PERGAMON

Available at
www.ElsevierComputerScience.com
POWERED BY SCIENCE DIRECT

Pattern Recognition 37 (2004) 2281–2291

PATTERN RECOGNITION
THE JOURNAL OF THE PATTERN RECOGNITION SOCIETY

www.elsevier.com/locate/patcog

Similarity-based classification of sequences using hidden Markov models

Manuele Bicego[a,*], Vittorio Murino[a], Mário A.T. Figueiredo[b]

[a]*Dipartimento di Informatica, Università di Verona, Ca' Vignal 2, Strada Le Grazie 15, 37134 Verona, Italy*
[b]*Instituto de Telecomunicações, Instituto Superior Técnico, 1049-001 Lisboa, Portugal*

Received 10 December 2002; accepted 12 April 2004

Abstract

Hidden Markov models (HMM) are a widely used tool for sequence modelling. In the sequence classification case, the standard approach consists of training one HMM for each class and then using a standard Bayesian classification rule. In this paper, we introduce a novel classification scheme for sequences based on HMMs, which is obtained by extending the recently proposed similarity-based classification paradigm to HMM-based classification. In this approach, each object is described by the vector of its similarities with respect to a predetermined set of other objects, where these similarities are supported by HMMs. A central problem is the high dimensionality of resulting space, and, to deal with it, three alternatives are investigated. Synthetic and real experiments show that the similarity-based approach outperforms standard HMM classification schemes.

Keywords: Hidden Markov models; Distance-based classification; 2D shape recognition; Face classification; Maximum-likelihood classification; Matching pursuit

1. Introduction

The analysis of sequential data is an interesting and important research area. Probabilistic modelling and classification is intrinsically more difficult when each observation is a sequence, compared to the standard scenario where each observation is a set (vector) of features. In fact, since the length of the sequences may vary, it is not possible to directly use standard pattern recognition techniques. Moreover, sequence classification problems usually involve very large data sets.

Hidden Markov models (HMMs) are commonly employed probabilistic models of sequential data [1]. HMMs can be viewed as stochastic generalizations of finite-state automata, when both the transitions between states and the generation of output symbols are governed by probabilistic mechanisms [1]. Although the basic theory and inference tools were developed in the late 1960s [2,3], HMMs have only been extensively applied in the last decade. Speech recognition [1], DNA and protein modelling [4,5], handwritten character recognition [6], gesture recognition [7], and behavior analysis and synthesis [8] are examples of problems for which HMMs have been exploited.

The standard HMM-based approach to sequence classification consists in training one HMM for each class, which are subsequently used as class-conditional densities in a standard Bayes classification paradigm. For example, assuming a priori equiprobable classes, an unknown sequence

* Corresponding author. Present address: DEIR, University of Sassari, Via Sardegna, 58 07100 Sassari - Italy.
E-mail addresses: bicego@uniss.it, bicego@sci.univr.it (M. Bicego), vittorio.murino@univr.it (V. Murino), mtf@lx.it.pt (M.A.T. Figueiredo).

0031-3203/$30.00
doi:10.1016/j.patcog.2004.04.005

is classified into the class whose model shows the highest probability (*likelihood*) of having generated this sequence (this is the well-known *maximum-likelihood* (ML) classification rule).

In this paper, an alternative classification scheme is proposed, by extending the similarity-based paradigm [9–14] to HMM-based classification. This paradigm, which has been introduced recently, differs from typical pattern recognition approaches where objects to be classified are represented by sets (vectors) of features. In the similarity-based paradigm, objects are described using pairwise (dis)similarities, i.e. distances from other objects in the data set. In this way, objects are not constrained to be explicitly represented in a feature space, and all that is necessary is a way to compute (dis)similarities between pairs of objects. The goal is then to learn a classifier only from these relational data.

The literature on similarity-based classification is not vast [9–14] (a brief review is given in Section 2.1). The general idea behind all these approaches is basically the same: given a set of pairwise dissimilarity values, a new representation space can be built, in which each object is described by these values. In Ref. [13], a simple synthetic experiment shows that a complex problem in a 2D space (requiring a quadratic classifier to achieve almost correct separation), becomes a linearly separable problem in a dissimilarity space.

In this paper, we extend this dissimilarity-based classification paradigm to HMM-based sequences classification problems. We propose to build a similarity[1] space, representing each object (sequence) by the vector of its similarities with respect to a predetermined set of objects (this can be the whole data set, in the simplest approach), called the *representatives set*; the classification is then performed in this new representation space. The similarities are derived by considering the likelihood $P(\mathbf{O}|\boldsymbol{\lambda})$ as a measure of the *similarity* between the sequence $\mathbf{O}$ and the HMM specified by the set of parameters $\boldsymbol{\lambda}$. This similarity measure was previously used in sequence clustering applications [15,16].

The similarity-based classification paradigm seems to be particularly well suited to HMMs, as it can be seen as a natural extension of the standard HMM classification scheme. Specifically, the standard ML approach assigns an unknown sequence $\mathbf{O}$ to the class whose model shows the highest likelihood. To do so, the likelihoods of $\mathbf{O}$ with respect to the HMMs of all classes are evaluated, each stating a *likelihood-based* measure of the similarity between that class and the observed sequence. In other words, HMMs are used to compute *similarities* between sequences and classes, with each class being represented by a single HMM. Subsequently, only the maximum of these values is used to take the classification decision. In the similarity-based approach, the classification decision is taken using the *whole* set of similarities between each observed sequence and all the other sequences. We will show that this strategy results in a substantial improvement in the classification performance, compared to standard HMM-based approaches. Moreover, with the use of HMMs and the similarity representation, the problem of classification of sequences is reduced to a more standard classification task (where each object is described by a fixed-length feature vector), for which arbitrarily sophisticated techniques can be used, allowing to increase even more the classification performance.

The proposed approach was successfully tested on both synthetic and real data, involving 2D shape recognition and face recognition problems. In comparison with the standard HMM-based ML classification approach, our method showed a significant performance improvement, confirming all the potential of the similarity-based classification approach.

The main problem of the similarity-based approach, of particular relevance in practical applications, is the high dimensionality of the resulting similarity space. Actually, in the basic approach, this dimensionality is equal to the cardinality of the whole training data set, possibly leading to a huge computational burden. In the literature, two types of solutions of this problem could be identified, summarized in Section 2.2. In this paper, three methods to face this problem are proposed. The first one aims at removing redundancy from the data by applying linear dimensionality reduction techniques, such as Fisher discriminant analysis (FDA) [17] and principal component analysis (PCA) [19]. The second proposed method is based on a greedy strategy known as *matching pursuit* [20], which selects a subset of representatives based on which the similarity values are computed. These two approaches are very general, and can be applied in all distance-based classification contexts. The third proposed approach is more specific to the HMM case, and is based on a simple adaptation of the similarity-based classification approach to the standard HMM learning procedure. All these approaches were experimentally evaluated, confirming the discriminative power of the similarity space, even when the dimensionality is reduced to more manageable numbers.

Summarizing, the main contribution of this paper is the introduction of the similarity-based recognition paradigm in an HMM context, resulting in a significant performance improvement with respect to standard HMM-based classification. The mapping to the similarity space proposed in this paper allows us to reduce complex problems of sequence classification to a more standard point classification problem, for which arbitrarily sophisticated techniques could be used. From the point of view of similarity-based recognition, we propose two different approaches for dealing with the high dimensionality of the similarity space, which is one of the main problems of the method. First, the potential of linear reduction techniques, as PCA and FDA, is exploited, showing that they are able to reduce the curse of dimensionality impact on the classification process. Second, we address the choice of a set of appropriate representatives using the matching

[1] Note that we refer indifferently to similarity or dissimilarity.

pursuit algorithm, which proves to be a robust and effective approach.

The rest of the paper is organized as follows. In Section 2, the state of the art related to the similarity-based classification and to the dimensionality issue is summarized. In Section 3, HMMs are introduced, together with the standard classification scheme. The proposed strategy is described in Section 4, and Section 5 reports the experiments and the related results are discussed. Finally, in Section 6, conclusions are drawn and future perspectives are envisaged.

2. State of the art

2.1. Similarity-based classification

The literature on similarity-based classification is not vast. The approach seems to have been firstly introduced by Jain and Zongker [9], who have obtained a dissimilarity measure, based on deformable templates, for the hand-written digit recognition problem. A multidimensional scaling approach was then used to project this dissimilarity space onto a low-dimensional space, where a 1-nearest-neighbor (1-NN) classifier was employed to classify new objects. In Ref. [10], Graepel et al. investigate the problem of learning a classifier based on data represented in terms of their pairwise proximities, using an approach based on Vapnik's structural risk minimization [21]. Jacobs and Weinshall [11] studied the use of distance-based classification with non-metric distance functions (i.e. that do not satisfy the triangle inequality). Duin and Pekalska are very active researchers in this area[2] having recently produced several papers [12–14]. Motivation and basic features of similarity-based methods were first described in Ref. [12]: it was shown, by experiments in two real applications, that a Bayesian classifier (the RLNC—regularized linear normal density-based classifier) in the dissimilarity space outperforms the NN rule. These aspects were more thoroughly investigated in Ref. [14], where other classifiers in the dissimilarity space were studied, namely on digit recognition and bioinformatics problems. Finally, in Ref. [13], a generalized kernel approach was introduced, dealing with classification aspects of the dissimilarity kernels.

2.2. The dimensionality issue

The main problem of the similarity-based approach, of particular relevance in practical applications, is the high dimensionality of the resulting similarity space. Two types of solutions have been proposed in order to address this problem. The first consists of building the similarity space using all available patterns, and subsequently applying some standard dimensionality reduction technique. One example of this kind of approach is the multidimensional scaling method used in Ref. [9]. Another recent example is presented in Ref. [22], where a reduction of the dimensionality of the dissimilarity space is obtained by a modified multidimensional scaling scheme, able to reduce the computational burden and to allow generalization to new data. The second type of solution works by directly choosing a small set of representatives. An example of this type of solution can be found in Ref. [14], where random selection, *most-dissimilar* rule and the *condensed* NN (CNN) rule were employed. Other examples can also be found in Ref. [11], where a new type of CNN method is proposed, or in a recent paper [23], where a greedy approach is proposed, able to find prototypes encoding the principal components of the similarity space.

[2] See http://www.ph.tn.tudelft.nl/Research/neural/index.html

3. Hidden Markov models

A discrete-time HMM is a probabilistic model that describes a random sequence $\mathbf{O} = O_1, O_2, ..., O_T$ as being an indirect observation of an underlying (hidden) random sequence $\mathbf{Q} = Q_1, Q_2, ..., Q_T$, where this hidden process is Markovian, even though the observed process may not be. Due to lack of space, HMM theory will not be covered in detail here; for a comprehensive tutorial, see Ref. [1]. Basically, an HMM $\boldsymbol{\lambda}$ is a 4-tuple $\boldsymbol{\lambda} = (S, \mathbf{A}, \boldsymbol{\pi}, \mathbf{B})$, where S is the set of states, $\mathbf{A}$ is the transition matrix (representing the probabilities of transition between states), $\boldsymbol{\pi}$ is a vector of initial state probabilities, and $\mathbf{B}$ is the emission model, which describes the probability (density or mass) function of symbol emission from each state. All HMMs used in this paper are continuous valued ($O_i \in \mathbb{R}$), with the emission probability of each state assumed Gaussian. Training is performed using the standard Baum–Welch algorithm [2,3], initialized using a Gaussian mixture model (as in Ref. [24]).

As mentioned above, the typical HMM-based classification approach adopts the ML criterion, where an unknown sequence $\mathbf{O}$ is assigned to the class showing the highest likelihood, i.e.

$$Class(\mathbf{O}) = \arg\max_i P(\mathbf{O}|\boldsymbol{\lambda}_i), \tag{1}$$

where $\boldsymbol{\lambda}_i$ is the HMM corresponding to the ith class. This requires training C HMMs for a C-class problem. In the sequel, we will call this the ML_{OPC} approach (with *OPC* standing for "one per class").

A somewhat different rule could also be considered. Instead of training one HMM for each class, we could train one model for each training sequence, and assign an unknown sequence $\mathbf{O}$ to the class of the model showing the highest likelihood. More formally, let $\boldsymbol{\lambda}_i^{(k)}$ denote the HMM model trained on sequence $\mathbf{O}_i^{(k)}$, which belongs to class k. The classification rule under this approach is then

$$Class(\mathbf{O}) = \arg\max_k \left(\max_i P(\mathbf{O}|\boldsymbol{\lambda}_i^{(k)}) \right). \tag{2}$$

We call this the ML_{OPS} approach (with *OPS* standing for "one per sequence"). Notice that this may be seen as 1-NN classifier, with the proximity measure defined by the likelihood function.

4. The similarity-based strategy

4.1. Introduction

The basic issue of a similarity-based strategy is how to define similarities in an HMM framework. Recall that, given an HMM $\boldsymbol{\lambda}$ and a sequence $\mathbf{O}$, there is a standard method (*forward–backward procedure* [2]) to compute $P(\mathbf{O}|\boldsymbol{\lambda})$, i.e. the probability (density) that sequence $\mathbf{O}$ was generated by model $\boldsymbol{\lambda}$. This quantity is called the likelihood, and measures how well the sequence $\mathbf{O}$ "fits" the model $\boldsymbol{\lambda}$. A natural choice is then to define the similarity $D_{ij} = \mathscr{D}(\mathbf{O}_i, \mathbf{O}_j)$ between two sequences $\mathbf{O}_i$ and $\mathbf{O}_j$ as

$$D_{ij} = \mathscr{D}(\mathbf{O}_i, \mathbf{O}_j) = \frac{\log P(\mathbf{O}_i|\boldsymbol{\lambda}_j)}{T_i}, \tag{3}$$

where $\boldsymbol{\lambda}_j$ is the HMM trained on sequence $\mathbf{O}_j$, and T_i is the length of the sequence $\mathbf{O}_i$. The $1/T_i$ is a normalization factor introduced to take into account sequences of different length. Notice that this similarity is not symmetric.

The idea at the basis of the proposed approach is conceptually simple: to build a new representation space, using the similarity values between sequences obtained via the HMMs according to Eq. (3), and construct a classifier in that space. One of the justifications for this approach lies in the fact that similarity is high for similar objects/sequences, i.e. belonging to the same class, and low for objects of different classes, making discrimination possible [13]. Therefore, we can interpret the similarity measure $\mathscr{D}(\mathbf{O}, \mathbf{O}_i)$ between a sequence $\mathbf{O}$ and another "reference" sequence $\mathbf{O}_i$ as a "feature" of the sequence $\mathbf{O}$. This fact suggests the construction of a feature vector for $\mathbf{O}$ by taking the similarities between $\mathbf{O}$ and a set of reference sequences $\mathscr{R} = \{\mathbf{O}_k\}$, so that $\mathbf{O}$ is characterized by a *pattern* (i.e. a set of features) $\{\mathscr{D}(\mathbf{O}, \mathbf{O}_k), \mathbf{O}_k \in \mathscr{R}\}$.

This approach is well suited for HMMs. Given a sequence $\mathbf{O}$, the standard rule defined by Eq. (2) uses HMMs to compute the similarities between $\mathbf{O}$ and all the sequences in the training set. It then looks for the most similar training sequence, and classifies $\mathbf{O}$ as belonging to the class of this sequence (exactly as in a 1-NN classifier). Therefore, this process does not use all the information contained in the complete set of similarities, as done in the similarity-based approach. Notice that the fact that two sequences, say $\mathbf{O}_i$ and $\mathbf{O}_j$, present similar degrees of similarity to several other sequences (e.g., they are both very similar to some sequences, and also both very dissimilar to some other sequences) enforces the hypothesis that $\mathbf{O}_i$ and $\mathbf{O}_j$ belong to the same class.

4.2. Formal definition

Formally, the proposed strategy is defined as follows. Consider a classification problem with C classes; for each class $k \in \{1, 2, ..., C\}$, we have a set of N_k training sequences $\mathscr{T}_k = \{\mathbf{O}_1^{(k)}, ..., \mathbf{O}_{N_k}^{(k)}\}$; thus, $N = \sum_k N_k$ is the total size of the training set $\mathscr{T} = \bigcup_{k=1}^{C} \mathscr{T}_k$.

Let $\mathscr{R} = \{\mathbf{P}_1, ..., \mathbf{P}_R\}$ be a set of R "reference" or "representative" objects; these objects may belong to the set of training sequences ($\mathscr{R} \subseteq \mathscr{T}$) or may be otherwise defined. Now, let $\mathscr{D}_{\mathscr{R}}(\mathbf{O})$ be a function that returns the vector of similarities between an arbitrary sequence $\mathbf{O}$ and all the sequences in $\mathscr{R}$, which is

$$\mathscr{D}_{\mathscr{R}}(\mathbf{O}) = \begin{bmatrix} \mathscr{D}(\mathbf{O}, \mathbf{P}_1) \\ \vdots \\ \mathscr{D}(\mathbf{O}, \mathbf{P}_R) \end{bmatrix} \in \mathbb{R}^R. \tag{4}$$

We will designate the space $\mathbb{R}^R$ in which the dissimilarity vector exists as the "similarity space" and denote it as $\mathscr{S}_{\mathscr{R}}$, where the subscript $\mathscr{R}$ is used to emphasize the dependence of the similarity space on the set $\mathscr{R}$. Once this similarity space is defined, any standard classifier can, in principle, be used.

Regarding the choice of $\mathscr{R}$, different approaches can be adopted; the basic one, described in the next subsection, is to choose $\mathscr{R} = \mathscr{T}$, the whole training set. With this choice, the dimensionality of $\mathscr{S}_{\mathscr{R}} = \mathscr{S}_{\mathscr{T}}$ is equal to N, the cardinality of the training set $\mathscr{T}$. This is obviously a problem, because it makes the proposed method inapplicable in most cases; nevertheless, it is interesting to investigate the discrimination ability of this space.

Subsequently, the problem of reducing the dimensionality of the space is addressed by three different approaches: in the first, linear projection techniques are applied to the whole similarity space $\mathscr{S}_{\mathscr{T}}$; in the second one, we will modify the strategy used to compute of the distance $\mathscr{D}(\cdot, \cdot)$; finally, we use a greedy strategy, based on a *matching pursuit* algorithm, in order to choose a "good" set of representatives.

4.3. Basic approach: $\mathscr{R} = \mathscr{T}$

When we take $\mathscr{R} = \mathscr{T}$, the dimensionality of $\mathscr{S}_{\mathscr{R}}$ is equal to N, the cardinality of $\mathscr{T}$. Notice that in this case we are required to design a classifier on an N-dimensional space using only N training sequences; this is an extreme case of the *curse of dimensionality* [18], suggesting that some dimensionality reduction technique should be adopted. Linear transformations, such as PCA (see Ref. [19]) or FDA (see Ref. [17]), were conceived as means of reducing the dimensionality of a space while preserving almost all the "relevant information" contained in a data set. The concept of "relevant information" is different in PCA and FDA. In PCA, the information to be preserved is the variance of the data, that is, PCA seeks a data projection of lower

dimensionality that preserves most of its variance; it is thus an unsupervised learning technique since it does not use the class labels of the training samples. In contrast, FDA is a supervised technique that looks for a low-dimensional projection that best preserves the class separability of the data. In FDA, several criteria can be adopted to quantify the concept of "class separability" [17]; in this paper, we adopt the classical one proposed by Fisher [25]. The reduction of the space dimensionality absorbs in some way the impact of the curse of dimensionality; moreover, it could sometimes eliminate some redundancy present in the data (as shown in the experiments), leading to a better classification performance.

4.4. Choice of the set of representatives $\mathscr{R}$

If we want to avoid the curse of dimensionality without having to resort to PCA or FDA, smarter ways of choosing $\mathscr{R}$ have to be devised. Clearly, the choice of $\mathscr{R}$ is critical as only if this set is adequately chosen, the discrimination power of the space $\mathscr{S}_{\mathscr{R}}$ will be large. In this paper, we propose two methods, namely, the OPC and the *matching pursuit* (MP) procedures.

4.4.1. The "OPC" approach

In this approach, which is similar to the ML_{OPC} scheme described in Section 3, instead of training one HMM for each sequence, a model is trained for each class using all sequences of that class. Using these HMMs, the feature vector of a sequence $\mathbf{O}$ is a C-dimensional (for a C-class problem) vector given by

$$\mathscr{D}_{OPC}(\mathbf{O}) = \frac{1}{T}\begin{bmatrix} \log\ P(\mathbf{O}|\boldsymbol{\lambda}_1) \\ \vdots \\ \log\ P(\mathbf{O}|\boldsymbol{\lambda}_C) \end{bmatrix}, \tag{5}$$

where $\boldsymbol{\lambda}_j$ is the HMM estimated from the set of all training sequences from class j, and T is the length of sequence $\mathbf{O}$. In this case, $\mathscr{D}_{OPC}(\mathbf{O})$ can be seen as containing the similarities between $\mathbf{O}$ and each of the C classes. We can imagine the set $\mathscr{R}$ as containing C sequences $\{\mathbf{P}_1, ..., \mathbf{P}_C\}$, such that $\mathbf{P}_j$ is an (imaginary) sequence such that if we applied the learning algorithm to $\mathbf{P}_j$ we would still obtain $\boldsymbol{\lambda}_j$. In the following, we will denote the similarity space obtained with this approach as $\mathscr{S}_{OPC}$.

4.4.2. The MP approach

The MP approach is based on the following ideas: instead of using all sequences of the training set, one can choose those that are more "useful" in classification, i.e. more discriminant in some sense. This choice is made incrementally, starting with an empty set, and adding at each step the object that yields the largest "performance improvement". The process is stopped by some convergence criterion.

The MP algorithm was introduced in the signal processing community as an algorithm to decompose a signal into a linear combination of basis functions from a redundant dictionary [20]. It is a general, greedy, approximation scheme that works by sequentially appending functions to an initially empty set. At each step, the basis function appended is the one that produces the largest decrease in the approximation error. Recently, Vincent and Bengio [26] used MP to obtain kernel-based solutions to machine-learning problems.

Formally, the MP algorithm is defined as:

- Set $\mathscr{R} = \emptyset$ (the empty set);
- Until some stopping criterion is met, repeat:
  - For each sequence $\mathbf{O}_i^{(k)} \notin \mathscr{R}$, compute the *leave one out* (LOO) classification error rate of the 1-NN classifier using the feature vector $\mathscr{D}_{\{\mathscr{R}\cup\mathbf{O}_i^{(k)}\}}(\cdot)$. Let us denote this error as $E_{\mathscr{R}}(\mathbf{O}_i^{(k)})$.
  - The new representative set is $\mathscr{R} = \mathscr{R}\cup\{\mathbf{O}_{i^*}^{(k^*)}\}$, where

$$(i^*, k^*) = \arg \min_{(i,k):\mathbf{O}_i^{(k)}\notin\mathscr{R}} E_{\mathscr{R}}(\mathbf{O}_i^{(k)}).$$

In the following, we denote the similarity space obtained with this approach as $\mathscr{S}_{MP}$. Note that, unlike the OPC approach, this scheme is very general, and can be used in all other instances of similarity-based classification.

5. Results and discussion

In this section, experimental results are reported, in order to validate the proposed approach. Firstly, we investigate the discriminative power of the space $\mathscr{S}_{\mathscr{R}}$ with $\mathscr{R} = \mathscr{T}$, i.e. using as reference set the whole training set $\mathscr{T}$. The standard ML classification scheme and the proposed approach are compared, with both synthetic and real data. The use of PCA and FDA is investigated in this context, also with the aim of visualizing the data. Secondly, experimental results concerning the two different choices of $\mathscr{R}$ (OPC and MP) are reported. All the experiments are repeated 10 times and the results are averaged, so as to decrease the dependence of the results from the training of the HMMs.

5.1. Basic approach: $\mathscr{R} = \mathscr{T}$

5.1.1. Synthetic data

We consider a 3-class synthetic problem, defined by the parameters given in Fig. 1. The training set is composed of 30 sequences (of length 400) from each of the three classes; the dimensionality of the similarity space $\mathscr{S}_{\mathscr{R}}$ is thus $N = 90$. Notice that this classification task is not easy, as the three HMMs are very similar to each other, only differing slightly in the variances of the emission densities.

$$\mathbf{A} = \begin{array}{|c|c|c|}\hline 1/3 & 1/3 & 1/3 \\ \hline 1/3 & 1/3 & 1/3 \\ \hline 1/3 & 1/3 & 1/3 \\ \hline \end{array} \quad \boldsymbol{\pi} = \begin{array}{|c|}\hline 1/3 \\ \hline 1/3 \\ \hline 1/3 \\ \hline \end{array} \quad \mathbf{B} = \begin{array}{|c|c|}\hline \mu_1 = 1 & \sigma_1^2 = 0.6 \\ \hline \mu_2 = 3 & \sigma_2^2 = 0.6 \\ \hline \mu_3 = 5 & \sigma_3^2 = 0.6 \\ \hline \end{array}$$

(a)

$$\mathbf{A} = \begin{array}{|c|c|c|}\hline 1/3 & 1/3 & 1/3 \\ \hline 1/3 & 1/3 & 1/3 \\ \hline 1/3 & 1/3 & 1/3 \\ \hline \end{array} \quad \boldsymbol{\pi} = \begin{array}{|c|}\hline 1/3 \\ \hline 1/3 \\ \hline 1/3 \\ \hline \end{array} \quad \mathbf{B} = \begin{array}{|c|c|}\hline \mu_1 = 1 & \sigma_1^2 = 0.5 \\ \hline \mu_2 = 3 & \sigma_2^2 = 0.5 \\ \hline \mu_3 = 5 & \sigma_3^2 = 0.5 \\ \hline \end{array}$$

(b)

$$\mathbf{A} = \begin{array}{|c|c|c|}\hline 1/3 & 1/3 & 1/3 \\ \hline 1/3 & 1/3 & 1/3 \\ \hline 1/3 & 1/3 & 1/3 \\ \hline \end{array} \quad \boldsymbol{\pi} = \begin{array}{|c|}\hline 1/3 \\ \hline 1/3 \\ \hline 1/3 \\ \hline \end{array} \quad \mathbf{B} = \begin{array}{|c|c|}\hline \mu_1 = 1 & \sigma_1^2 = 0.4 \\ \hline \mu_2 = 3 & \sigma_2^2 = 0.4 \\ \hline \mu_3 = 5 & \sigma_3^2 = 0.4 \\ \hline \end{array}$$

(c)

Fig. 1. Generative HMMs for synthetic data testing: **A** is the transition matrix, $\boldsymbol{\pi}$ is the initial state probability and **B** contains the parameters of the emission density (Gaussians with the indicated means and variances).

We compare the standard ML classification criterion with a simple classifier in the similarity space $\mathcal{S}_{\mathcal{R}}$, the *k*-nearest-neighbor (*k*-NN), for $k = 1$ (1-NN) and $k = 3$ (3-NN), using Euclidean distance. This classical technique assigns a given object **O** to the class having the largest number of representatives in the set of the *k* objects in the training set that are nearest to **O**. This classifier is widely used, as it is simple, fast and reasonably accurate. The major drawback of NN classifiers is their sensitivity to noisy patterns on the training set, and the need to store all the training samples.

Accuracies were computed using the LOO procedure. This means that the dissimilarity space $\mathcal{S}_{\mathcal{R}}$ is actually built by using the representatives set $\mathcal{R}$ consisting of 89 sequences, while one sequence is left out and used for testing. The procedure is repeated until all sequences have been tested (i.e. 90 times), and results are averaged. Results of different classifiers are shown in Table 1 (a). We can observe that there is an improvement when using the simple classifier in the similarity space. Recall that, as mentioned above, the three classes are very similar and the classification task is very difficult.

In order to get a better insight into the structure of our similarity space, we have applied PCA and FDA to the space $\mathcal{S}_{\mathcal{T}}$. Plots of the 2D projections of the training set thus obtained are shown in Fig. 2. It is clear that FDA is really effective in separating the classes, and even PCA leads to a satisfactory result, even if it ignores the class labels. In both cases, the three classes in the training set would be easily separable, although generalization would clearly be better with the FDA projection.

Table 1
Classification accuracies using the basic approach on: (a) synthetic data and (b) synthetic data projected using PCA and FDA

Classifier	Accuracy (%)			
(a)				
ML_{OPS}	95.7			
1-NN on $\mathcal{S}_{\mathcal{T}}$	98.9			
3-NN on $\mathcal{S}_{\mathcal{T}}$	98.9			
	Dimensionality (%)			
	2	3	4	5
(b)				
1-NN on PCA space	98.9	98.9	98.9	98.9
MC on PCA space	98.9	97.8	97.8	96.7
1-NN on FDA space	100	—	—	—
MC on FDA space	100	—	—	—

Classification accuracies were also obtained in these reduced spaces, in order to investigate discrimination ability of the similarity space. In this case, we use 1-NN and the Mahalanobis classifier (MC), which classifies an unknown observation as belonging to the class whose mean is nearest, using a Mahalanobis distance [18]. Accuracies (again computed with the LOO procedure) are presented in Table 1(b). For FDA, the maximum dimensionality allowed is $C - 1$, where C is the number of classes [17]. In this case, therefore, the maximum dimensionality is two. Comparing Table 1(b) with Table 1(a) we can conclude that the performances on the FDA reduced space is increased, reaching a perfect classification rate (which is not surprising in view of Fig. 2(b)).

5.1.2. Real data

The proposed approach has been tested on two real applications: a 2D shape recognition task, detailed in Ref. [24], and a face recognition problem, using HMMs as proposed in Ref. [27].

In the 2D shape experiment, each object is represented by the sequence of the curvature coefficients, computed as follows: first, the contours are extracted by using the *Canny* edge detector; the boundary is then approximated by segments of approximately fixed length d_L. The resulting sequences show different lengths, ranging from 267 for the smallest object to 559 for the largest. Finally, the curvature value at point x is computed as the angle between the two consecutive segments intersecting at x. For a non-occluded object, the initial point is the rightmost point lying on the horizontal line passing through the object centroid, following the boundary in a counterclockwise manner. If the object is occluded, the endpoint allowing the contour to be

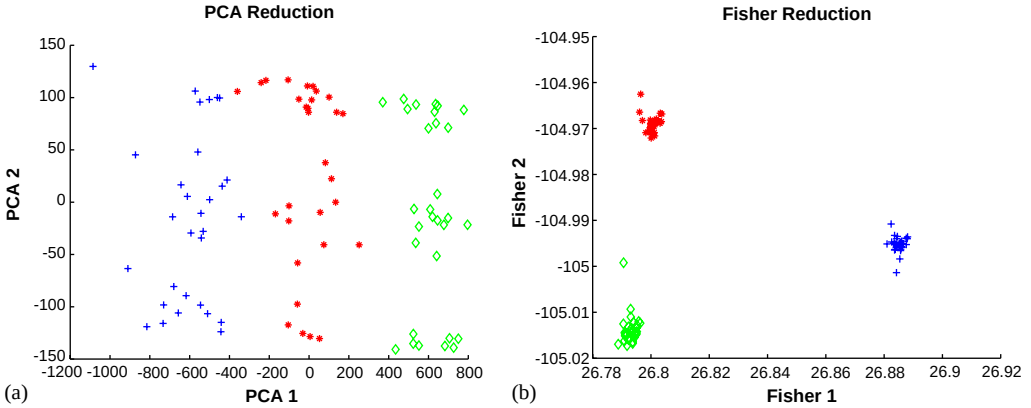

Fig. 2. 2D projections of the synthetic training set using: (a) PCA and (b) FDA.

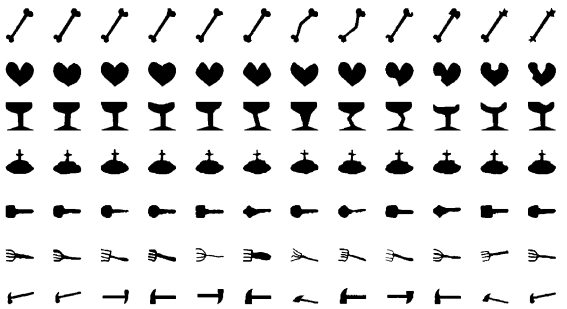

Fig. 3. Objects set used for testing.

Table 2
Classification accuracies using the basic approach on real data: (a) 2D shape recognition and (b) face recognition

Classifier	Accuracy (%)
(a)	
ML_{OPS}	80.9
1-NN on $\mathscr{S}_{\mathscr{F}}$	98.8
3-NN on $\mathscr{S}_{\mathscr{F}}$	93.2
(b)	
ML_{OPS}	50.6
1-NN on $\mathscr{S}_{\mathscr{F}}$	72.1
3-NN on $\mathscr{S}_{\mathscr{F}}$	60.5

followed in an counterclockwise way is considered as the initial point. A thorough analysis of the HMMs' capabilities in classifying 2D shapes is presented in Ref. [24], where the standard ML approach was tested in cases of translation, rotation, noise, occlusions, shearing transformations and combination of the above perturbations, showing really promising results.

In this paper we compare our similarity-based approach with a simplified version of the system described in Ref. [24]: unlike in that paper, we do not use here any model selection technique, the number of states was fixed to three for all experiments. Testing was performed on part of the object set used in Ref. [28], composed by seven classes, each containing 12 different shapes. As before, accuracies are computed using the LOO scheme. The database used is shown in Fig. 3.

For the face recognition task, HMMs were used as proposed in Ref. [27], considering DCT coefficients as features. Given a sequence of sub-images of the face image, the DCT coefficients of each sub-image are computed, and vectorized using a *zig-zag* scan. The number of coefficients chosen determines the dimensionality of the observation, and 10 coefficients are used in our experiment. The sequence of sub-images is obtained by sliding over the face image a square fixed size window, in a raster scan fashion, with a predefined overlap. The window size and the overlap ratio were fixed to 8% and 50%, respectively. Testing was performed using the Bern face database,[3] which consists of 30 subjects with 10 face images each. For each subject, five faces were used for training and the others for testing. We have chosen to use this database, instead of the ORL used in Ref. [27], because with that database HMMs are able to reach an almost perfect classification, so without any possibility of improvement.

The classical ML classification criterion was compared with the similarity-based approach, using a k-NN rule (for $k = 1$ and 3) in the similarity space $\mathscr{S}_{\mathscr{F}}$. Accuracies are presented in Table 2 (a) and (b), for 2D shape recognition and for face classification tasks, respectively.

In the 2D shape case, the improvement in classification rates is very large, of about 18% for the 1-NN classifier and of about 13% for the 3-NN. This shows that this similarity-based feature space is very well suited for this real problem.

[3] Downloadable from ftp://iamftp.unibe.ch/pub/Images/FaceImages

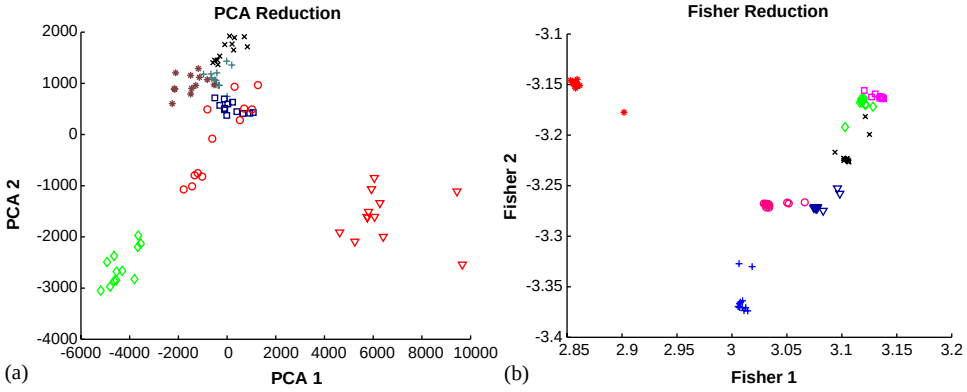

Fig. 4. Data set for 2D shape recognition experiment, reduced and plotted using: (a) PCA and (b) FDA.

Table 3
LOO accuracies for the 2D shape recognition task after PCA and FDA projections

	Dimensionality (%)			
	2	3	4	5
1-NN on PCA space	80.1	97.7	98.2	98.3
MC on PCA space	81.2	92.9	91.7	93.3
1-NN on FDA space	92.5	95.0	96.3	97.1
MC on FDA space	86.5	92.6	90.1	91.3

Table 4
Classification accuracies on data set formed by occluded shapes, at different occlusion levels

	Occlusion level		
	10%	30%	50%
ML_{OPS}	76.9%	71.5%	60.9%
NN on $\mathcal{S}_{\mathcal{F}}$	91.5%	76.4%	64.2%
3-NN on $\mathcal{S}_{\mathcal{F}}$	91.9%	73.0%	64.1%

Actually, using the standard ML approach, most of the errors occur in the "key" class: looking at Fig. 3, it is worth noting that the instances of this class represent the same object only semantically, but the related shapes are indeed very different. This aspect, which is negative in the ML scheme, is the key feature of our approach: since there are many differences among items in the same class, the use of all similarities between items may add a lot of discriminative power to the method. The additional discriminative power increases more when the differences among items of same class are larger.

Also in the face recognition case there is a noticeable improvement in the accuracies of classification (about 12% and 10% for the 1-NN and 3-NN, respectively), confirming the wide applicability of this method to real cases.

FDA and PCA were also studied in the case of the 2D shape recognition experiment. Plots of projected training set are shown in Fig. 4. As in the previous subsection, classification accuracies were calculated for different dimensionalities, using the LOO procedure, and the results are reported in Table 3 . In this case, the reduction of dimensionality to 2 decreases the classification performance, which, in any case, is about equal or still better than the results obtained using the standard ML criterion. The similarity feature space is complex in this case due to the presence of very dissimilar elements in the same class. For low dimensionality, part of this information is lost, but, by slightly increasing the dimensionality, this information is correctly recovered, and the performance returns to a very good level.

To investigate the robustness of the approach, we have also tested the method behavior in the presence of shape occlusions. Occlusion is one of the most severe limitations to the application of typical object recognition techniques. Recently [24,29], it has been shown that HMMs are very effective in dealing with object occlusions. Here we show that the approach proposed in Ref. [24] can be further improved by using the similarity space representation.

Object occlusion is simulated by considering a fragment of the object boundary, starting at a randomly chosen location. Each object was occluded 5 times, resulting in 420 sequences. Occlusion percentages considered were 10%, 30% and 50%. Notice that in the latter case, one half of the whole boundary is missing. An LOO scheme was again adopted, which means that these results are obtained in a really complex task, as the left out sequence (the occluded one) was not used for building the similarity space. This choice makes all experiments uniform throughout the paper, even if it can be seen as somewhat strange, since typically to recognize an occluded object, also the original shape is available (this obviously results in a great improvement in the performances, see Refs. [24,29]).

Results for the different occlusion levels considered, using 1-NN and 3-NN classifiers, are shown in Table 4 . We

Table 5
Accuracies of the OPC and MP approaches in: (a) 2D shape experiment, with entire shapes; (b) 2D shape experiment, with occluded shapes, for different occlusion levels (OL) and (c) face recognition experiment

Classifier	Accuracy (%)	Dim. of $\mathcal{S}$
(a)		
ML_{OPC}	89.3	—
ML_{OPS}	80.9	—
1-NN on $\mathcal{S}_{\mathcal{F}}$	98.8	84
1-NN on $\mathcal{S}_{OPC}$	97.4	7
1-NN on $\mathcal{S}_{MP}$	92.9	4.1

Classifier	OL = 10%	OL = 30%	OL = 50%	Dim. of $\mathcal{S}$
(b)				
ML_{OPC}	83.0%	78.0%	69.2%	—
ML_{OPS}	76.9%	71.5%	60.9%	—
1-NN on $\mathcal{S}_{\mathcal{F}}$	91.5%	76.4%	64.2%	84
1-NN on $\mathcal{S}_{OPC}$	86.1%	71.1%	57.8%	7
1-NN on $\mathcal{S}_{MP}$	85.9%	72.1%	56.1%	4.26

Classifier	Accuracy (%)	Dim. of $\mathcal{S}$
(c)		
ML_{OPC}	51.67	—
ML_{OPS}	50.60	—
1-NN on $\mathcal{S}_{\mathcal{F}}$	72.1	150
1-NN on $\mathcal{S}_{OPC}$	69.4	30
1-NN on $\mathcal{S}_{MP}$	68.9	10.1

observe a clear improvement in the classification accuracies of the classifiers in the similarity space.

5.2. Choice of representatives set $\mathcal{R}$

In this section, the two approaches for the choice of $\mathcal{R}$ described in Section 4.4 are tested. These approaches were applied to the 2D shape recognition (using both the entire and occluded shapes) and to the face classification experiments. Classification accuracies were calculated as in the previous section. We used 1-NN classifiers in the similarity spaces $\mathcal{S}_{OPC}$ and $\mathcal{S}_{MP}$.

A comparison between the proposed approaches and ML classification is reported in Tables 5 (a)–(c), for the entire shapes, the occluded shapes and the face experiments, respectively. For the sake of clarity, results for 1-NN on $\mathcal{S}_{\mathcal{F}}$ (entire similarity space) are also shown, in order to quantify the loss in classification accuracy determined by the reduction. Moreover, the dimension of the resulting similarity space $\mathcal{S}$ is included in the tables, in order to emphasize the amount of the reduction obtained.

In summary, we can conclude that both approaches are able to preserve most of the performance of the basic approach (classification on the whole similarity space $\mathcal{S}_{\mathcal{F}}$), while achieving a drastic dimensionality reduction. Regarding the 2D shape recognition experiment, by comparing the performance of the ML_{OPC} method, with the standard ML_{OPS} criterion, we can notice that the use of all sequences to learn each HMM enhances the accuracy of the standard ML classification. HMM is really suitable to be trained using many sequences, as it is able to deal with their possible different lengths. Nevertheless, this could reduce the expressivity of the resulting similarity space, especially in some real cases, where items of the same class present remarkable differences between each other. From Table 5(b) we can also notice that when increasing too much the occlusion level, the performances on reduced similarity spaces (MP and OPC approaches) are lower than standard ML classification level. This is probably due to the fact that when the percentage of occlusion increases, the HMMs are less accurately estimated. The obtained similarity space is thus noisy, and the 1-NN rule (that is the simplest classifier) is not able to perform well in a such noisy space. To verify this explanation, we recompute the LOO classification accuracies on the experiment with the occluded shapes, with occlusion level 50%. We used a carefully trained multilayer feed forward neural network on the MP reduced similarity space: 1-NN accuracies were about 56% in that reduced space. Accuracies obtained with neural network is around 88%, confirming the large potentialities of this approach: the mapping onto the similarity space allows us to reduce complex sequence classification into easier standard point classification, for which one could use arbitrarily sophisticated techniques.

In conclusion, the two approaches for the choice of representatives set $\mathcal{R}$ are both effective. OPC seems to be more interesting, as it results directly from the standard HMM training, without any need to postprocess the space. Nevertheless, the resulting dimensionality is equal to the number of classes, reducing the usefulness of the approach in problems with many classes (e.g., face recognition). Moreover, the training of one HMM for each class can drastically reduce the discrimination ability of the similarity space when items of the same class are very different. On the other hand, the MP approach seems better at identifying the representatives that are *really* useful for the similarity-based classification purpose. The higher computational burden introduced with this approach is its major drawback.

5.3. Computational aspects

The similarity-based technique introduced in this paper is more computational demanding than the ML scheme. More specifically, with our approach, in order to build the similarity space of the training sequences, in the learning phase all the training sequences should be evaluated, whereas this is not needed by the standard ML approach. The training phase, nevertheless, is performed only once, typically offline, so the overall impact is minor. Regarding the testing case, both schemes compute the likelihoods of the testing sequence for all the trained HMMs. Subsequently, the ML

Table 6
Computational requirements (in seconds on a 800 MHz processor with 256 MB of RAM) of the ML scheme and our approach (using 1-NN) in the 2D shape recognition experiment

Strategy	Training time (s)	Testing time (s)
ML_{OPS}	62.01	2.23
ML_{OPC}	62.55	0.20
1-NN on $\mathscr{S}_{\mathscr{T}}$	308.16	2.67
1-NN on $\mathscr{S}_{MP}$	318.78	0.11
1-NN on $\mathscr{S}_{OPC}$	83.57	0.23

scheme looks for the maximum, while our approach uses the whole likelihood vector as feature pattern, using a standard fixed feature vector strategy for the classification. The overhead introduced by our approach strictly depends on the classification strategy chosen: the more complex this strategy, the larger the computational burden which is added. If needed (e.g., when using a neural network), an additional training should be performed on the similarity space of the training sequences. Nevertheless, as shown in the experimental part, using a simple classifier (as the *K*-NN) which does not need any training, we could obtain a large improvement in the results' accuracies.

The second consideration regards the reduction of the dimensionality of the similarity space, which, in the basic approach, is equal to the cardinality of the training set, inapplicable for practical situations. In this paper two techniques have been introduced in order to address this problem, namely the OPC and the MP approaches. The former reduces the needed training time, since only few HMMs have to be trained (equal to the number of classes), and there is no need of postprocessing the resulting space. Moreover the testing phase is sped up, since the similarity space has a smaller dimensionality, hence resulting in a faster classification. The latter is more accurate, but a computational overhead in the training phase is introduced, as choosing the appropriate representatives is quite onerous.

Practically, all the experiments have been conducted on a machine with a 800 MHz processor and 256 MB RAM. The code was entirely developed under the MATLAB environment. We reported in Table 6 the time needed by the standard ML schemes and our approach (using 1-NN) in the 2D shape recognition experiment, for the training phase (all 84 sequences) and for the testing phases (one testing sequence). We can note that most part of the overhead is in the training phase, while the testing phase remains quite fast, becoming faster using the MP and OPC approaches. The best compromise between computational overhead and accuracy seems to be the OPC approach.

6. Conclusions

In this paper we have proposed a novel sequence classification scheme by combining hidden Markov models (HMM) with the similarity-based paradigm. This approach creates a representation space for sequences in which standard feature-based classification techniques can be used. We showed that a simple classifier in a such space outperforms standard HMM-based classification schemes. Three approaches to deal with the high dimensionality of the resulting space were also considered and investigated, showing that the similarity-based representation is still effective when its dimensionality is reduced in order to make it more manageable.

Future directions consist in applying and investigating more ad hoc similarity space classifiers, as those proposed in Refs. [13,14], and in studying novel techniques for reducing space dimensionality.

References

[1] L.R. Rabiner, A tutorial on hidden Markov models and selected applications in speech recognition, Proc. IEEE 77 (2) (1989) 257–286.
[2] L.E. Baum, T.E. Petrie, G. Soules, N. Weiss, A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains, Ann. Math. Stat. 41 (1) (1970) 164–171.
[3] L.E. Baum, An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes, Inequality 3 (1970) 1–8.
[4] R. Durbin, S. Eddy, A. Krogh, G.J. Mitchison, Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids, Cambridge University Press, Cambridge, 1998.
[5] R. Hughey, A. Krogh, Hidden Markov model for sequence analysis: extension and analysis of the basic method, Comput. Appl. Biosci. 12 (1996) 95–107.
[6] J. Hu, M.K. Brown, W. Turin, HMM based online handwriting recognition, IEEE Trans. Pattern Anal. Mach. Intell. 18 (10) (1996) 1039–1045.
[7] S. Eickeler, A. Kosmala, G. Rigoll, Hidden Markov model based continuous online gesture recognition, Proceedings of International Conference on Pattern Recognition, Vol. 2, 1998, pp. 1206–1208.
[8] T. Jebara, A. Pentland, Action reaction learning: automatic visual analysis and synthesis of interactive behavior, In: Proceedings of International Conference on Computer Vision Systems, 1999.
[9] A.K. Jain, D. Zongker, Representation and recognition of handwritten digits using deformable templates, IEEE Trans. Pattern Anal. Mach. Intell. 19 (12) (1997) 1386–1391.
[10] T. Graepel, R. Herbrich, P. Bollmann-Sdorra, K. Obermayer, Classification on pairwise proximity data, In: D. Cohn, M. Kearns, S. Solla (Eds.), Advances in Neural Information Processing, Vol. 11, MIT Press, Cambridge, MA, 1999, pp. 438–444.
[11] D.W. Jacobs, D. Weinshall, Classification with nonmetric distances: image retrieval and class representation, IEEE Trans. Pattern Anal. Mach. Intell. 22 (6) (2000) 583–600.
[12] E. Pekalska, R.P.W. Duin, Automatic pattern recognition by similarity representations, Electron. Lett. 37 (3) (2001) 159–160.

[13] E. Pekalska, P. Paclik, R.P.W. Duin, A generalized kernel approach to dissimilarity-based classification, J. Mach. Learning Res. 2 (2) (2002) 175–211.
[14] E. Pekalska, R.P.W. Duin, Dissimilarity representations allow for building good classifiers, Pattern Recogn. Lett. 23 (8) (2002) 943–956.
[15] P. Smyth, Clustering sequences with hidden Markov models, In: M. Mozer, M. Jordan, T. Petsche (Eds.), Advances in Neural Information Processing, Vol. 9, MIT Press, Cambridge, MA, 1997, pp. 648–654.
[16] A. Panuccio, M. Bicego, V. Murino, A hidden Markov model-based approach to sequential data clustering, In: T. Caelli, A. Amin, R.P.W. Duin, M. Kamel, D. de Ridder (Eds.), Structural, Syntactic and Statistical Pattern Recognition, Lecture Notes in Computer Science, Vol. 2396, Springer, Berlin, 2002, pp. 734–742.
[17] K. Fukunaga, Introduction to Statistical Pattern Recognition, 2nd Edition, Academic Press, New York, 1990.
[18] R.O. Duda, P.E. Hart, D.G. Stork, Pattern Classification, 2nd Edition, Wiley, New York, 2001.
[19] I.T. Jollife, Principal Component Analysis, Springer, New York, 1986.
[20] S. Mallat, Z. Zhang, Matching pursuit with time-frequency dictionaries, IEEE Trans. Signal Process. 41 (12) (1993) 3397–3415.
[21] V. Vapnik, Statistical Learning Theory, Wiley, New York, 1998.
[22] E. Pekalska, R.P.W. Duin, Spatial representation of dissimilarity data via lower-complexity linear and non linear mappings, In: T. Caelli, A. Amin, R.P.W. Duin, M. Kamel, D. de Ridder (Eds.), Structural, Syntactic and Statistical Pattern Recognition, Lecture Notes in Computer Science, Vol. 2396, Springer, Berlin, 2002, pp. 488–497.
[23] E. Pekalska, R.P.W. Duin, Prototype selection for finding efficient representations of dissimilarity data, In: Proceedings of the International Conference on Pattern Recognition, Vol. 3, 2002, pp. 37–40.
[24] M. Bicego, V. Murino, Investigating Hidden Markov models' capabilities in 2D shape classification, IEEE Trans. Pattern Anal. Mach. Intell. 26 (2) (2004) 281–286.
[25] R.A. Fisher, The use of multiple measurements in taxonomic problems, Ann. Eugenics 7 (1936) 179–188 Republished in Contributions to Mathematical Statistics, Wiley, New York, 1950.
[26] P. Vincent, Y. Bengio, Kernel matching pursuit, Mach. Learning 48 (1) (2002) 165–187.
[27] V.V. Kohir, U.B. Desai, Face recognition using DCT-HMM approach, Proceedings of the Workshop on Advances in Facial Image Analysis and Recognition Technology (AFIART), Freiburg, Germany, 1998.
[28] T.B. Sebastian, P.N. Klein, B.B. Kimia, Recognition of shapes by editing Shock graphs, Proceedings of International Conference on Computer Vision, 2001, pp. 755–762.
[29] M. Bicego, V. Murino, 2D shape recognition by hidden Markov models, Proceedings of the International Conference on Image Analysis and Processing, 2001, pp. 20–24.

About the Author—MANUELE BICEGO received his Laurea degree and Ph.D. degree in Computer Science from University of Verona in 1999 and 2003, respectively. Since 2000 he is working in VIPS (Vision, Image Processing and Sound) laboratory of the Computer Science Department of University of Verona. His research interests include statistical pattern recognition, artificial vision, electronic noses, neural networks, hidden Markov models and video analysis. Manuele Bicego is member of the IAPR-IC society and student member of the IEEE Systems, Man, and Cybernetics society.

About the Author—VITTORIO MURINO is professor and chairman of the Department of Computer Science of the University of Verona, Italy. He received the Laurea degree in Electronic Engineering in 1989 and the Ph.D. in Electronic Engineering and Computer Science in 1993, both at the University of Genoa. He was a Post-Doctoral Fellow at the University of Genoa, working in the Signal Processing and Understanding Group of the Department of Biophysical and Electronic Engineering, as supervisor of research activities on signal and image processing in underwater environments. From 1995 to 1998, he was assistant professor at the Department of Mathematics and Computer Science of the University of Udine, Italy, supervising research activities regarding multisensorial underwater vision for object recognition and virtual reality applications. From 1998 he is at the University of Verona, where he founded the Vision, Image Processing and Sound (VIPS) laboratory. He worked at several national and European projects, especially in the context of the MAST (MArine Science and Technology) programme concerning with the investigation of underwater scenes by visual and acoustical sensors. He is an evaluator for the European Commission of project proposals related to several programmes. His main research interests include: 3D computer vision and pattern recognition, acoustic and optical underwater vision, probabilistic techniques for image processing (specifically, Hidden Markov models, Markov random fields, Bayesian networks), data fusion, and neural networks with applications on surveillance, autonomous driving, visual inspection, and robotics. He is also interested in the integration of image analysis and synthesis methodologies for object recognition and 3D modelling. Dr. Murino is author of more than 100 papers in the above subjects, and associate editor of the Pattern Recognition and IEEE Transactions on Systems, Man, and Cybernetics journals, and the electronic journal ELCVIA (Electronic Letters on Computer Vision and Image Analysis). He is also referee for many international journals, member of IAPR, and senior member of IEEE.

About the Author—MÁRIO A.T. FIGUEIREDO received the E.E., M.Sc. and Ph.D. degrees in electrical and computer engineering, all from "Instituto Superior Tecnico" (IST), the engineering school of the Technical University of Lisbon, Portugal, in 1985, 1990, and 1994, respectively. He has been an Assistant Professor with the Department of Electrical and Computer Engineering of IST, since 1994. He is also a researcher with the Communication Theory and Pattern Recognition Group at the Institute of Telecommunications, Lisbon. In 1998, he was a visiting professor with the Department of Computer Science and Engineering, at Michigan State University. His research interests are in the fields of image analysis, computer vision, and statistical pattern recognition. He is currently an Associate Editor of the IEEE Transactions on Image Processing, the IEEE Transactions on Mobile Computing, and Pattern Recognition Letters. He is co-chair of the 2001 and 2003 editions of the International Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition. Mário Figueiredo is a Senior Member of the IEEE and received the 1995 Portuguese IBM Scientific Prize for his work on unsupervised image restoration.