(C) IN PROC. BRITISH MACHINE VISION CONFERENCE BMVC97 UK, 1997, A. F. CLARK ED.

Fast Face Localisation and Verification

J. Matas, K. Jonsson and J. Kittler
Centre for Vision, Speech and Signal Processing
University of Surrey
Guildford, Surrey GU2 5XH, United Kingdom
{g.matas,k.jonsson}@ee.surrey.ac.uk

Abstract

We propose a method for fast face localisation and verification (identification) based on a robust form of correlation. Geometric and photometric normalisation of face images is achieved by direct minimisation. During optimisation, the correlation is estimated from a set of samples drawn from a Sobol sequence. This Monte-Carlo technique speeds the evaluation of correlation approximately twenty five times and makes the optimisation process near-real time. In recognition experiments, the optimised robust correlation outperformed two standard techniques based on the Dynamic Link Architecture [11].

1 Introduction

Personal identification (authentication, verification of identity) is an important issue in many security applications. In this paper we focus on personal identification from frontal face images. Identification is closely related to recognition, but differs in at least three fundamental aspects. Firstly, a client – an authorised user of a personal identification system – is assumed to be co-operative and makes an identity claim. Computationally this means that it is not necessary to consult the complete set of models (reference images in our case) in order to verify a claim. A test image is thus compared to a small number of reference images of the person whose identity is claimed and not, as in the recognition scenario, with every image (or some descriptor of an image) in a potentially large database. Secondly, an automatic authentication system must operate in near-real time to be acceptable to users. And finally, in recognition experiments[1] only images of people from the training database are presented to the system, whereas the case of an imposter (most likely a previously unseen person) is of utmost importance for authentication.

In this paper we propose an identification method based on optimised robust correlation. We show that in the context of the identification task it offers some advantages over standard face recognition approaches, eg. the dynamic link architecture [11] and methods based on principal component analysis [12, 17, 13]. High recognition rates for methods based on correlation of grey-level distributions in selected areas of the face have been reported [4, 5]. But since direct correlation is sensitive to changes in scale, rotation, and illumination conditions, these methods have to rely on pre-normalisation and

[1] At least as commonly reported in the field of face recognition.

pre-segmentation. The segmentation and the normalisation process typically depends on detectors of facial features. With this strategy, recognition performance depends critically on the reliability of the detector, because a failure of the detector almost certainly implies recognition failure.

In the method proposed in this paper we avoid this dependence by using an integrated approach, where localisation, normalisation (geometric and photometric) as well as identification is achieved simultaneously. To that end, a robust form of correlation is evaluated inside an optimisation loop. In the optimisation, we search the space of all affine transformations between the test image and reference images augmented to the space of all linear mappings between their corresponding grey-level values. Such direct approach clearly must evaluate hundreds of correlations per verification. This seems to be extremely computationally inefficient and therefore bound to fail the near real-time requirement of practical identification systems. However, this is not so. By evaluating the correlation in a Monte-Carlo fashion, ie. by *estimating* the correlation from a small sample of suitably chosen points, we are able to speed up the evaluation of the cost function inside the optimisation loop. In our experiments, it was sufficient to take a sample of 2-4% of pixels to converge to essentially the same solution that would have been obtained had a full image correlation been evaluated. The complete minimisation process that simultaneously achieves intensity normalisation, registration of the test and reference images and detection of outliers (occluded parts of the face, hair and beard changes) terminates in the time it would take to evaluate approximately ten full correlations (ie. computed using every pixel) of the face image. For the image resolution used in our experiments (appr. 280×350), the optimisation takes on average only a fraction of a second (see Section 3.2). Comparing this with the principle component approach, we see that the computational effort involved is similar to computing about ten projections onto eigenvectors.

The framework has a number of attractive features. It does neither require pre-registration nor does it depend on the success of a face detector (localisation). If some a priori knowledge is available, eg. estimates of scale or head positions, the optimisation process can take advantage by starting close to the optimum in the search space and thus run faster. Unlike the eigen-face methods, no manual model-building is necessary. Normally, a video stream, rather than a single image, is available to the identification system. The speed of the robust optimised correlation allows to repeat the identification process and thus achieve higher reliability.

The rest of the paper is organised as follows. In Section 2 details of the formulation of the optimisation problem are given, including the definition of the search space and the cost function (Section 2.1), the description of the search algorithm (Section 2.2) and the randomised sampling technique used for fast evaluation of correlation (Section 2.3). Experiments on recognition performance and run-time efficiency are presented in Sections 3 and, finally, results are summarised in Section 4.

2 Optimised Robust Correlation

The objective of the optimised robust correlation is to find the global extremum in a multi-dimensional search space that corresponds to the best match between a pair of images. This search space is defined by the set of all valid geometric and photometric transformations. In our implementation of the proposed method the geometric transformations are

translation, scaling and rotation[2]. Given a point in the multi-dimensional search space, a combined score function is evaluated. This function and some of its properties are described in Section 2.1. To find the global optimum of the score function a search technique based on random exponential perturbations is employed. This optimisation method, which was shown to be particularly suitable for the given search problem (see Section 3), is discussed in Section 2.2. Finally in Section 2.3, a quasi-random sampling technique used for estimating the value of the score function is outlined.

2.1 Score Function

Given a transformation t, a match score s is computed as a weighted sum of two scores, an area score s_{area} and a grey level score s_{grey}:

$$s(t) = \alpha \cdot s_{area}(t) + (1 - \alpha) \cdot s_{grey}(t)$$

where α denotes a constant in the interval $[0, 1]$ [3]. The area score is included to encourage large overlaps and is defined as

$$s_{area}(t) = \frac{|S_t \cap S_r| - |S_t \cap S_r^c|}{|S_t|}$$

where S_r and S_t denote the sampling sets corresponding to the reference and test images, I_r and I_t, respectively. The grey level score, which measures the similarity between the intensity distributions, is defined as

$$s_{grey}(t) = \frac{\sum_{p_r \in S_r \cap S_t} f_k(f_i(I_r(p_r)), I_t(f_p(t, p_r))}{f_k^{max} \cdot |S_r \cap S_t|}$$

where f_k denotes the robust kernel, f_i the intensity transformation, f_p the projection function and f_k^{max} the maximum response of the robust kernel. The kernel function f_k used in the experiments reported in Section 3 is a simple quadratic function and is defined as

$$f_k(g_1, g_2) = \begin{cases} -(g_1 - g_2)^2 + d_c^2 & \text{if} \quad |g_1 - g_2| < d_c \\ 0 & \text{otherwise} \end{cases}$$

where g_1 and g_2 denote the compared grey levels and d_c the cut-off distance. The latter defines the width of the kernel and grey level differences greater than this constant will not contribute to the final score. The intensity transformation f_i implements a linear mapping between grey levels in the reference and test images. It is defined as

$$f_i(t, g) = g \cdot t_{slope} + t_{offs}$$

where g denotes a grey level. Pixels are projected from the reference image to the test image using an affine projection function $f_p(t, p) = p'$ where p denotes the pixel to

[2]The transformations can also be represented by 3×2 matrices. However, the resulting search space is of higher dimensionality which implies a more complicated search and it is less intuitive to enforce constraints on the possible transformations in terms of eg. rotation angles.

[3]Refer to [8] for an evaluation of the impact of α on the recognition performance.

be projected and $p' = (p'_x, p'_y)$ the result of the projection. The horizontal and vertical coordinates of the projection are defined as

$$p'_x = cos(t_{rot}) \cdot t_{scale} \cdot (p_x - p^c_x) - sin(t_{rot}) \cdot t_{scale} \cdot (p_y - p^c_y) + t_{hori} + p^c_x$$
$$p'_y = sin(t_{rot}) \cdot t_{scale} \cdot (p_x - p^c_x) + cos(t_{rot}) \cdot t_{scale} \cdot (p_y - p^c_y) + t_{vert} + p^c_y$$

where $p^c = (p^c_x, p^c_y)$ denotes the centre of gravity computed from the sampling set.

2.2 Optimisation Method

The search technique we employ is based on random exponential perturbations (see Algorithm 1). In each iteration, the transformation between reference and test image is perturbed by adding a random vector drawn from an exponential distribution. The new transformation is accepted only if the score was increased.

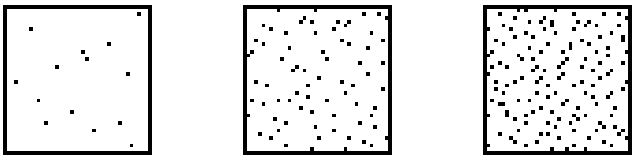
Algorithm 1 Random exponential perturbations

1: Let t_{curr} and s_{curr} denote the transformation and the score, respectively, in the current iteration. These variables are initialised by aligning the centres of gravity of the reference and test sample sets.
2: Let n_f denote the number of failed perturbations in the current iteration. This counter is initialised to zero. Furthermore, let P denote the finite set of exponentially distributed perturbations.
3: **while** $n_f < |P|$ **do**
4: Randomly select an element p from the subset of P consisting of all perturbations not yet applied in the current iteration. Optionally, the selection can be biased by the success rate of the different perturbations computed over the last n iterations (this option was not used in the experiments reported in Section 3). Create the new transformation $t_{new} = t_{curr} + p$.
5: Evaluate the score function by applying t_{new} to the reference image and comparing the result with the test image. Let s_{new} denote the obtained score.
6: **if** $s_{new} > s_{curr}$ **then**
7: $t_{curr} = t_{new}$
8: $n_f = 0$
9: **else**
10: $n_f = n_f + 1$
11: **end if**
12: **end while**

The approach described above is similar to simulated annealing [10] at zero temperature. Successful applications of simulated annealing within the areas of object detection and recognition have been reported in [9] and [1].

2.3 Random Sampling

If the object under consideration can be adequately represented using only a fraction of the pixels then this is clearly advantageous since the execution time of the method is directly dependent on the number of samples used. In our application this is certainly the



(a) 0.01 (b) 0.05 (c) 0.10

Figure 1: Two-dimensional Sobol sequences at three different sampling rates.

case since an image of a face contains a high degree of redundant information, and large regions can often be represented with only a few points.

A sampling technique commonly used in Monte Carlo integration is based on Sobol sequences [15]. A Sobol sequence is a quasi-random sequence of numbers maximally spread out over a given hyper-cube. The sequence is generated number-theoretically, rather than randomly, and successive points at any stage fill in the gaps in the previously generated distribution. The use of Sobol sequences leads to faster convergence compared to uniformly distributed random numbers since the fractional error of the approximation decreases as $ln(N)^d/N$ instead of $1/\sqrt{N}$ [15], where N is the number of samples and d the dimensionality of the approximated function. In Figure 1, three Sobol sequences are shown.

Combining the optimised robust correlation method with random sampling yields several benefits. The quasi-random nature of the process implies that the sampling points will not interfere with (and possibly cancel out) a specific frequency. In contrast, if sampling points are positioned on a grid, aliasing is much more likely. Furthermore, it is possible to continuously increase the sampling density — until some convergence criterion is met — and at the same time maintain approximately uniform density throughout the image. This is because the sampling points are avoiding the chance clustering that occurs with random points drawn from a uniform distribution. A direct consequence of this property is that Sobol sequences can be used for continuous multi-resolution matching. By varying the sampling density the image can be represented in different resolutions.

3 Experiments

The experiments summarised below were all performed on images from the M2VTS multi-modal database [14]. This publicly available database contains facial images and recordings of speech from 37 persons. For each person, 5 'shots'[4] acquired over a period of several weeks are available. A single shot is made up of 3 sequences: (1) a frontal-view sequence in which the person is counting from 0 to 9, (2) a rotation sequence in which the person is moving his or her head and (3) a rotation sequence identical to the previous one except that, if glasses are present, they are removed. Some sample images from the M2VTS database are shown in Figure 2.

[4] A take is called a shot in the M2VTS terminology.

(a) Shot 1 (b) Shot 2 (c) Shot 3 (d) Shot 4 (e) Shot 5

Figure 2: Sample images from the M2VTS database illustrating changes in the appearance of a client.

3.1 Recognition Performance

To demonstrate the overall recognition performance of the optimised robust correlation method an experiment was performed using frontal-view images from one of the two rotation sequences of the first four shots. Several different search methods were implemented and evaluated: the technique based on random perturbations described in Section 2.2, the Simplex algorithm due to Nelder and Mead [15], a direction set method due to Powell [15] and simulated annealing combined with Simplex [15][5]. Only two of these fulfill the near real-time requirements, the random perturbations and the Simplex, and the results obtained using these are presented here. The recognition performance was estimated using the *leave-one-out* methodology in which training and testing sets are disjoint. The receiver operating characteristics (ROC) are shown in Figure 3a. The equal error rates (EERs) for the random perturbations and the Simplex are 5.4% and 9.6%, respectively.

An example of a client test output is shown in Figure 4. The combined image in Figure 4c was obtained by transforming the reference image and then selecting rows interchangeably from the transformed image and the test image. The response image in Figure 4d was computed by applying the robust kernel to each pixel in the overlapping region between the transformed reference image and the test image. Mismatches appear in areas with hair change and non-rigid deformations (ie. the mouth region) as well as in the parts of the face not visible in both frames. Due to the robust kernel these mismatches do not have a disproportionate influence on the match score and the client test is successful.

The sampling density used for computing the ROC curves shown in Figure 3a was established experimentally. Using the same dataset as in the above experiment, the sampling density was increased from 0.5% until no further improvement of performance was achieved. This point was reached at a sampling rate of 4%. The ROC curves for different sampling densities are shown in Figure 5a. The difference in EER between the two extreme cases (0.5% and 4%) is 4.3%. To experimentally confirm the convergence properties of the optimised robust correlation the relative estimation error with respect to a full correlation was computed for a subset of client and impostor tests and for different sampling densities (see Figure 5b). The median relative error at a sampling rate of 4% was 1.5% and 1.4% for the client and impostor tests, respectively.

To illustrate the benefits of applying the method to image sequences an experiment was performed using several test images per shot. Since a single image-to-image comparison is completed in near real time (see Figures 3c and 3d), it is possible to repeatedly apply the method to a continuous stream of test images. For standard video equipment

[5] Refer to [8] for a complete evaluation.

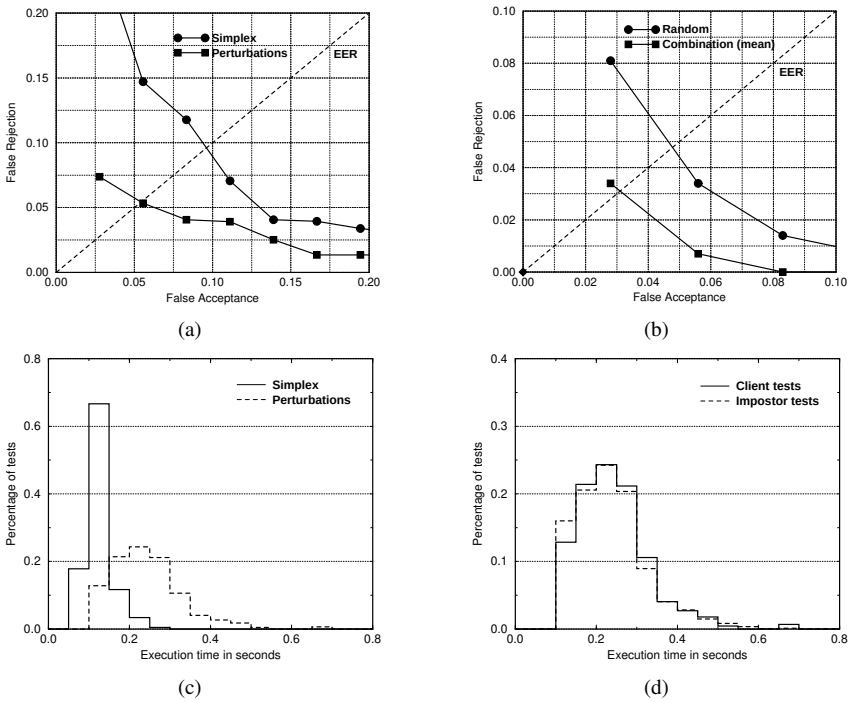
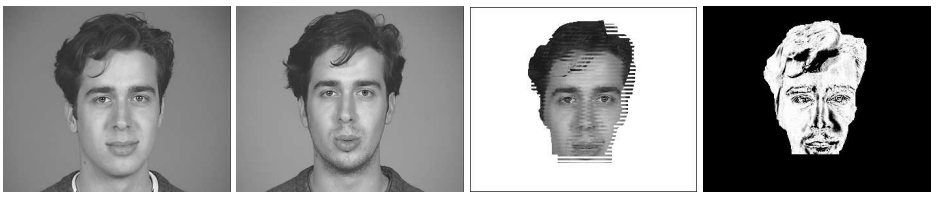
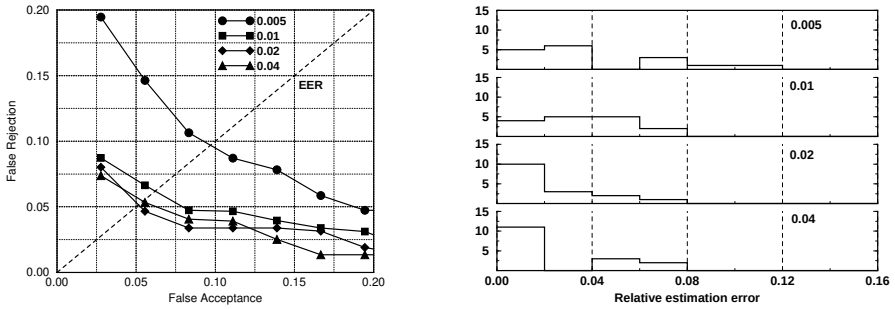


Figure 3: Performance of the optimised robust correlation: recognition performance as a function of (a) search method and (b) number of test images used; execution times on SGI Power Challenge for (c) two different search methods (client tests only) and for (d) client and impostor tests using random perturbations for optimisation.



(a) JR shot 1 (b) JR shot 3 (c) Combination (d) Response

Figure 4: An example of a client test: Person JR shot 1 against shot 3.



(a) (b)

Figure 5: The impact of sampling density on (a) recognition performance and (b) relative estimation error.

the method allows every fourth or fifth frame to be matched. By combining the scores obtained on the sequence, this approach will on average outperform the one in which a single, randomly-chosen test image is used.

For this experiment the images were selected from the frontal-view sequences of the first four shots in the M2VTS database. A lip tracker described in [16] was used to select 'shut-mouth' images. The ROC curves obtained when using sequences of test images and single, randomly-chosen ones are shown in Figure 3b. The EERs for the two cases are 3.1% and 4.8%, respectively. Note that this approach effectively includes normalisation for 3D rotation (assuming that the state of the reference image with respect to rotation will always be present in the test sequence) and changes in facial expression. Thus, the method may be used for selection of the best image for frontal-face recognition, eg. as a front-end to a more reliable, but slower, method.

3.2 Efficiency

The execution time for the optimised robust correlation depends on the sampling density and the number of optimisation steps (which is a function of the similarity of the compared images and the starting point in the multi-dimensional search space). The histogram of execution times shown in Figure 3d was obtained from more than 16000 randomly selected imposter and client tests. On average, a single identification test took 0.24 seconds. The trade-off between recognition performance and execution time is apparent when comparing Figures 3a and 3c.

4 Conclusions

The recognition experiments performed on the M2VTS database show (see Table 1) that the optimised robust correlation outperformed methods based on the dynamic link architecture [6, 2]. The speed of the method is adequate for the identification scenario. Moreover, the near real-time response allows repeated application of the method to increase the reliability of the reject/accept decision, especially in border-line cases. The performance of the method depends neither on pre-registration of the images to be matched, nor on the

Partner	*EER*	*Source*
EPFL	7.4%	AVBPA '97 [6]
	6.3%	Personal communication
AUT	13.5%	M2VTS Deliverable 3.2.1 [2]
	9.3%	Personal communication

Table 1: Results obtained by other partners within the M2VTS project.

success of feature detectors for localisation. On the contrary, registration of the compared images is achieved as an integral part of the identification process. The method is robust and needs no manually built models.

The method seems promising as implemented, but, in our opinion, it can be further strengthened by the following two improvements. At present, the same sampling density is used throughout the image. Non-uniform sampling controlled by the discriminative power of the different regions of the face is likely to improve the recognition performance [7]. Secondly, inclusion of colour information in the proposed method is fairly straightforward. Given a metric for measuring the distance between two pixels in colour space the optimised robust correlation can be applied directly without any modifications.

Acknowledgements

The research reported in this paper was carried out within the framework of the European Union ACTS project M2VTS and ESPRIT RETINA.

References

[1] M. Betke and N. C. Makris. Fast object recognition in noisy images using simulated annealing. In *Fifth International Conference on Computer Vision (Cambridge, MA, June 20–23, 1995)*, volume 1, pages 523–530, Washington, DC., 1995. Computer Society Press.

[2] J. Bigün. EU ACTS-M2VTS Deliverable 3.2.1. Technical report, Signal Processing Laboratory, Swiss Federal Institute of Technology, CH-1015 Lausanne, Switzerland, Jan 1997.

[3] J. Bigün, Gerard Chollet, and Gunilla Borgefors, editors. *Audio- and Video-based Biometric Person Authentication*, volume 1206 of *Lecture Notes in Computer Science*. Springer, 1997.

[4] R. Brunelli and T. Poggio. Face recognition: Features versus templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(10):1042–1052, Oct 1993.

[5] N. Costen, I. Craw, and S. Akamatsu. Automatic face recognition: What representation? Technical report, U. of Aberdeen, Aberdeen, UK, 1996.

[6] B. Duc, G. Maitre, S. Fischer, and J. Bigün. Person authentication by fusing face and speech information. In Bigün et al. [3], pages 21–26.

[7] K. Etemad and R. Chellappa. Discriminant analysis for recognition of human face images. In Bigün et al. [3], pages 127–142.

[8] K. Jonsson, J. Matas, and J. Kittler. Fast face localisation and verification by optimised robust correlation. Technical report, U. of Surrey, Guildford, Surrey, United Kingdom, 1997.

[9] C. Kervrann, F. Davione, P. Pèrez, H. Li, R. Forchheimer, and C. Labit. Generalized likelihood ratio-based face detection and extraction of mouth features. In Bigün et al. [3], pages 27–34.

[10] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi. Optimization by simulated annealing. *Science*, 220(4598):671–680, May 1983.

[11] M. Lades, J. C. Vorbrüggen, J. Buhmann, J. Lange, C. v.d. Malsburg, R. P. Würtz, and W. Konen. Distortion invariant object recognition in the dynamic link architecture. *IEEE Transactions on Computers*, 42(3):300–311, Mar 1993.

[12] A. Lanitis, C. J. Taylor, and T. F. Cootes. Unified approach to coding and interpreting face images. In *IEEE International Conference on Computer Vision*, pages 368–373. IEEE, Piscataway, NJ, USA, 1995.

[13] B. Moghaddam and A. Pentland. Probabilistic visual learning for object detection. In *IEEE International Conference on Computer Vision*, pages 786–793. IEEE, Piscataway, NJ, USA, 1995.

[14] S. Pigeon. The M2VTS database. Technical report, Laboratoire de Télécommunications et Télédétection, Université catholique de Louvain, Louvain-La-Neuve, Belgium, http://www.tele.ucl.ac.be/M2VTS, 1996.

[15] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling. *Numerical Recipes in C*. Cambridge University Press, 1992.

[16] M. U. Ramos Sánchez, J. Matas, and J. Kittler. Statistical chromaticity-based lip tracking with b-splines. In *International Conferene on Acoustics, Speech and Signal Processing, Munich, Germany, (April 21-24)*, volume 4, pages 2973–2976, 1997.

[17] M. A. Turk and A. P. Pentland. Face recognition using eigenfaces. In *Proceedings of the 1991 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Maui, HI, 03-06 Jun 1991, (Conf. code 16244)*, pages 586–591. IEEE, IEEE Service Center, Piscataway, NJ, USA, 1991.