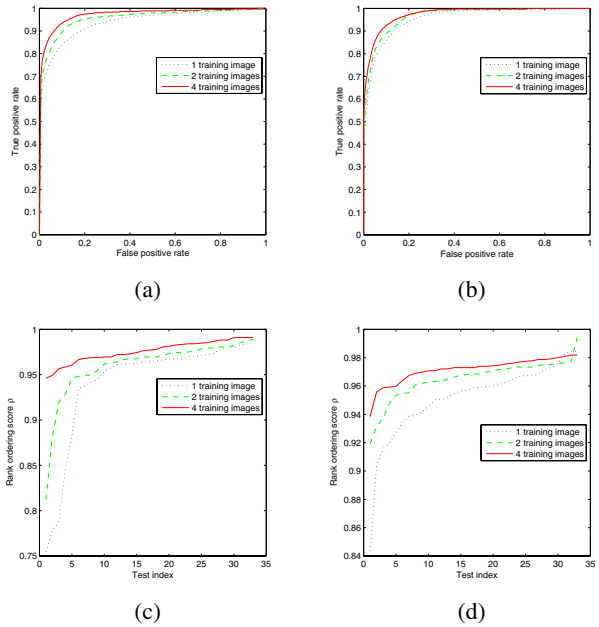


Figure 12. *(a, b) ROC curves for the retrieval of Basil and Sybil in "Fawlty Towers". The corresponding rank ordering scores across 35 retrievals are shown in (c) and (d), sorted for clarity.*

3.2. Results and Discussion

Typical Receiver Operator Characteristic (ROC) curves obtained with the proposed method are shown in Figure 12 (a, b). Excellent results are obtained using as little as 1-2 query images, typically correctly recalling 92% of the faces of the query person with only 7% of false retrievals. As expected, more query images produced better retrieval accuracy, also illustrated in Figure 12 (c, d). Note that as the number of query images is increased, not only is the ranking better on average but also more robust, as demonstrated by a decreased standard deviation of rank order scores. This is practically very important as it implies that less care needs to be taken by the user in the choice of query images. For the case of multiple query images, we compared the proposed subspace-based matching with the k-nearest neighbours approach, which was found to consistently produce worse results. The improvement of recognition with each stage of the proposed algorithm is shown in Figure 13.

Example retrievals are shown in Figures 14-16. Only a single incorrect face is retrieved, and this is with a low matching confidence (i.e. ranked amongst the last in the retrieved set). Notice the robustness of our method to pose, expression, illumination and background clutter.

4. Summary and Conclusions

The proposed approach of systematically removing particular imaging distortions – pose, background clutter, il-

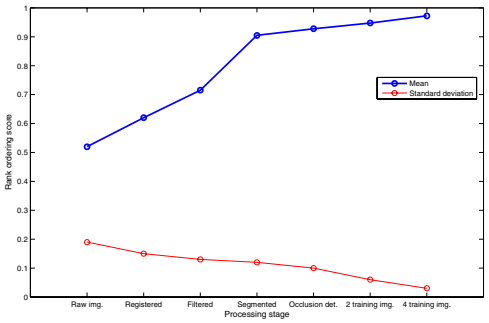


Figure 13. *The average rank ordering score of the baseline algorithm and its improvement as each of the proposed processing stages is added. The improvement is demonstrated both in the increase of the average score, and also in the decrease of its standard deviation averaged over different queries. Finally, note that the averages are brought down by few very difficult queries, which is illustrated well in Figure 12 (c,d).*

Figure 14. *The result of a typical retrieval of Basil in "Fawlty Towers". Query images are outlined. There are no incorrectly retrieved faces.*

Figure 15. *The result of a typical retrieval of Julia Roberts in "Pretty woman". Query images are outlined by a solid line, the incorrectly retrieved face by a dashed line. The performance of our algorithm is very good in spite of the small number of query images used and the extremely difficult data set – this character frequently changes wigs, makeup and facial expressions.*

lumination and partial occlusion has been demonstrated to consistently achieve high recall and precision rates.

The main research direction we intend to pursue in the future is the development of a flexible model for learning person-specific manifolds, for example due to facial expression changes. Another possible improvement to the method that we are considering is incorporating temporal information in the existing recognition framework.

Acknowledgements. We are very grateful to Mark Everingham for a number of helpful discussions and suggestions, and Krystian Mikolajczyk and Cordelia Schmid of INRIA Grenoble who supplied face detection code. Funding was provided by EC Project CogViSys.

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] W. A. Barrett. A survey of face recognition algorithms and testing results. *Systems and Computers*, 1, 1998.

[3] T. L. Berg, A. C. Berg, J. Edwards, M. Maire, R. White, Y. W. Teh, E. Learned-Miller, and D. A. Forsyth. Names and faces in the news. *CVPR*, 2004.

[4] M. J. Black and A. D. Jepson. Recognizing facial expressions in image sequences using local parameterized models of image motion. *IJCV*, 26(1), 1998.

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

[6] C. J. C. Burges. A tutorial on support vector machines for pattern recognition. *DMKD*, 2(2), 1998.

[7] D. Cristinacce, T. F. Cootes, and I. Scott. A multistage approach to facial feature detection. *BMVC*, 1, 2004.

[8] G. Edwards, C. Taylor, and T. Cootes. Interpreting face images using active appearance models. *AFG*, 1998.

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

[10] P. F. Felzenszwalb and D. Huttenlocher. Pictorial structures for object recognition. *IJCV*, 61(1), 2005.

[11] A. Fitzgibbon and A. Zisserman. On affine invariant clustering and automatic cast listing in movies. *ECCV*, 2002.

[12] G. R. Grimmett and D. R. Stirzaker. *Probability and Random Processes.* Clarendon Press, Oxford, 2nd edition, 1992.

[13] E. Hjelmås. Face detection: A survey. *CVIU*, (83), 2001.

[14] Z. Jing and R. Mariani. Glasses detection and extraction by deformable contour. *ICPR*, 2, 2000.

[15] G. Salton and M. J. McGill. *Introduction to modern information retrieval.* McGraw Hill, New York, 1983.

[16] H. Schneiderman. *A statistical approach to 3D object detection applied to faces and cars.* PhD thesis, Robotics Institute, Carnegie Mellon University, 2000.

[17] V. Vapnik. *The nature of statistical learning theory.* Springer-Verlag, 1995.

[18] P. Viola and M. Jones. Robust real-time face detection. *IJCV*, 57(2), 2004.

[19] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld. Face recognition: A literature survey. *ACM Computing Surveys*, 35(4), 2004.

(a)

(b)

(c)

Figure 16. *(a) The "Groundhog day" data set – every 30th detected face is shown for compactness. Typical retrieval results are shown in (b) and (c) – query images are outlined. There are no incorrectly retrieved faces.*