User Authentication based on Face Recognition with Support Vector Machines

Paolo Abeni, Madalina Baltatu, Rosalia D'Alessandro
Telecom Italia, Security Innovation
Via Reiss Romoli 274, 10148 Torino, Italy
{paolo.abeni, madalina.baltatu, rosalia.dalessandro}@telecomitalia.it

Abstract

The present paper proposes an authentication scheme which relies on face biometrics and one-class Support Vector Machines. The proposed recognition procedures are based on both a global approach and on a combination of a global and a component-based approaches. Two different features extraction methods and three light compensation algorithms are tested. The combined system outperforms the global system and yields a significant performance enhancement with respect to the prior results obtained with the one-class Support Vector Machines approach for face recognition.

1. Introduction

In a context in which identity theft attacks represent an important part of the reported security incidents, biometrics-based techniques are becoming a very interesting authentication and identification method for both IT Security providers and, actually, for an increasing number of sectors of the civil society.

Biometrics identifiers are built from an unique, physical or behavioral trait of an individual for automatically recognizing or verifying her identity [18]. As such, they provide a solution to the problem of unequivocal identification of users, and, hence, can efficiently prevent identity theft and un-authorized access attacks. Biometric identifiers have some advantages over conventional, symmetric key mechanisms (like passwords, one-time passwords, tokens and the like): they are uniquely and permanently associated with their owners, they cannot be lost or forgotten, are difficult to forge (generally, it requires more resources to forge biometrics than compromise passwords), and, last but not least they are easier to use, eliminating the need to carry tokens or remember secrets.

Among biometrics, face-based techniques are perceived as one of the less intrusive and more convenient recognition methods, which require a moderate degree of user collaboration and are usually well accepted by the population.

The research on face recognition has recently proposed a new method [3] based on one-class Support Vector Machines [17], which we consider very promising and worth looking at into more detail. Support Vector Machines, the learning approach originally developed by Vapnik et al. [19], represent a powerful pattern recognition method, able to deal with sample sizes of order of hundred of thousands instances [10]. They have been used till now for solving several practical problems, like isolated handwritten digit recognition [14], speaker identification [13], face detection [11] and text categorization [9].

Mainly, Support Vector Machines are used for multi-class classification, in which any new object is assigned to one of a pre-defined set of classes [4]. In one-class classification though, it is assumed that only information of the target class is available. Therefore, only examples of the target class can be used for training, and no information on the class of outliers is present. The goal is to define a boundary around the target class, such that it accepts as much of the target objects as possible, while minimizing the probability of accepting outliers objects. From a practical point of view, this means that negative examples are not required for one-class SVM training, which makes this method very interesting for the face recognition task, where the choice of the set of negative examples can have an important impact on the recognition performances, as shown in [3].

The proposed face recognition procedure is built on the work presented in [3], but also introduces a new component-based approach, in which the one-class SVM algorithm is applied to the main components of the human face (eyes, nose and mouth). The combined global and local approach performs better that the original algorithm. The paper also proposes a different feature extraction method instead of raw gray level features and presents the experimental results obtained with three different light normalization procedures.

The paper is organized as follows: the next section presents the basic operational flow of the recognition proce-

Proceedings of the 3rd Canadian Conference on Computer and Robot Vision (CRV'06)
0-7695-2542-3/06 $20.00 © 2006 IEEE

dure and outlines the novelty of the proposed approach with respect to the prior work. In the third section the experimental results are presented. The paper ends with a conclusions and future work section.

2 Method description

There are mainly three types of models for one-class classification: density estimators, reconstruction methods and boundary methods [17]. The model we use was proposed by Scholkopf et al. in [15] for the estimation of the support of high-dimensional distributions. Considering the training vectors $\mathbf{x}_i \in R^n, i = 1, \cdots, l$, where l is the cardinality of the training set, and Φ the features mapping function, Scholkopf et al. develop an algorithm that returns a model f that takes the value $+1$ in a small region (hyper sphere) containing most of the training vectors, and -1 elsewhere. For a new vector $\mathbf{x}$, the value $f(\mathbf{x})$ is determined by evaluating on which side of the hyper sphere it falls on, in the features space. To find the hyper sphere, the following quadratic programming problem has to be solved:

$$\begin{array}{ll} min_{w,\xi,\rho} & \frac{1}{2}\mathbf{w}^T\mathbf{w} - \rho + \frac{1}{\nu l}\sum_{i=1}^{l}\xi_i, \\ \text{subject to} & \mathbf{w}^T\Phi(x_i) \geq \rho - \xi_i \\ & \xi_i \geq 0. \end{array} \quad (1)$$

The meaning of the parameter ν is explained further on. The slack variables ξ_i encode the empirical error contribution (ξ_i larger than zero are penalized in the objective function). Introducing the Lagrangian [15], the dual problem is:

$$\begin{array}{ll} min_{\alpha} & \frac{1}{2}\sum_{i,j}\alpha_i\alpha_j K(x_i, x_j), \\ \text{subject to} & 0 \leq \alpha_i \leq \frac{1}{\nu l} \\ & \sum_i \alpha_i = 1. \end{array} \quad (2)$$

Considering the Gaussian kernel, which only uses the distances between the objects (training set vectors), we have:

$$K(x_i, x_j) = (\Phi(x_i) \cdot \Phi(x_j)) = e^{-\frac{\|x_i - x_j\|^2}{s^2}}. \quad (3)$$

The decision function in the features space is:

$$f(x) = sgn(\sum_{i=1}^{l}\alpha_i K(\mathbf{x}_i, \mathbf{x}) - \rho). \quad (4)$$

The parameter $\nu \in [0, 1)$ controls the trade-off between the error and the number of support vectors (the size of the solution). The parameter s of the Gaussian kernel is also important in determining the number of resulting support vectors. This approach and the SVDD approach described in [17] are identical when the Gaussian kernel with width s and $C = \frac{1}{\nu l}$ is used. s is optimized to obtain a given fraction of support vectors, while the value of C indicates the fraction of objects which should be rejected. For small values of s (smaller than the average distance between the objects in training set) all objects tend to become support vectors with $\alpha = \frac{1}{l}$, while for large s (in the order of maximum distance between the training objects) the solution approximates a hyper sphere (like in the case of a polynomial kernel with degree $n = 1$ [17]). For moderate values of s, the Lagrange multipliers α_i tend to become 0 for x_i and x_j objects that are close, and larger than 0 for most dissimilar objects, i.e., objects found at the boundary of the solution, and hence objects that eventually become support vectors.

2.1 Global approach

The enrollment and the verification modules work on feature vectors extracted from live video sequences. For each video frame the following common operations are performed:

- face detection (based on the Viola et al. [20] Haar detector the Ada Boosting algorithm for training)
- image normalization: color conversion from RGB color space to gray levels, image down scaling to a fixed size (128 × 128 pixels) by means of bilinear interpolation, image enhancement based either on simple histogram equalization or the adaptative histogram equalization method proposed by [8] combined with the light direction compensation algorithm proposed in [2]
- feature extraction based on two different algorithms: simple vectorization of the gray level image matrix (like in [3]), and the use of Fourier coefficients vectors obtained from the bi-dimensional Fourier transform of the normalized gray level images.

If $L \times L$ is the dimension of the normalized image, and $f_{x,y}$ the gray value at position (x, y) in the image, then the bi-dimensional Fourier transform coefficient value at frequency location (u, v) depends on the entire image pixels and is given by the formula:

$$F_{u,v} = \sum_{y=0}^{L-1}\sum_{x=0}^{L-1} f_{x,y} e^{-2\pi j(\frac{xu}{L} + \frac{yv}{L})}, j^2 = -1. \quad (5)$$

The Fourier spectrum of real face images is concentrated around the origin in a rhombus-like region as illustrated in Figure 1 (a). Because of the hermitian property of the bi-dimensional Fourier transform [12], it is sufficient to consider the lower spatial frequencies from two of the quadrants of the spectrum (e.g., the two upper quadrants) to obtain good recognition rates. Actually, from the mentioned quadrants, the coefficients corresponding to the spatial frequencies included in two isosceles right triangles of side N

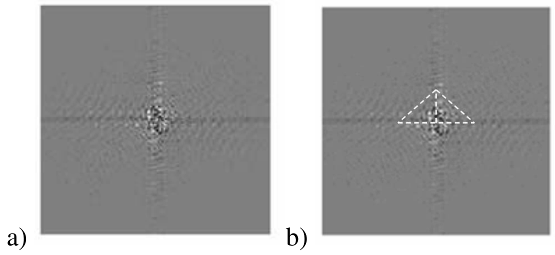

Figure 1. Fourier face spectrum (a) and frequencies selection (b).

around the origin (i.e., the upper half of the rhombus-shaped region) are sufficient.

The Fourier feature vectors consist of the concatenation of the real and imaginary part of the N lowest frequencies of the selected triangles of the spectrum (a frequency is lower if it is closer to the origin). The number N can vary from a minimum of 6 to a maximum of $\frac{1}{4}L$ frequencies, where L is the size of the quadratic Fourier matrix row (in our case $L = 128$). Note that the actual size of the Fourier feature vectors is given by $4\frac{N(N+1)}{2} - 2$, since the continuous component is not considered and the imaginary part of the Fourier coefficient corresponding to the origin is always 0. The selected Fourier coefficients are the most variable over the entire database of faces, hence they encode the most discriminant information about the user's face. If the continuous component is considered as part of the feature vectors, the inter-user variance decreases. As a consequence the Fourier coefficient corresponding to the origin is not selected. The configuration of the Fourier feature vectors was chosen after thorough tests on the ORL database [5] and our proprietary database presented in Section 3. Figure 1 (b) illustrates the coefficients selection for the real part of the Fourier spectrum, for the imaginary part the procedure is analogous.

For both gray level and Fourier methods, a one-class SVM is trained for each user, from a training set of 50 feature vectors extracted from concatenated video sequences of 300 face frames. The 50 frames included in the training set are selected either uniformly from the input sequence, or in a manner that excludes the presence of too similar frames (the degree of similarity is computed based on the cross-correlation of two frames). For the Fourier feature vectors another tested configuration consists in training two separate one-class SVMs for the real and the imaginary part of the selected Fourier spectrum.

In the verification phase, all the face frames of the input sequence are considered, each producing a score (i.e., a distance from the user's class region) matched against a threshold. A simple majority voting scheme is then used to produce the final authentication decision.

2.2 Combined global and components-based approach

In the components-based approach, separate one-class SVMs are trained for the local salient regions of the user's face and for the face itself. The concept of components-based face recognition was introduced in [7], but the approach is different: 10 face components are individually detected in the original face image, then they are assembled in a new normalized image. This image produces a single feature vector, subsequently used for training a binary linear SVM classifier.

In the proposed approach, each video frame in which all the components are detected is used to produce five different feature vectors, as shown in Figure 2, and the user's template includes 5 one-class SVM trained classifiers, and hence 5 different sets of support vectors together with their relative configuration parameters values.

The face detection module is enriched with four additional modules for eyes, mouth and nose detection. These detectors are also based on the Viola et al. Haar algorithm. The normalization and feature extraction procedures are performed but means of the algorithms mentioned above. The dimension of the resulting feature vectors is normalized to pre-defined optimal values: smaller for eyes, mouth and nose, for which fewer frequencies are sufficient to obtain good recognition rates. Again, the lowest frequencies are more representative, the information bore by the high frequencies of the spectrum being mostly related to the noise inherently present in the input images. The verification module combines the scores produced by each individual one-class SVM according to a weighted sum rule, the more accurate classifier being assigned the higher weight. Fundamentally, the weight assigned to a classifier is inversely proportional to its corresponding equal error rate:

$$w_i = \frac{\frac{1}{\sum_{i=1}^{n} \frac{1}{eer_i}}}{eer_i}, \tag{6}$$

where eer_i is the equal error rate of the classifier i, and $i \in \mathbf{0}, ..., \mathbf{n}$, with n the number of classifiers. These settings also guarantee that the sum of the weights of the classifiers is 1. The scores of the classifiers are also normalized to the $[0, 1]$ interval.

The weights are computed by testing the single classifiers performances on the ORL database. Independently from the light normalization and feature extraction configurations used, the most precise classifier is the global face classifier, closely followed by the eyes classifiers, and, at some distance, by the nose classifier and the mouth classifier. The weakness of the nose and mouth classifiers are in

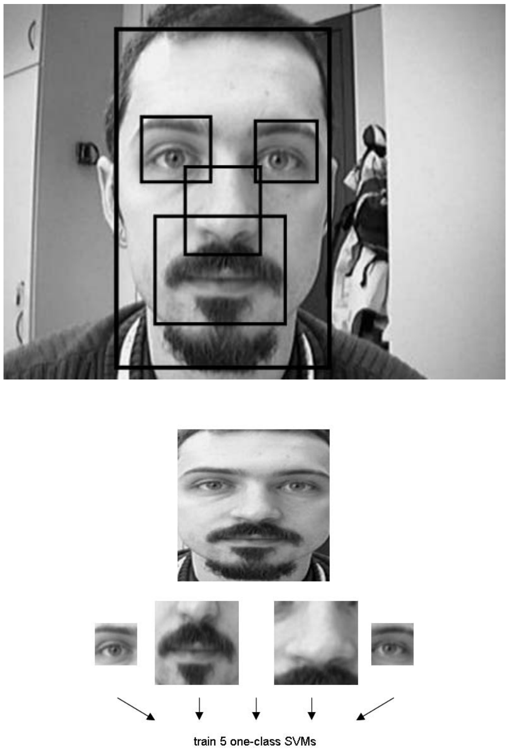

Figure 2. Components-based approach detection and training.

part due to the lesser accuracy of the corresponding detection modules.

2.3 Some practical considerations

In all the tested configurations, the value of parameter s of the kernel is set to the average distance between the objects of each training set, and this value is automatically computed during enrollment for each subject. The value of C is determined empirically for each type of feature vector and represents a trade-off between the size of the user's solution and the error rate (in particular the false rejection rate). Usually, adjusting s with $C = 1.0$ is enough for good classification results, but if a low false acceptance error is required, C must also be adjusted ($C < 1.0$) according to the training set data. After a careful algorithm tuning on the ORL database and our proprietary database, we chose to work with a fixed C of 0.3 (for gray level feature vectors) and 0.2 for Fourier feature vectors.

3 Experimental results

The proposed approaches have been tested using a proprietary database of 32 subjects with 300 video frames for enrollment and 300 video frames for test. The database was built according to the guidelines presented in [3]. The enrollment frames are obtained by the concatenation of two video sequences of the same subject, acquired in different days (two or three days apart), containing different illumination conditions, slight facial expression changes and scale modifications. The video sequences used for tests are also acquired in different days, on average one week after the enrollment session. The acquisition environment is not controlled: different day light and artificial light intensities and directions are present in both enrollment and verification sequences. The intention was to simulate real use scenarios, where is quite difficult to work in controlled conditions. Figure 3 and 4 illustrate some enrollment and test video sequences for three subjects of the database .

With all the face frames from the verification video sequences used for tests, we have 9600 client tests and 297600 impostor tests for all the configurations included in the testbeds. For the combined local and global approach, the number of tests is slightly smaller though, because we reject all the frames in which one of the components is not present. This happens mostly with the nose and mouth traits on which the Haar detector fails more easily. As a consequence, for the composed one-class SVM classifier an average of 9300 client tests and 287900 impostor tests were performed.

The results obtained with the global approach for the two feature extraction methods are summarized in Table 1. The table presents the Equal Error Rates obtained by the one-class SVM classifier for each feature extraction method (Gray Levels, Fourier and split spectrum Fourier), and each available light normalization method (simple histogram equalization - HISTEQ, Adaptive Image Enhancement - AIE and Light Direction Compensation - LDC).

	HISTEQ	**AIE and LDC**
Gray levels	10.32	6.19
Fourier	7.35	5.14
split Fourier	6.75	4.54

Table 1. Global approach: comparative EERs with different combinations of feature extraction and light normalization methods.

The application of adaptive light normalization methods improved the recognition rate, and the Fourier-based features yield better results than simple gray levels-based features. The Fourier method with split real and imaginary

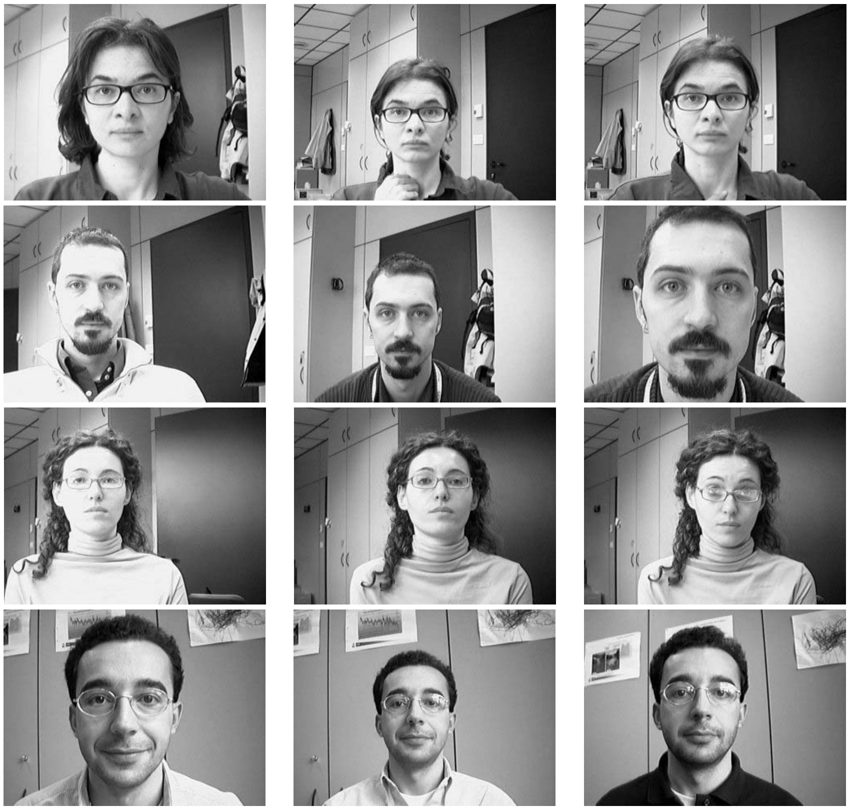

Figure 3. Examples of enrollment sequences.

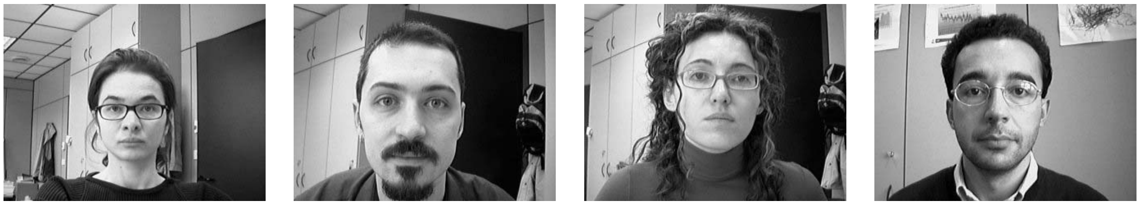

Figure 4. Examples of test sequences.

feature vectors outperforms the other features extraction approaches for the tested database. On average, the subjects' one-class SVM solutions included 16 support vectors for the gray levels approach and 10 support vectors for the Fourier-based approaches (with the adaptive light compensation). For the Fourier feature extraction method the precision of the classifier increases with the number of frequency coefficients used, reaching a maximum when the size of the considered triangle of the spectrum equals a quarter of the original row size (L) of the quadratic matrix of the image Fourier transform (i.e., $N = 32$ for $L \times L = 128 \times 128$ pixels images). Table 2 presents the Equal Error rates obtained with the best light normalization method and various feature vector sizes.

N	6	12	16	32	64
EER	6.75	5.87	5.33	5.14	5.15

Table 2. Equal Error Rate for different sizes of the Fourier feature vectors with AIE and LDC.

Note that we have tested the algorithm with the Fourier configuration suggested in [16], but considering two quadrants of the spectrum yields far better results. With the described database, the equal error rate increases on average with 4% if only the frequencies from one quadrant (i.e., one triangle) are selected from the Fourier spectrum.

As far as the combined global and local approach is concerned, we present the results obtained with the best configuration (i.e., the configuration based on Fourier feature vectors). The considered Fourier spectra sizes are $N_{face} = 32, N_{eyes} = 7, N_{mouth} = N_{nose} = 9$ and the light compensation methods are AIE and LDC. AIE and LDC have been applied both globally, to the entire face image, and locally, to each image representing the detected component. As Table 3 shows, the results are better if the light compensation algorithms are applied locally, to the single component image (i.e., light normalization is applied after detection).

	global	**local**	**local**
	LDC	**AIE**	**AIE and LDC**
Fourier	4.35	3.72	3.14
split Fourier	1.87	1.33	1.02

Table 3. Combined approach: comparative EERs with different combinations of feature extraction and light normalization methods.

We can conclude that the composed classifier and the split Fourier feature vectors with AIE and LDC normalization yield the best results on our database. With this configuration we obtain good equal error rates ($EER = 1.33\%$ and $EER = 1.02\%$), as illustrated in Figures 5 and 6.

Note that in all the tested approaches we do not use personalized thresholds. If we set different decision thresholds for each subject (as suggested in [3]) we obtain yet lower equal error rates (below 1% as also reported in [3]). Actually, this could be a logical configuration option, since the radii of the solutions computed by the one-class SVM classifier differ from one subject to another. There is only one inconvenient: during enrollment some tests should also be conducted to determine the optimal threshold distance that successfully discriminates between the true objects and the outliers. This is equivalent to introducing some negative examples in the training set, like explained in [17], thus altering the pure "one-class" nature of the approach.

The advantage of using one-class SVM is preserved though, because the number of the negative examples required for a correct tuning is quite low. In practice, we have determined that only one impostor video sequence is enough to reach a Equal Error Rate as low as 0.6% for the global approach. For five of the subjects of the database we found optimal thresholds for which the recognition rate over the entire database is 100%, and the false acceptance rate is 0. We also want to outline that with individual thresholds, the error rates of the global approach are comparable to the error rates of the component-based system with global threshold.

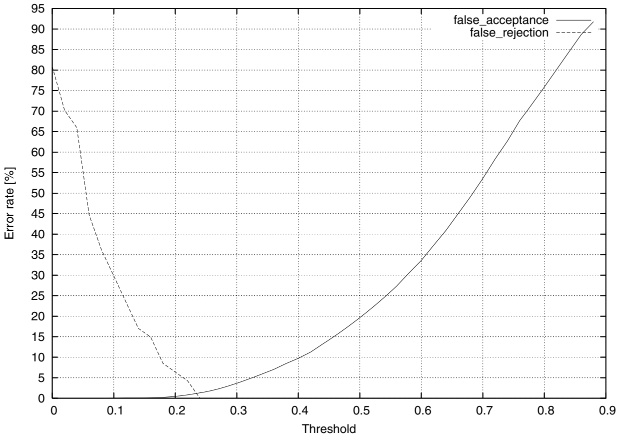

Figure 5. The best results of the composed one-class SVM classifier and AIE.

The performances of the combined approach would also increase if individual thresholds were used. Since this would require more elaborated tuning of the system, we are currently studying automatic approaches from multi modal biometrics systems [1] for the configuration of user thresh-

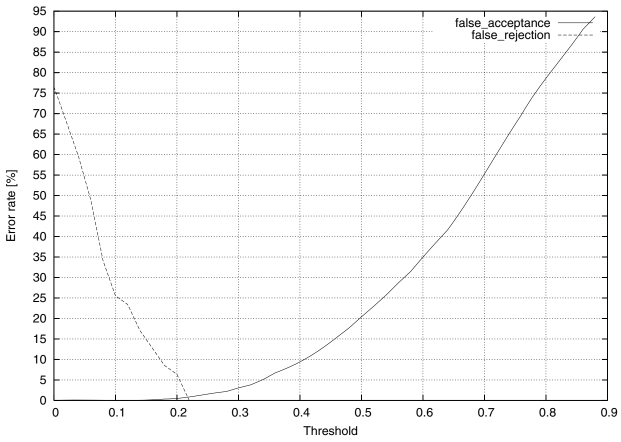

Figure 6. The best results of the composed one-class SVM classifier with AIE and LDC.

olds and also of classifiers' weights.

Finally, we outline the importance of the training procedure for SVM-based recognition. From our experiments, the subjects who have more variability in expression and (slightly) in pose in their training video sequences obtained far better results, with equal error rates around 0.2%.

4 Conclusions and future work

In the present paper we described a global technique and a components-based technique for face recognition with one-class SVM classifiers and evaluated their performance with respect to various feature extraction approaches and different light normalization methods. The input to the system in both enrollment and verification is represented by video sequences, which contain an average of 150 frames.

The global system produces one feature vector per frame and uses at least 50 feature vectors for the training of the one-class Gaussian SVM classifier. The component-based system detects four facial components and produces four sets of feature vectors, each of which is used for the training of a separate one-class SVM. The global and component-based approaches are then combined in a unique recognition system.

We tested different configurations of the proposed system with different feature extraction and light normalization algorithms, varying from simple histogram equalization to adaptive techniques. The combined system with Fourier features, adaptive image enhancement and light direction compensation outperforms the other approaches, reaching equal error rates as low as 1.02%. The global approach performances though become comparable to the results obtained by the combined approach if individual decision thresholds are introduced. The average equal error rates are below 1% in this case.

We are also investigating the use of Fourier Mellin features with the one-class SVM classier. Preliminary error rates are still high at the moment (around 7%), following the feature extraction approach presented in [6]. Simultaneously, we are trying to increase the accuracy of the mouth and nose detectors and to implement new light normalization procedures.

References

[1] A. J. amd A. Ross. Learning user-specific parameters in a multibiometric system, 2002. Proceedings of the International Conference on Image Processing (ICIP).

[2] A.Tankus and Y.Yeshurun. Convexity-based visual camouflage breaking. *Computer Vision and Image Understanding*, 84(3):234–778, 2001.

[3] M. Bicego, , E. Grosso, and M. Tistarelli. Face authentication using one-class support vector machines, 2005. Springer Verlag.

[4] C. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2:121–167, 1998.

[5] A. L. Cambridge. The orl database of faces, 1994. http://www.cl.cam.ac.uk/Research/.

[6] S. Derrode and F. Ghorbel. Robust and efficient fourier-mellin transform approximations for gray-level image reconstruction and complete invariant description. *Computer Vision and Image Understanding*, 83(1):57 –78, 2001.

[7] B. Heisele, P. Ho, and T. Poggio. Face recognition with support vector machines: Global versus component-based approach. *Proceedings of IEEE Int. Conf. on Computer Vision and Pattern Recognition*, 2(1):688–694, 2001.

[8] L. Jin, S. Satoh, and M. Sakauchi. A novel adaptive image enhancement algorithm for face detection. *Proceedings of the 17th International Conference on Pattern Recognition 2004*, 14(1):234–778, 2004.

[9] T. Joachims. Text categorization with support vector machines, 1997. Technical Report, LS VIII Number 23, Unoversity of Dortmund.

[10] J. S.-T. N. Cristianini. An itroduction to support vector machines and other kernel-based learning methods, 2000. Cambridge University Press, ISBN 0 521 78019 5.

[11] E. Osuna, R. Freund, and F. Girosi. Training support vector machines: an application to face detection, 1998. in IEEE Conference on Computer Vision and Pattern Recognition.

[12] W. Pratt. Digital image processing, 2001. John Wiley & Sons, Inc.

[13] M. Schmidt. Identifying speaker with support vector networs, 1996. in Proceedings of Interface '96, Sydney.

[14] B. Scholkopf, B. Burges, and V. Vapnik. Incorporating invariances in support vector learning machines. *Artificial Neural Networks, ICANN '96*, 1112:47–52, 1995.

[15] B. Scholkopf, J. C. Platt, J. Shawe-Taylor, and A. J. Smola. Estimating the support of a high dimensional distribution. *Neural Computation, MIT*, (13):1443–1471, 2001.

[16] H. Spies and I. Ricketts. Face recognition in fourier space, 2000. In the Proceedings of the Vision Interface Conference (VI 2000), Montreal, Canada.

[17] D. Tax. One-class classification: Concept learning in the absence of counter examples, 2001. Ph.D. Thesis, University of Delft.

[18] U. Uludag, S. Prabhakar, and A. Jain. Biometrics cryptosystems: Issues and challenges. *Proceedings of the IEEE Special issue on Multimedia Security for Digital Rights Management*, 92(6):948–960, 2004.

[19] V. Vapnik. The nature of statistical learning theory, 1995. Springer Verlag.

[20] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. *Proceedings of IEEE Int. Conf. on Computer Vision and Pattern Recognition*, 1:511–518, 2001.