Saliency-Based Robust Correlation for Real-Time Face Registration and Verification[*]

K. Jonsson, J. Matas, J. Kittler and S. Haberl
Centre for Vision, Speech and Signal Processing
University of Surrey
Guildford, Surrey GU2 5XH, United Kingdom
K.Jonsson@ee.surrey.ac.uk

Abstract

We propose a novel person verification system for real-time face identification. The main features of the system include accurate registration of face images using a robust form of correlation, a framework for global registration of a face database using a minimum spanning tree algorithm and a method for selecting a subset of features optimal for discrimination between clients and impostors. The results indicate that the image registration is of high accuracy and the feature selection is successfully improving on the verification performance.

1 Introduction

Verification of person identity based on biometric information is important for many security applications. Examples include access control to buildings, surveillance and intrusion detection. Furthermore, there are many emerging fields that would benefit from developments in person verification technology such as advanced human-computer interfaces and tele-services including tele-shopping and tele-banking. Comparing verification to recognition there are several aspects which differ. First, 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 database images (denoted *model* images below) in order to verify a claim. An incoming image (referred to as a *probe* image) is thus compared to a small number of model 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. Second, an automatic authentication system must operate in near-real time to be acceptable to users. 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 focus on person verification based on frontal face images. The method proposed here has several properties in common with the pose estimator and face recogniser presented in [2]. In both cases, localisation is performed using correlation with

[*] The research reported in this paper was funded by ACTS-M2VTS and ESPRIT RETINA.

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

BMVC 1998 doi:10.5244/C.12.5

templates in a multi-resolution fashion. However, the approach described in [2] is computationally expensive due to the large number of templates required to handle variations in scale, rotation and pose. Furthermore, the result of the registration is highly dependent on the output of feature detectors. Our work is also related to the face recognition system presented in [5]. Different robust estimators are evaluated using template matching on a pre-normalised database of face images. Other traditional approaches to face recognition include the eigenface approach [14] and the dynamic link architecture [10].

The work presented below is a development of [11] but there are a number of important differences. First, the registration is more accurate due to the inclusion of gradient information in the search process. Second, the method does not require a segmentation of the incoming image. However, an initial segmentation of the database images is still required but this can be performed off-line and does, therefore, not affect the run-time performance. Third, we present a novel framework for global registration of a database of face images and, fourth, we show how feature selection can be used to improve the accuracy and efficiency of the verification process.

In the proposed method, model images are pairwise registered using a robust form of correlation. The registration process is treated as an optimisation problem and the corresponding search space is defined by the set of all possible geometric and photometric transformations. In each point of the search space, a score function is evaluated and the optimum of this function is localised using a combined gradient-based and stochastic optimisation technique. Using gradient information we can efficiently find the local maxima and the stochastic component allows us to jump between different “hills” in the search space. To meet the real-time requirements and to ensure high registration accuracy, we use a multi-resolution scheme in both the image and parameter domains.

The verification procedure can be based directly on the output of the optimisation by thresholding the match score (this approach was evaluated in [11]). However, a more accurate and efficient approach is to select a subset of features which are optimal for discrimination between clients and impostors. A prerequisite for the feature selection is global registration since we need to match corresponding features across different model images. To achieve this we use a minimum spanning tree algorithm which reduces the complete graph obtained by matching all possible pairs of model images to a tree structure. Using this tree, we can find a unique transformation for any pair of images.

Given a global registration, we employ a feature selection procedure based on minimising the intra-class variance and at the same time maximising the inter-class variance. A feature criterion is evaluated for each pixel and the subset of pixels that best discriminates a given client from the other clients in the database (effectively modelling the impostor distribution) are selected. This feature subset is then used in verification allowing accurate and efficient registration and identification of the probe image.

The rest of the paper is organised as follows. In Section 2, we describe the technique used for pairwise registration of images, including details about the definition of the search space, the score function and the combined gradient-based and stochastic optimisation method. Furthermore, we also present the minimum spanning tree algorithm used to achieve global registration and the technique used to select a subset of features optimal for discrimination between client and impostors. In Section 3, we present results obtained through experiments on registration accuracy, selection of features and verification performance. Finally, we draw conclusions and outline future work in Section 4.

2 Saliency-Based Robust Correlation

The objective of the 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. Given a point in the multi-dimensional search space, a score function is evaluated. This function and some of its properties are described in Section 2.1. To find the global extremum of the score function a stochastic search technique incorporating gradient information 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. Matching of corresponding features across clients requires global registration. This is achieved using a minimum spanning tree algorithm described in Section 2.3. Finally, the techniques used for selecting an optimal subset of features are outlined in Section 2.4.

2.1 Score Function

Given an affine transformation $\vec{a}$, the error function expressing the intensity difference between a pixel p in the model image I_m and its projection in the probe image I_p is defined as

$$\epsilon(p, \vec{a}) = I_m(p) - I_p(T_{\vec{a}}(p))$$

where $T_{\vec{a}}$ denotes the geometric transformation function. In the current implementation, three different transformation functions are used (see Section 2.2):

$$\begin{aligned} T_{\vec{a}}(x,y) &= (x + a_1, y + a_2) && (1)\\ T_{\vec{a}}(x,y) &= (a_1 x - a_2 y + a_3, a_2 x + a_1 y + a_4) && (2)\\ T_{\vec{a}}(x,y) &= (a_1 x + a_2 y + a_3, a_4 x + a_5 y + a_6) && (3) \end{aligned}$$

Equation 1 describes a simple translational model whereas Equations 2 and 3 describe more sophisticated models incorporating (uniform or non-uniform) scaling and rotation. These affine transformations are applied globally to all pixels treating the face as a rigid object (see Section 4 for a discussion on how to extend the method to incorporate non-rigid transformations). The score function used to evaluate a match between the transformed model image and the probe image is

$$S(\mathcal{R}, \vec{a}) = \frac{1}{|R| \cdot \rho_{max}} \sum_{p \in \mathcal{R}} \rho(\epsilon(p, \vec{a}))$$

where ρ denotes a robust kernel. In other words, this function is the average percentage of the maximum kernel response taken over some region R typically obtained by segmenting the model image[2]. Possible kernel functions are the Huber Minimax and the Hampel (1,1,2) [7]. Experiments reported in [4] showed that the choice of kernel is not critical.

The parameters of the score function are purely geometrical and the intensity values are not subjected to any transformation. There are several ways in which photometric normalisation can be incorporated into the proposed method. One simple approach would be to pre-normalise the images to zero mean and unit variance. This technique has

[2] Note that this can be done off-line and there is no need for segmenting the probe image.

the obvious benefits of being fast and simple to implement but the global nature of the method in combination with the inability to model transformations of higher complexity is a clear disadvantage. Another possible approach which allows for more sophisticated transformations amounts to extending the dimensionality of the search space by including parameters controlling the transformation of intensity values (with the number of parameters deciding the complexity of the model; this approach was evaluated in [11]). For efficiency reasons, we decided to adopt a less sophisticated approach in which we (for each point in the search space) shift the histogram of residual errors using the median error. Of course, in a general context with highly-varying and non-uniform illumination this might not be adequate and an extension of the method which features local normalisation of intensity values is discussed in Section 4.

2.2 Optimisation Method

To find the global extremum of the score function we employ a stochastic search technique incorporating gradient information. The basic idea is to use the gradient to efficiently find the local optima[3] and then to jump between different hills (or valleys in the case of minimisation) by randomly perturbing the current point. The gradient-based search is implemented using steepest descent on a discrete grid with the resolution of the grid being changed during the optimisation (multi-resolution in the parameter domain) following a predefined schedule. In the four-parameter case (see Equation 2), the different components of the gradient (the partial derivatives with respect to the affine coefficients) are

$$\begin{array}{rcl}
\frac{\partial S(\mathcal{R},\vec{a})}{\partial a_1} & = & -\sum_{p\in\mathcal{R}} \Psi(\epsilon) \left(\frac{\partial I_p(p')}{\partial x} x + \frac{\partial I_1(p')}{\partial y} y \right) \\
\frac{\partial S(\mathcal{R},\vec{a})}{\partial a_2} & = & -\sum_{p\in\mathcal{R}} \Psi(\epsilon) \left(-\frac{\partial I_p(p')}{\partial x} y + \frac{\partial I_1(p')}{\partial y} x \right) \\
\frac{\partial S(\mathcal{R},\vec{a})}{\partial a_3} & = & -\sum_{p\in\mathcal{R}} \Psi(\epsilon) \left(\frac{\partial I_p(p')}{\partial x} \right) \\
\frac{\partial S(\mathcal{R},\vec{a})}{\partial a_4} & = & -\sum_{p\in\mathcal{R}} \Psi(\epsilon) \left(\frac{\partial I_p(p')}{\partial y} \right)
\end{array}$$

where Ψ denotes the influence function of the robust kernel (obtained by differentiating the kernel). The stochastic search is performed by adding a random vector drawn from an exponential distribution (meaning that small perturbations are more likely than larger ones). This optimisation technique is effectively a special case of simulated annealing [9] which has been successfully applied within the areas of object detection and recognition as reported in [8, 1].

To meet the real-time requirements of the verification scenario, we employ a multi-resolution scheme in the spatial domain. This is achieved by applying the combined gradient-based and stochastic optimisation as described above to each level in a Gaussian pyramid. The estimate obtained on one level is used to initialise the search at the next level. In addition to the speed-up, this multi-resolution search also has the benefit of removing local optima from the search space effectively improving the convergence characteristics of the method.

[3] The assumption is that the score function is sufficiently smooth close to the optimum to enable the use of the gradient.

2.3 Global Registration

The verification of a client can be based directly on the score obtained from the pairwise registration. For example, given a threshold determined using the training data (the database of model images), one possible approach would be to consider the client access as authentic if the corresponding match score exceeds the threshold. However, further improvements can be made by first selecting the optimal subset of features that best discriminates a client from the impostors (see Section 2.4). This selection procedure requires global correspondence between features and we therefore need to disambiguate our registration which amounts to reducing the complete graph obtained by matching all possible pairs of model images to a tree structure. In the current implementation, we are using a minimum spanning tree algorithm as a suboptimal way of performing the graph reduction. The technique is suboptimal since is does not take into account any information about the quality of the resulting feature selection (see Section 4 for possible extensions of the method). When constructing the tree the clients are not treated separately and each model image is linked to the closest neighbour (corresponding to the highest match score). An alternative would be to build a minimum spanning tree for each client and then to merge the set of client trees using a similar technique. This method forces client frames to be linked together which should improve the client-wise correspondence. However, there is a trade-off between client-wise and global correspondence and an improvement in one is likely to worsen the other. Examples of a client mean image (obtained by building a minimum spanning tree for the set of client frames) and a global mean image (computed from the global registration of all clients) are given in Section 3.

2.4 Feature Selection

Given a global registration of the model database, the optimal subset of features can be identified for each client and used in the subsequent verification. In this way, the verification can be made more robust and efficient since noisy features are suppressed and the dimensionality of the client representation is significantly reduced. The basic idea is to minimise the intra-class variance and at the same time maximise the inter-class variance. In the following we briefly present the theoretical framework underlying this idea.

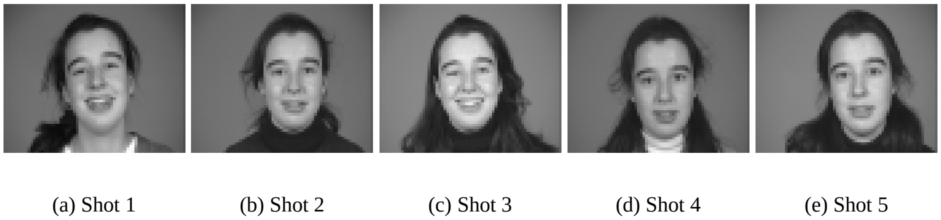
The method we have adopted uses the inter-class scatter f and the intra-class scatter g, computed for each pixel p as

$$f(p) = \frac{1}{N-1} \sum_{i=1}^{N} (\mu_i(p) - \mu(p))^2 \tag{4}$$

$$g_c(p) = \frac{1}{M-1} \sum_{i=1}^{M} (I_{(c,i)}(p) - \mu_c(p))^2 \tag{5}$$

where N denotes the number of clients, M the number of images per client, c the current client, $I_{(c,i)}$ the ith image of c, μ_c the mean image corresponding to c and μ the overall mean image. Given the inter-class scatter and the intra-class scatter, we define the criterion for class separability as

$$K(p) = \frac{f(p)}{g_c(p)}$$



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

Figure 1: Frontal-face images from the M2VTS database.

Since we want to minimise the intra-class variance and at the same time maximise the inter-class variance we are looking for pixels with high values in K. However, before the optimal subset of pixels is extracted, a non-maximum-suppression technique is applied on the criterion matrix with the aim of decorrelating the features (neighbouring pixels are likely to be correlated and can therefore be represented by a single pixel corresponding to the highest value in K).

Given the optimal set of pixels, the verification is performed by computing the score function over the feature subset. There are several issues in this context that need to be addressed. For example, how to combine the scores obtained on different pixels. One possibility would be to weight the pixels according to the corresponding feature criteria. However, the natural choice for the robust correlation is to select the optimal set of pixels and then to adjust the width of the kernel function according to the variance of the individual pixels. The results reported in Section 3 were obtained using straight averaging but future implementations of the method will incorporate pixel-dependent kernel width. Another issue is how many features should be used to represent a given client. It should be noted that this is not a traditional feature selection problem since we are aiming for a redundant set of features to be robust against occlusion, slight mis-registrations and so on. Furthermore, some clients may require more features than others for an adequate representation.

3 Experimental Results

The experiments summarised below were all performed on images from the M2VTS multi-modal database [12]. This publicly available database contains facial images and recordings of speech from 37 persons. For each person, 5 shots 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 1.

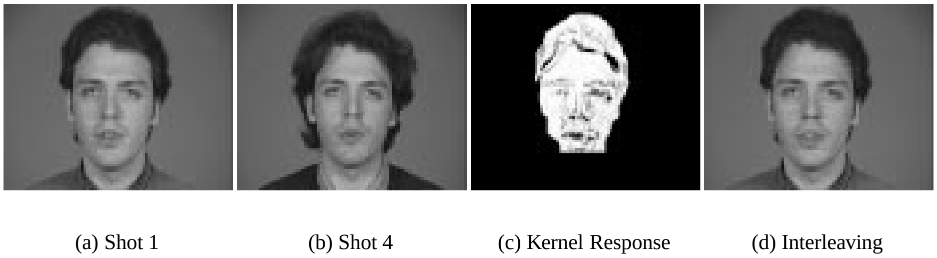

(a) Shot 1 (b) Shot 4 (c) Kernel Response (d) Interleaving

Figure 2: An example of a client registration: Person BP shot 1 against shot 4.

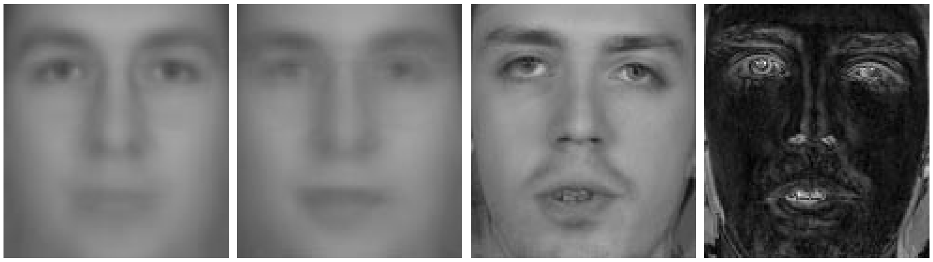
3.1 Registration and Verification Performance

An example of a client registration is shown in Figure 2. The response image in Figure 2c was computed by applying the robust kernel to each pixel in the overlapping region between the transformed model image and the probe image. The combined image in Figure 2d was obtained by transforming the model image and then selecting rows interchangeably from the transformed image and the probe image. Mismatches appear in areas with hair change and non-rigid deformations (i.e. 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 registration is successful. The result of a more quantitative analysis of the registration error is shown in Figure 4c. The accuracy of the pairwise alignment was estimated as the Euclidean distance between the manually located eye coordinates of the probe image and the corresponding coordinates predicted by the robust correlation. The median deviation is 1.87 pixels.

The accuracy of the registration can also be verified by examining the mean images shown in Figure 3. The mean image obtained using automatic registration (Figure 3b) is more or less equally blurred compared to the one obtained using manual registration (Figure 3a). The main difference is in the eye regions which is a direct consequence of the fact that the manual registration was performed by aligning the eye centres. Figures 3c and 3d show the mean and variance[4] images for client BP obtained using automatic registration.

To demonstrate the overall verification performance of the robust correlation a set of experiments were performed using frontal-view images from one of the two rotation sequences of the first four shots. The verification performance was estimated using the *leave-one-out* methodology in which training and testing sets are disjoint. Four different experiments were performed: manual registration followed by feature selection, automatic registration with and without feature selection and automatic registration using the mean image for each client instead of the individual frames in verification. The receiver operating characteristics (ROC) for these four cases are shown in Figure 4a. As expected, the manual registration performs slightly better than the automatic one. The equal error rate (EER) in the two cases are 9.3% and 10.1% (using the mean image for each client). In the case of automatic registration, there is a clear improvement when the feature se-

[4] The variance image was obtained using gamma correction to emphasise areas of large variance.



(a) Mean, Man Reg (b) Mean, Auto Reg (c) BP Mean, Auto Reg (d) BP Variance, Auto Reg

Figure 3: Mean and variance images after registration: (a) global mean image (manual registration), (b) global mean image (automatic registration), (c) mean image for client BP (automatic registration) and (d) variance image for client BP (automatic registration).

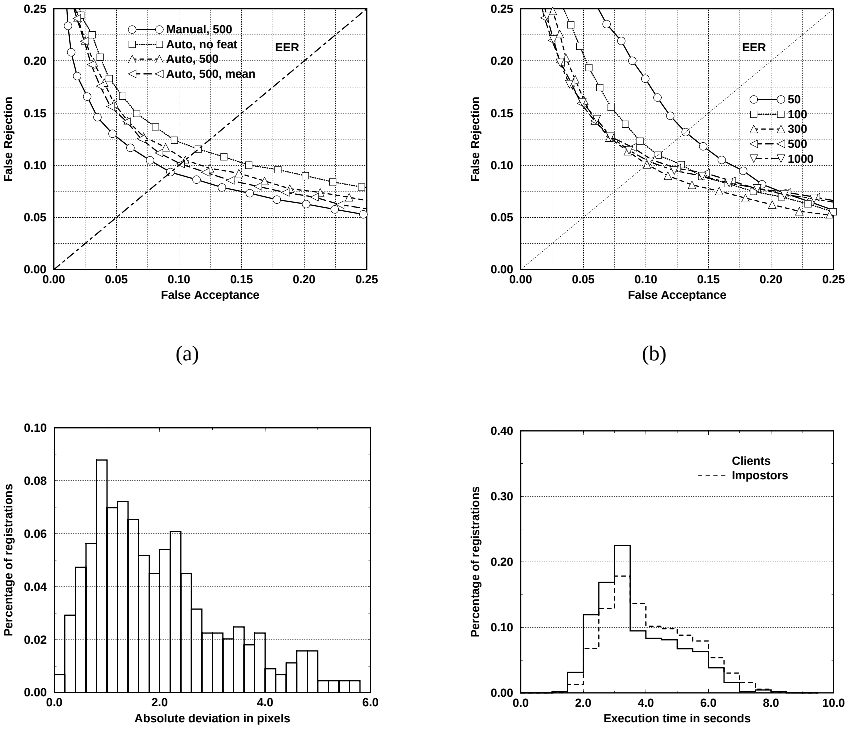
lection is included. The EERs with and without feature selection are 10.5% and 11.5%, respectively. Furthermore, the EER is slightly reduced when the mean image is used instead of the individual frames. In the latter case, we adjust the inter-class scatter $f(p)$ by adding a correction factor. Following the notation in Section 2.4, the corrected inter-class scatter is

$$f'(p) = f(p) + \frac{N}{N-1}(\mu(p) - \mu_c(p))^2$$

The dependence of the verification performance on the number of features used to represent the clients is shown in Figure 4b. The number of features was increased until no further improvements in verification performance were achieved. The saturation point is at approximately $400 - 500$ features which corresponds to a sampling rate of 3%. Results obtained by other research groups on the M2VTS database include EER 3.7% and 5.4% using elastic graph matching based on mathematical morphology [13] and local discriminants [6], respectively. These methods are, however, computationally expensive and the typical run-time is a factor of ten higher than the robust correlation (depending on how many features are used to represent the clients).

3.2 Efficiency Considerations

The execution time for the robust correlation depends on the number of features used for representation of the client 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). In Figure 4d, an histogram of execution times of client and impostor registrations is shown. These run-times should be considered an upper bound since all pixels were used (this is the case for the pairwise registration of the model images which takes place before the feature selection). The execution times were obtained from more than 16000 registrations and the median run-times for client and impostors were 3.4 and 3.9 seconds, respectively.



(a)

(b)

(c)

(d)

Figure 4: Verification performance as function of (a) registration method (manual versus automatic) and feature selection scheme (no selection versus selection followed by verification using individual frames and mean image, respectively) and (b) number of features; (c) registration accuracy shown as absolute deviations in eye coordinates; (d) histogram of client and impostor registration times on Sun Ultra Enterprise 450.

4 Concluding Remarks and Future Work

We have described a novel person verification system based on frontal face images. The main features of the system include accurate registration of face images using a combined gradient-based and stochastic optimisation technique, a framework for global registration of a face database using a minimum spanning tree algorithm and a method for selecting a subset of features optimal for discrimination between clients and impostors. The results indicate that the pairwise registration is of high accuracy and the feature selection is successfully improving on the verification performance. However, the global registration needs further improvements and one way of achieving this is to incorporate information about the quality of the resulting feature selection. This can be carried out using a feedback system in which the discrimination criterion is iteratively fed back from the feature selection module to the registration module. The proposed method can also be extended to handle non-rigid transformations in a way similar to [2]. The basic idea is to apply a two-step procedure in which the image is first registered using 2D affine transformations and then, in a second step, the optical flow is computed and a 2D warp is performed on the output of the flow computation. Finally, pose invariance can be achieved by applying techniques similar to the one described in [15]. The 2D shape and texture of the probe image are decomposed into example 2D shapes and textures. This results in a representation which is invariant under any 3D affine transformation. A WWW demonstration of the robust correlation can be reached at

http://www.ee.surrey.ac.uk/Research/VSSP/demos/face-ver/

References

[1] M. Betke and N. C. Makris. Fast object recognition in noisy images using simulated annealing. In *ICCV'95*, volume 1, pages 523–530, Washington, DC., 1995. Computer Society Press.

[2] D. J. Beymer. Face recognition under varying pose. In *CVPR'94*, pages 756–761. IEEE, Los Alamitos, CA, 1994.

[3] J. Bigün, Gerard Chollet, and Gunilla Borgefors, editors. *AVBPA'97*, volume 1206 of *Lecture Notes in Computer Science*. Springer, 1997.

[4] M. Bober and J. Kittler. Robust motion analysis. In *CVPR'94*, pages 947–952, Washington, DC., Jun 1994. Computer Society Press.

[5] R. Brunelli and S. Messelodi. Robust estimation of correlation with applications to computer vision. *Pattern Recognition*, 28:833–841, 1995.

[6] B. Duc, E. S. Bigun, J. Bigun, G. Maitre, and S. Fischer. Expert conciliation for multi modal person authentication systems by bayesian statistics. *Pattern Recognition Letters*, 18:835–843, 1997.

[7] F. R. Hampel, E. M. Ronchetti, P.J. Rouseseeuw, and W.A. Stahel. *Robust Statistics*. John Wiley, 1986.

[8] 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.

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

[10] 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.

[11] J. Matas, K. Jonsson, and J. Kittler. Fast face localisation and verification. In A. Clark, editor, *BMVC'97*, pages 152–161. BMVA Press, 1997.

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

[13] A. Tefas, C. Kotropoulos, and I. Pitas. Variants of dynamic link architecture based on mathematical morphology for frontal face authentication. In *CVPR'98*, pages 814–819, Washington, DC., June 1998. Computer Society Press.

[14] M. A. Turk and A. P. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1):71–86, 1991.

[15] T. Vetter. Learning novel views to a single face image. In *FG'96*, volume 1, pages 22–27, Los Alamitos, CA, USA, 1996. Computer Society Press.