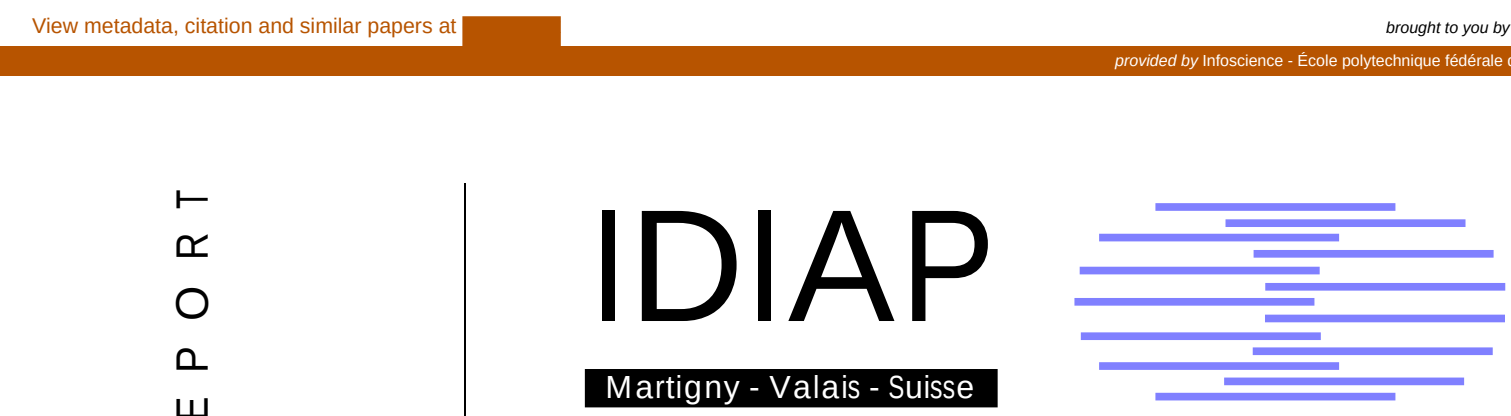
IDIAP RESEARCH REPORT



Multi-Modal Data Fusion for Person Authentication using SVM

Souheil Ben-Yacoub [a]

IDIAP–RR 98-07

SUBMITTED FOR PUBLICATION

Dalle Molle Institute for Perceptual Artificial Intelligence • P.O.Box 592 • Martigny • Valais • Switzerland

phone +41 – 27 – 721 77 11
fax +41 – 27 – 721 77 12
e-mail secretariat@idiap.ch
internet http://www.idiap.ch

[a] IDIAP, CP592 1920 Martigny, Switzerland. email: sby@idiap.ch

IDIAP Research Report 98-07

Multi-Modal Data Fusion for Person Authentication using SVM

Souheil Ben-Yacoub

SUBMITTED FOR PUBLICATION

Abstract. In the context of multi-modal person authentication, a set of experts (face recognizer, speaker recognizer, etc.) give their opinion about the identity of an individual. The opinions of the experts can be combined to form a final decision (rejecting or accepting the claim). We show that the final decision is a binary classification problem and propose to solve it by a Support Vector Machine (SVM). We compare our approach with other proposed methods for an identical verification task and show that it leads to considerably higher performance.

Multi-Modal Data Fusion for Person Authentication using SVM

Abstract

In the context of multi-modal person authentication, a set of experts (face recognizer, speaker recognizer, etc.) give their opinion about the identity of an individual. The opinions of the experts can be combined to form a final decision (rejecting or accepting the claim). We show that the final decision is a binary classification problem and propose to solve it by a Support Vector Machine (SVM). We compare our approach with other proposed methods for an identical verification task and show that it leads to considerably higher performance.

1 Introduction

The authentication of persons is a complex task with high performance and robustness requirements when used in practical applications. Identity verification by means of visual information (like face recognition) or audio information (speaker recognition) are very mature areas. When used separately, each modality reaches some limitations or shows a lack of robustness. Increasing the reliability of the person authentication process can be achieved through a combination of the different modalities, yielding a multi-modal authentication process.

The critical step in this process is the "fusion" task which combines different expert opinions in order to make a final decision. The main difficulty in this task is due to the different behavior of the experts that, for example, can give opposite opinions.

We present the "fusion" task as a binary-classifier problem (the two classes beeing the impostor and the client class). We propose to use as an efficient binary-classifier the support vector machine or SVM which is gaining more attention [16] and which has shown higher performance than alternative methods. Another advantage of the SVM is that very few parameters need to be fixed by the user, almost all parameters are determined internally by the algorithm.

The proposed method will be compared to another approach [2] using the same data and the same test protocol.

2 Person authentication

The authentication and identification of persons are important tasks for a wide field of applications (law enforcement, secure buildings access, server access, etc ...). Authentication (or verification) and identification can be defined as follow:

- Verification is concerned with validating the claimed identity of a person from an open set of persons and to either accept or reject the claim.
- Identification is concerned with determining that person from a closed set, whose features best match the features of the person to identify. It assumes that only enrolled persons will access the system.

The proposed approach focuses on the *authentication* problem. Different biometrics features can be used to identify a person (fingerprints, iris, voice, etc ...). Vocal and facial modalities for person identification are very interesting since they are none invasive methods. They are easily accepted by users and they do not require expensive specialized hardware (a camera, microphone and a computer

are enough). On the opposite side, fingerprints and iris require very specialized hardware and are not well accepted by users. Moreover face recognition and voice identification are very common task for humans, hence monitoring an identification system based on these feature is an easy task.

In the work we are presenting we are dealing with multi-modal biometric features: voice and face recognition. Using different modalities increases the reliability of the identification process. The identification system relies on 3 modules:

- The face identification expert which delivers a score ranging from 0 to 1 (0 reject person and 1 accept person).
- The voice identification expert which delivers a score ranging from 0 to 1.
- The supervisor which takes as input the scores delivered by the experts and given these scores must take the final decision (i.e. reject or accept person).

The proposed approach focuses on the last module (the supervisor). Although the presented work combines only two modalities, it can be extended to any number of modalities.

2.1 Face identification expert

The problem of face identification has been addressed by different researcher and with different methods. For a complete survey and comparison of different approaches see [4, 17].

The face identification expert which generated the data (i.e scores) we are using, is based on *Elastic Graph Matching* [11]. The method consists in positioning a regular grid over a face and computing features on the graph's nodes. The features that were used in this case are modulus of complex Gabor responses (with 6 orientations and 3 resolutions) [6]. A database of graph-features is stored for each client. During a request, the graph is positioned on the user's face and a set of Gabor features are computed. The resulting graph and features are compared to the stored models. To measure the difference between the user and the claimed identity, the deformation of the graph, in order to achieve the matching, as well as the distance between the features are taken into account. In the final stage the expert delivers a score ranging from 0 to 1 which reflects its "soft opinion" on the request made by the user.

Extensive experiments have been made to assess the accuracy and reliability of this method, for more details see [2].

2.2 Speech Identification Expert

Speech is a very convenient feature to identify people, it is a 1-D signal which is unique to every person. The recognition and identification using speech is a very active research area [9, 12].

One of the most used feature in speech is the LPC-C (Linear Predictive Coefficients-Cepstrum) [7]. In our experiments, during the training, the users are asked to pronounce a sequence of digits from 0 to 9. The input signal is segmented into phonemes and LPC-C vectors are computed. A Hidden Markov Model [14] is used to model the digit-user couple. The LPC-C vectors distribution is modeled by a Gaussian (parameters are estimated during training) for each digit and for each person [8]. Another Hidden Markov Model is computed to represent the "impostor model" using a different database of a large number of persons: a digit-world model is obtained for each digit.

During the test sequence, the user claims an identity Id and pronounces the sequence of digits 0 to 9. For each digit, the likelihood of the sample being produced by the corresponding HMM model is estimated using the Viterbi algorithm. The similarity measure is the summed log-likelihood over the digits and normalized by the number of LPP-C vectors in the sequence:

$$S_{Id} = \sum_{i=0}^{9} \frac{Log(L_{Id}i)}{N}$$

where $Log(L_{Id}i)$ is the log-likelihood of the user Id model for digit i and N the number of LPC-C vectors in the sequence.

In order to normalize the similarity measure, the likelihood of the test sequence being generated by the "impostor model" w is computed in the same way as for a client model:

$$S_w = \sum_{i=0}^{9} \frac{Log(L_w i)}{N}$$

The final normalized similarity measure (or score) S that will be used by the supervisor fusion algorithm is:

$$S = f(S_{Id} - S_w), \text{where} f(x) = \frac{1}{1 + e^{-x}}$$

which maps the difference of the two log-likelihoods onto a sigmoid function.

2.3 Supervisor

A classical scenario of a multi-modal identity claim is shown on Figure 1.

A potential **User** claims an **Identity** and the different data associated to the user are acquired (face image, voice, profile image etc...). The number of different modalities used is **N**. Each modality expert has an access to a database of models associated to each client. The experts compare the data associated to the user and the data associated to the claimed **Identity**: they produce a score in the range [0..1] which reflects the opinion of the expert on the match between the **User** and the claimed **Identity**.

In the final stage, the **Supervisor** combines the different opinions (or scores) and makes a final decision: accept or reject the **User**.

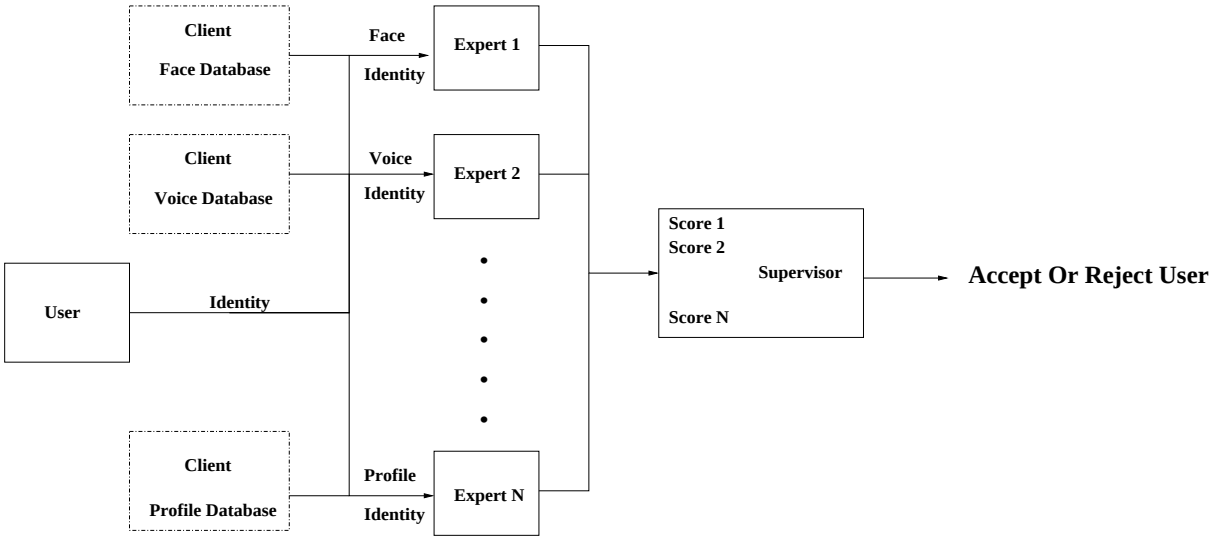


Figure 1: Multi-Modal access request

The problem of the design of an efficient supervisor can be formulated as a binary classifier problem. Given two classes,the impostor class (denoted as "−1") and the client class (denoted as "+1"), and given the expert scores associated to a member of each class, find the optimal function f that separates the two classes. This is equivalent to the formulation of the binary classifier as stated in Section 3. Formally, the problem is formulated as follow:

$$Score_i = Expert_i(\textbf{User}), \forall i \in \{1..N\}, \text{ Find } f \text{ verifying :}$$

$$\textbf{User} \in +\textbf{1}, \quad f(Score_1, Score_2, ..., Score_N) = +1$$
$$\textbf{User} \in -\textbf{1}, \quad f(Score_1, Score_2, ..., Score_N) = -1$$

We propose to use the Support Vector Machine to find the optimal function f to separate the two classes. One of the parameters of the SVM algorithm is the choice of the kernel that will be used to define the decision surface (see Section 3).

3 Support Vector Machines

The Support Vector Machines (SVM) is a new technique in the field of statistical learning theory. It is based on the principle of *Structural Risk Minimization* [16]. Classical learning approaches are designed to minimize the empirical risk (i.e error on a training set) and therefore follow the *Empirical Risk Minimization* principle. The SRM principle states that better generalization capabilities are achieved through a minimization of the bound on the generalization error.

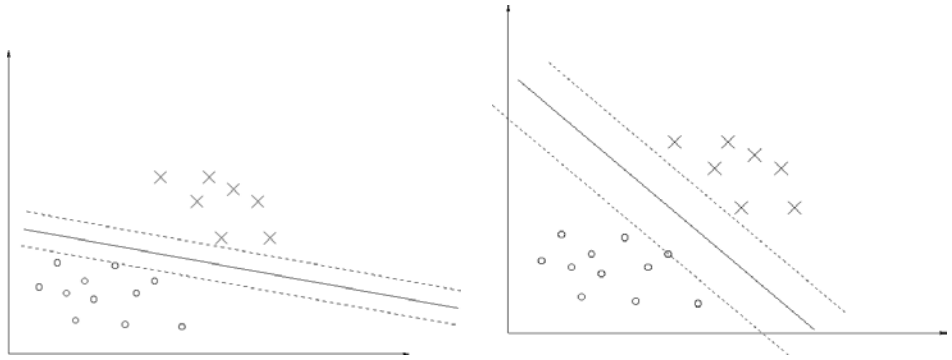
We assume that we have a data set $\mathcal{D}$ of M points in a n dimensional space belonging to two different classes **+1** and **-1**:

$$\mathcal{D} = \{(X_i, y_i) | i \in \{1..M\}, X_i \in \mathbb{R}^n, y_i \in \{+1, -1\}\}$$

A binary classifier should find a function f that maps the points from their data space to their label space:

$$\begin{array}{rcl} f : \mathbb{R}^n & \longrightarrow & \{+1, -1\} \\ X_i & \longmapsto & y_i \end{array}$$

For the sake of simplicity, we assume that the data space is $\mathbb{R}^2$ and that a hyperplane (i.e affine function) separates the data. There are actually an infinite number of hyperplanes that could partition the data into two sets. According to the SRM principle, there will be just one optimal hyperplane: the hyperplane with the maximal margin[1]. Figure 2 illustrates the concept of optimal separating hyperplane. The margin in this case is the distance separating the dashed lines.



(a) Separating Hyperplane (b) Optimal Separating Hyperplane

Figure 2: Examples of separating hyperplanes: Margin of (b) is larger.

In Figure 2 (b), the separating hyperplane has a larger margin than the one in Figure 2 (a). According to the SRM principle, the hyperplane with the largest margin minimizes the misclassification error.

[1] The margin is defined as the sum of distances from the hyperplane to the closest points of the two classes

It has been shown [16] that the optimal separating hyperplane is expressed as:

$$f(x) \quad = \quad sign(\sum_{i \in SV} \alpha_i y_i K(X_i, x) + b) \tag{1}$$

where SV is a subset of the data points (this subset contains the points with $\alpha_i > 0$), K(x,y) is a positive definite symmetric function, α_i are the solutions of the following Quadratic Programming (QP) problem:

$$\begin{cases} \min_{\mathcal{A}} W(\mathcal{A}) = -\mathcal{A}^t I + \frac{1}{2}\mathcal{A}^t D \mathcal{A} \\ \text{with the constraints:} \\ \sum_i \alpha_i y_i = 0 \text{ and } \alpha_i \geq 0 \\ \\ \text{where:} \\ (i,j) \in [1..M] \times [1..M] \\ (\mathcal{A})_i = \alpha_i \\ (I)_i = 1 \\ (D)_{ij} = y_i y_j K(X_i, X_j) \end{cases}$$

The SVM is a SRM-based binary classifier: it finds the optimal hyperplane separating the data into two classes by solving a QP problem. This result is valid regardless of the dimension of the data space.

We have so far discussed the case of linearly separable data. If no hyperplane can be found to separate the data, a non-linear mapping function is then needed. The data will be mapped non-linearly in a high-dimensional space and the optimal hyperplane is computed in the high-dimensional space. This is performed through specific kernel functions $K(x,y)$ [16] which actually define the nature of the decision surface that will separate the data.

The kernel functions must satisfy some constraints in order to be applicable (Mercer's conditions, see [16]). Some possible kernel functions are already identified (we assume (x,y) $\in \mathbb{R}^n \times \mathbb{R}^n$) :

- $K(x,y) = x^t y$ defines a linear classifier.
- $K(x,y) = (x^t y + 1)^p$ with $p \in \mathcal{N}$, this defines a polynomial decision surface of degree p.
- $K(x,y) = tanh(ax^t y + b)$ with $(a,b) \in \mathbb{R}^2$, this defines a multi-layer perceptron classifier.
- $K(x,y) = e^{-\frac{||x-y||^2}{\sigma^2}}$ is equivalent to a RBF classifier.

From Equation (1), we can see that the number of parameters that must be determined is always equal to the number of data points. The user-defined parameters are the type of kernel (Gaussian, polynomial etc...) and the parameters associated to the corresponding kernel (maximum of 2 parameters for MLP kernel). This has a major advantage from the practical point of view, since no special knowledge or extensive tests are needed to fix the values of the parameters.

The computational complexity of the SVM during the training depends on the number of data points rather than on their dimensionality. The number of computation steps is $O(n^3)$ [15] where n is the number of data points. At run time the classification step of SVM is a simple weighted sum.

The SVM algorithm requires a QP solver: we used the DONLP2[2] package.

4 Results and Experiments

The data we used for the experiments were extracted from the M2VTS multi-modal database [13]. In a first set of experiments, we use two modalities: face and text-dependent voice identification. The experts described in section 2.2 and 2.1 were used to provide the scores for the fusion supervisor.

[2]Provided by prof. Spellucci, http://plato.la.asu.edu/donlp2.htm

The database of scores consists in 37 different persons. Each person is present in 4 different shots. There is a total of 10656 identity claims (client and impostor claims). We used the same protocol as in [2]: a rotational scheme on users is performed. All the scores related to a user X (there are 576 of such score: 288 client and 288 impostor scores) are removed from the database in order to be used as test scores. The remaining scores (namely 10080 scores) are used to train the supervisor algorithm. Since we have 37 different users, the supervisor algorithm is tested on 37 different couples training-test sets.

There are two types of errors that can occur during a verification procedure:

- False acceptance (FA): an impostor is recognized as a client.
- False rejection (FR): a client is classified as impostor.

In order to evaluate the different identification algorithms the FA-rate and the FR-rate are used. The FA-rate is defined as the ratio between the total number of false-acceptances and the total number of impostor accesses. In the same manner, the FR-rate is the ratio between the total number of false-rejections and the total number of client accesses. The total error rate (TE) is defined as the sum of the FA-rate and FR-rate. The results achieved by the SVM are displayed in Table 1

SVM-Kernel	Number of Errors	FA	FR	TE
linear	8	$7.5\ 10^{-4}$	0	$7.5\ 10^{-4}$
polynomial (p=2)	24	$2.2\ 10^{-3}$	0	$2.25\ 10^{-3}$
RBF (σ^2=0.1)	12	$1.12\ 10^{-3}$	0	$1.12\ 10^{-3}$
MLP (a=0.5, b=-1)	16	$1.5\ 10^{-3}$	0	$1.5\ 10^{-3}$

Table 1: SVM classification rates

Using the same data and the same protocol as in [2] enables us to compare the SVM-based classifier to other approaches [5] (Bayesian supervisor, speech expert alone, face expert alone, arithmetic mean supervisor). The results are shown in Table 2

Supervisor	FA (%)	FR (%)	TE (%)
face	3.6	7.4	11.0
speech	6.7	0.0	6.7
arithmetic mean	1.2	2.1	3.3
Bayesian conciliation	0.54	0.0	0.54
Linear-SVM	0.07	0.0	0.07
polynomial-SVM	0.21	0.0	0.21
RBF-SVM	0.12	0.0	0.12
MLP-SVM	0.15	0.0	0.15

Table 2: Comparative results

The results show clearly that the SVM outperforms the other approaches and that it leads to a total error rate which is in the best case 8 times smaller than the Bayesian conciliation algorithm, and 2 times smaller in the worst case.

In another series of experiments we used 4 modalities. The two new experts are also based on speech and face, but different techniques are used. The face authentication expert is based on morphological filters [10], while speech expert is based on text-independent speaker recognition [3]. The results in this case did not change significantly. This can be explained by the fact that the two new experts are not as skillful as the first two experts and hence no real improvement of the results can be achieved.

5 Conclusion

We proposed to approach the problem of multi-modal person authentication as a binary-classifier problem. We suggested to use the Support Vector Machine, a new binary classification paradigm based on Structural Risk Minimization, in order to find the optimal decision surface. A series of experiments were performed using two modalities (face and voice). The results of the experiments show that SVM outperforms the Bayesian conciliation fusion algorithm and that it reduces the total error rate by a factor ranging from 2 to 8 (depending on the type of kernel used for the SVM). One of the major advantage also of SVM is that, unlike Bayesian conciliation, no assumption is made on the data distribution. In the case of Bayesian conciliation [1], the logarithm of the misidentification score is assumed to have a normal distribution. This strong constraint is not always verified and may explain why Bayesian conciliation achieves a lower performance than SVM. The proposed approach can be applied with any number of experts with constant computational cost since the complexity depends on the number of data points rather than on their dimensionality.

The promising results we obtained indicates that the SVM can be a good solution in the framework of multi-modal data fusion. Further experiments on a larger database (300 people) will focus on robustness with respect to missing data.

Acknowledgment

This work was done in the framework of the M2VTS project. The author thanks J. Lüttin for his comments and useful discussions.

References

[1] E. Bigün, J. Bigün, B. Duc, and S. Fischer. Expert conciliation for multi modal person authentication systems by bayesian statistics. In *Proceedings of the First International Conference on Audio- and Video-based Biometric Person Authentication (AVBPA'97)*, Lecture Notes in Computer Science, pages 291–300. Springer Verlag, 1997.

[2] J. Bigün, B. Duc, F. Smeraldi, S. Fischer, and A. Makarov. Multi-modal person authentication. In *Face Recognition: From Theory to Applications (NATO-ASI Workshop)*. Springer-Verlag, 1997. (to be published).

[3] F. Bimbot and L. Mathan. Second order statistical measures for text independent speaker verification. In *ESCA Workshop on Automatic Speaker Recognition*, pages 51–54, 1994.

[4] R. Chellappa, C.L Wilson, and C.S Barnes. Human and machine recognition of faces: A survey. Technical Report CAR-TR-731, University of Maryland, USA, 1994.

[5] Benoît Duc, Gilbert Maître, Stefan Fischer, and Josef Bigün. Person authentication by fusing face and speech information. In *Proceedings of the First International Conference on Audio- and Video-based Biometric Person Authentication (AVBPA'97)*, Lecture Notes in Computer Science, pages 311–318. Springer Verlag, 1997.

[6] S. Fischer, B. Duc, and J. Bigün. Face recognition with gabor phase and dynamic link matching for multi-modal identification. Technical Report LTS 96.04, EPFL, Lausanne, 1996.

[7] S. Furui. Cepstral analysis technique for automatic speaker verification. *IEEE Trans. on Acoustics, Speech, Signal Processing*, 29(2):254–272, 1981.

[8] D. Genoud, F. Bimbot, G. Gravier, and G. Chollet. Combining methods to improve speaker verification decision. In *Proceedings of the Fourth Intern. Conf. on Spoken Language Processing*, pages 1756–1759, October 1996.

[9] D. Gibbon, R. Moore, and R. Winski, editors. *Handbook of Standards and Resources for Spoken Language Systems*. Mouton de Gruyter, 1997.

[10] C. Kotropoulos, I. Pitas, Stefan Fischer, and Benoît Duc. Face authentication using morphological dynamic link matching. In *Proceedings of the First International Conference on Audio- and Video-based Biometric Person Authentication (AVBPA'97)*, Lecture Notes in Computer Science, pages 169 –176. Springer Verlag, 1997.

[11] M. Lades, J. Buhmann J. C. Vorbrüggen, 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, March 1993.

[12] C. Lee, F. Soong, and K. Paliwal, editors. *Automatic Speech and Speaker Recognition*. Kluwer Academic Publishers, 1995.

[13] S. Pigeon. The m2vts multimodal face database (release 1.00). *CEC ACTS/M2VTS Deliveralble AC102/UCL/WP1/DS/P/161*, 1996. http://www.tele.ucl.ac.be.

[14] L.R Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[15] P. Spellucci. Personal communication, 1998.

[16] V. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, 1995.

[17] J. Zhang, Y. Yan, and M. Lades. Face recognition: Eigenfaces, elastic matching, and neural nets. *Proceedings of IEEE*, 85:1422–1435, 1997.