Available online at www.sciencedirect.com
SCIENCE DIRECT®
Pattern Recognition Letters 26 (2005) 2588–2599
Pattern Recognition Letters
www.elsevier.com/locate/patrec

A Hidden Markov Model approach for appearance-based 3D object recognition

Manuele Bicego [1], Umberto Castellani, Vittorio Murino *

Dipartimento di Informatica, Università di Verona Ca' Vignal 2, Strada Le Grazie 15 37134 Verona, Italia

Received 22 June 2004; received in revised form 30 May 2005
Available online 3 August 2005

Communicated by S. Dickinson

Abstract

In this paper, a new appearance-based 3D object classification method is proposed based on the Hidden Markov Model (HMM) approach. Hidden Markov Models are a widely used methodology for sequential data modelling, of growing importance in the last years. In the proposed approach, each view is subdivided in regular, partially overlapped sub-images, and wavelet coefficients are computed for each window. These coefficients are then arranged in a sequential fashion to compose a sequence vector, which is used to train a HMM, paying particular attention to the model selection issue and to the training procedure initialization. A thorough experimental evaluation on a standard database has shown promising results, also in presence of image distortions and occlusions, the latter representing one of the most severe problems of the recognition methods. This analysis suggests that the proposed approach represents an interesting alternative to classic appearance-based methods to 3D object classification.

Keywords: 3D object recognition; Hidden Markov Models; Model selection; Shape occlusion; Appearance-based recognition

1. Introduction

Three-dimensional object recognition is an active research area in computer vision. Several and variegate approaches have been proposed in the past, differentiating by the types of input data and models (2D, 2.5D, 3D), intermediate representations, and types of procedures manipulating such representations. Structural, probabilistic, or

* Corresponding author. Tel.: +39 045 8027996; fax: +39 045 8027068.
E-mail addresses: bicego@uniss.it, bicego@sci.univr.it (M. Bicego), castellani@sci.univr.it (U. Castellani), murino@sci.univr.it (V. Murino).
[1] Current address: DEIR, University of Sassari, via Torre Tonda 34, 07100 Sassari (Italy).

0167-8655/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2005.06.005

algebraic methods, graph-, feature-, and physics-based algorithms, and hybrid strategies have been proposed in the literature (Prokop and Reeves, 1992; Arman and Aggarwal, 1993; Pope, 1994; Loncaric, 1998; Batlle et al., 2000; Campbell and Flynn, 2001; Duda et al., 2001), so that a comprehensive taxonomy is difficult to organize properly. Nevertheless, focusing on the type of model, a general and commonly accepted subdivision of recognition methods can be stated into two main classes: object-centered and viewer-centered. Roughly speaking, the former approaches are characterized by the use of a 3D model of the object, which can be available or generated in some way (Pope, 1994). Therefore, the actual classification is typically performed by searching for the best alignment between the model and the observed view (Pope and Lowe, 2000). In the latter class of methods, also called aspect- or appearance-based, no 3D model is available, but only a set of model aspects, so that the recognition can be performed by directly analyzing and comparing the observed and the model views (Murase and Nayar, 1995). These approaches are in fact characterized by robustness against rotation and pose changes due to the use of multiple views of the same object model. Therefore, given an unknown view, the recognition is typically achieved by determining the most similar view. Appearance-based methods are then well-suited for dealing with recognition problems in which geometric models of the observed objects are difficult or impossible to obtain. The fact that a small set of 2D views of a complex 3D object may be sufficient to recognize the object in critical conditions (as illumination changes and variations of the point of view) is supported by both psychophysical evidence (Bulthoff et al., 1994) and theoretical speculations (Ullman and Basri, 1991). A brief review of recent literature regarding appearance-based approaches to object recognition is presented in Section 2.

In this paper, a new method to appearance-based 3D object recognition is proposed, based on Hidden Markov Models (HMMs). Hidden Markov Models are a widespread approach to probabilistic sequence modelling: they can be viewed as stochastic generalizations of finite-state automata, where both transitions between states and generation of output symbols are governed by probability distributions (Rabiner, 1989; Baum et al., 1970; Baum, 1970). Originally, these models were almost exclusively applied in the speech recognition context, and it is only in the last decade that they have been widely used for several other applications, as handwritten character recognition, DNA and protein modelling, gesture recognition, and behavior analysis and synthesis. Even if HMMs have been largely applied for classifying planar objects (He and Kundu, 1991; Arica and Yarman-Vural, 2000; Fred et al., 1997; Cai and Liu, 2001; Bicego and Murino, 2004), its use in the context of 3D object recognition has been poorly investigated, and only few papers exploring this research direction are appeared in the literature. A first approach was proposed in (Ham and Park, 1999), where range images are modelled using HMMs and Neural Networks, using 3D features such as surface type, moments and others. More recently, in (Trazegnies et al., 2003), a HMM was used to model the sequence of 2D views gathered from a moving camera, where each view is described using contour-based features. Even if interesting, the method presents some drawbacks: first, only one HMM was trained for each class of objects, requiring a re-training in case of availability of further views. Moreover, only the contour of the object was employed to compute features, discarding important information such as texture and colors. Finally, a thorough experimentation with a standard database is missing, and occlusions are not considered.

In this paper a different approach is proposed, which explicitly considers all the information contained in the object view, modelling it using a HMM. The image is visited in a raster-scan fashion with a squared window of fixed size, obtaining a sequence of overlapping sub-images. For each sub-image, wavelet coefficients are computed, discarding the less significant ones. The sequence of wavelet features associated to each sub-image is then modelled using a HMM. In the modelling, particular care is devoted to the training procedure initialization, which represents a crucial factor because of the locality of the optimization procedure, and to the model selection issue, which represents the problem of choosing the topology and the number of states of the HMM. These issues,

typically disregarded in the application-oriented HMM literature, are fundamental factors for obtaining an effective modelling, and have been carefully addressed in this paper.

A strategy similar to that proposed in this paper has been recently applied by the authors in the context of face recognition (Bicego et al., 2003a), showing promising results. Even if the modelling strategy is similar, there are two important differences between the approach in (Bicego et al., 2003a) and the one proposed in this paper: the classification strategy and the application domain. The classification strategy adopted in (Bicego et al., 2003a) is the standard rule: one model is trained for each class, and is subsequently used as class-conditional densities in a standard Bayes classification paradigm. Assuming a priori equiprobable classes, an unknown sequence 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, instead of training one HMM for each class, we train one model for each training sequence (each view), and assign an unknown view to the class of the model showing the highest likelihood. Notice that this may be seen as a 1-nearest-neighbor (1-NN) classifier, with the proximity measure defined by the likelihood function. This scheme is particularly suited in this application domain and some justifications have been discussed in the paper. The second important difference is the application domain: in (Bicego et al., 2003a), the strategy is applied to the face classification problem (inherently 2D), whereas in this paper it is applied to the 3D object recognition context.

The proposed approach has been thoroughly tested on the COIL database (Nene et al., 1996), which represents a standard in the object recognition literature. The results are really promising, especially in the case of occlusions, which represents one of the most severe problems of the appearance-based methods to 3D object recognition. A careful analysis of the minimum number of views needed by the system in order to work properly has also been presented, showing that the proposed system reaches a satisfactory accuracy also using only eight views.

Summarizing, the main features of the proposed approach are the following: firstly, it has been shown in the experimental part that the system is really robust to objects perturbations, like radial distortion or occlusions, the latter representing one of the most severe perturbations in the object recognition context. Secondly, the system is easily extendible to new objects or new aspects of objects, guaranteeing the scalability of the method. Thirdly, the system performed in a satisfactory way even if the number of views per object used for training is drastically reduced. Nevertheless, the main distinctive merit of the paper is that this represents the first study which thoroughly and systematically investigated the HMM capabilities in 3D object recognition. HMMs have been largely applied in several computer vision and pattern recognition problems, whereas a systematic analysis of its behavior in this context is missing in literature. The obtained results confirm the real effectiveness of the Hidden Markov Model approach, which is able to properly work even if the application domain in not completely "sequential", and the sequence has to be forcedly determined from the data.

The rest of the paper is organized as follows. In Section 3, the fundamental concepts of the proposed approach are briefly introduced. The core strategy is detailed in Section 4, and Section 5 describes experimental results. Finally, in Section 6, conclusions are drawn and future perspectives are envisaged.

2. Appearance-based approaches

There are several examples of appearance-based approaches to 3D object recognition. In (Murase and Nayar, 1995) the authors based their method on the notion of parametric eigenspace. This approach has been further exploited in (Leonardis and Bischof, 2000) in term of robustness of recognition performance, especially in presence of illuminance variations (Bischof et al., 2001). Moreover, in order to improve the robustness with respect to the occlusions, visual recognition using *local* appearance has been introduced (de Verdiere and Crowley, 1998; Schneiderman and Kanade,

1998; Moghaddam et al., 2003). Local features (*local descriptors*) are extracted from small windows (Schneiderman and Kanade, 1998; Moghaddam et al., 2003) or from interest points (de Verdiere and Crowley, 1998) by defining a *local appearance space* on which the recognition is carried out.

A wide class of appearance-based object recognition methods is based on Support Vector Machine (SVM) (Pontil and Verri, 1998; Roobaert et al., 2001; Barla et al., 2002). In (Pontil and Verri, 1998), the interesting issue is that feature extraction is not required, and the recognition is performed directly on images considered as points in a high dimensional space. Given fixed but unknown probability distributions, a SVM finds the hyperplane (called optimal separating hyperplane) that maximizes the margin between the classes. In (Roobaert et al., 2001), the robustness of SVMs on background variations is analyzed. Furthermore, the authors have tested the method by using just a few images per object during the learning phase. In (Barla et al., 2002), the authors have focused on the selection of the kernel functions. In order to introduce correctly the prior information of the learning system, a new class of kernels is introduced basing on similarity measures inspired by the Hausdorff distance (Huttenlocher et al., 1993).

Another class of methods is based on the measure of similarity between shapes (Hagedoorn, 2000; Cyr and Kimia, 2001; Belongie et al., 2002). In (Cyr and Kimia, 2001), an aspect graph-based method is proposed. The measure of the similarity between two views is given by measuring the distance between the projected and segmented shapes of the 3D object. This endows the viewing sphere with a metric which is used to group similar views into aspects, and to represent each aspect by a prototype. The same shape similarity metric is then used to rate the similarity between unknown views of unknown objects and stored prototypes to identify the object and its pose. In (Belongie et al., 2002), the authors proposed to proceed by (1) solving for correspondences between points on the two shapes, (2) using the correspondences to estimate an aligning transform. Then, the sum of matching error between corresponding points, together with a term measuring the magnitude of aligning transform, give the shapes' similarity. An extensive survey of shape matching in computer vision can be found in (Hagedoorn, 2000).

Finally, another interesting approach could be found in (Caputo et al., 2002), where a probabilistic strategy to 3D object recognition is proposed. The authors introduce a method that allows the use of the spin-glass theory (SGT) in the context of a maximum a posteriori-Markov random field (MAP-MRF). Features are extracted from the images using a multidimensional receptive field histogram (MFH) representation, and, by applying the SGT, the open question of defining automatically an appropriate neighborhood system on irregular sites is solved. Furthermore, the algorithm is very efficient as the solution is found analytically, so that it does not require any searching techniques to look for the absolute minima.

3. Fundamentals

In this section the fundamental tools of the proposed approach are briefly summarized: in Section 3.1 the theory of the HMM is presented, while a short introduction to the wavelets is given in Section 3.2.

3.1. Hidden Markov Models

A discrete-time first-order HMM (Rabiner, 1989) is a probabilistic model that describes a stochastic sequence $\boldsymbol{O} = O_1, O_2, \ldots, O_T$ as being an indirect observation of an underlying (hidden) random sequence $\boldsymbol{Q} = Q_1, Q_2, \ldots, Q_T$, where this hidden process is Markovian, though the observed process may not be so. More formally, a HMM is defined by the following entities (Rabiner, 1989):

- $S = \{S_1, S_2, \ldots, S_N\}$ the finite set of the hidden states;
- the transition matrix $\boldsymbol{A} = \{a_{ij}, 1 \leqslant j \leqslant N\}$ representing the probability to go from state S_i to state S_j,

  $a_{ij} = P[Q_{t+1} = S_j | Q_t = S_i] \quad 1 \leqslant i, j \leqslant N$

  with $a_{ij} \geqslant 0$ and $\sum_{j=1}^{N} a_{ij} = 1$;

- the emission matrix $\boldsymbol{B} = \{b(O|S_j)\}$, indicating the probability of the emission of the symbol O when system state is S_j; in this paper continuous HMM were employed: $b(O|S_j)$ is represented by a Gaussian distribution, i.e.

$$b(O|S_j) = \mathcal{N}(O|\mu_j, \Sigma_j). \quad (1)$$

  where $\mathcal{N}(O|\mu, \Sigma)$ denotes a Gaussian density of mean μ and covariance Σ, evaluated at O;
- $\pi = \{\pi_i\}$, the initial state probability distribution, representing probabilities of initial states, i.e.

$$\pi_i = P[Q_1 = S_i] \quad 1 \leqslant i \leqslant N$$

  with $\pi_i \geqslant 0$ and $\sum_{i=1}^{N} \pi_i = 1$.

For convenience, we denote a HMM as a triplet $\lambda = (\boldsymbol{A}, \boldsymbol{B}, \pi)$.

The training of the model, given a set of sequences $\{\boldsymbol{O}_i\}$, is performed using the standard Baum–Welch re-estimation procedure (Rabiner, 1989), able to determine the parameters $(\boldsymbol{A}, \boldsymbol{B}, \pi)$ that maximize the probability $P(\{\boldsymbol{O}_i\}|\lambda)$. This method is based on the well-known *Expectation Maximization* (EM) algorithm. The evaluation step, i.e. the computation of the probability $P(\boldsymbol{O}|\lambda)$, given a model λ and a sequence $\boldsymbol{O}$ to be evaluated, is performed using the *forward–backward procedure* (Rabiner, 1989).

A practical but fundamental issue to be addressed when using HMMs is the determination of their structure, namely the topology and the number of states. The former aspect regards the possibility of introducing some constraints in the HMM structure. No assumptions about the topology have been made in this paper, letting it free to be determined by the transition matrix derived from the training strategy. More interesting is the latter aspect, regarding the determination of the number of the states: this choice is the first a fundamental step in the model selection, mainly preventing overtraining situations. Unfortunately, even though good theoretical approaches have been proposed (Stolcke and Omohundro, 1993; Brand, 1999; Bicego et al., 2003b), the aforesaid issue is usually disregarded in the HMM literature. Another important issue is the initialization of the training procedure: this issue is crucial to the learning, because of the local behavior of the standard procedure used to estimate HMM parameters. Starting from some initial estimate, this technique converges to the nearest local maximum of the likelihood function, which is highly multimodal; therefore, a good initialization is needed to guarantee the convergence of the procedure to the global optimum. Both model selection and initialization issues have been addressed in this paper, making the learning particularly effective.

3.2. Wavelets

Wavelets can be defined as a mathematical method for hierarchically decomposing functions. The wavelet transform is aimed at describing a function in terms of a *coarse* overall shape, plus *details* that range from broad to narrow. The basic idea is to represent any arbitrary function $f(t)$ as a superposition of a set of basis functions. In particular, basis functions related to *coarse* coefficients are called *scaling functions*, while those related to *detail* coefficients are called *wavelets functions*. The *wavelets*, or *baby wavelets*, are obtained from a single prototype wavelet called the *mother wavelet*, by dilations or contractions (scaling) and translations (shifts). In this paper, we used the Haar wavelets (De Vore et al., 1992), representing the simplest wavelet basis. The advantage of wavelet transform is that often a large number of the *detail* coefficients turns out to be very small in magnitude. Truncating, or removing, these small coefficients from the representation introduces only small errors in the reconstructed image, giving a form of *lossy* image compression. The wavelet transform has been successfully applied in many contexts, especially in the field of image compression (De Vore et al., 1992). Wavelet-based coding provides substantial improvements in picture quality at higher compression ratios, with respect to standard DCT transform. Over the past few years, a variety of powerful and sophisticated wavelet-based schemes for image compression have been developed and implemented (Saha, 2000). The effectiveness of the wavelet transform has been widely demonstrated.

For example, in (Garcia et al., 1998), a wavelet-based method for face recognition has been intro-

duced. Each face is described by a subset of band filtered images containing wavelet coefficients, then, a probabilistic metric derived from the Bhattacharya distance is used for classification. In (Wu and Bhanu, 1995; Krüger and Peters, 1997), Gabor Wavelets have been used to set up an invariant representation of objects. A Gabor grid efficiently encodes the structural information of an object in a sparse multiresolution representation. The Gabor grid subsamples the Gabor wavelets decomposition of an object model and is deformed to allow the indexed object model match with the image data. Finally, in (Reinhold et al., 2001), an appearance-based method for 3D object localization and recognition using wavelets has been described. Local features are derived from the wavelet multiresolution analysis and are modelled statistically by normal distributions. The localization and classification of the objects are then performed hierarchically by maximum-likelihood estimation.

In this paper, wavelets coefficients are extracted as local descriptors, so as to improve robustness with respect to noise and lighting changes, while retaining the ability in grabbing essential information of the signal, by discarding non-important parts. They are used together with HMMs, an outstanding method able to capture the sequential nature of data, which, in this case, succeeds to describe the shape of an object from an unrolled sequence of its wavelet coefficients. In this way, a powerful framework for object classification can be constructed. Furthermore, strong arguments for the use of multiresolution decomposition can be found in psychovisual research which offers evidence that the human visual system processes the images in a multiscale way (Reinhold et al., 2001).

4. The proposed approach

In this section, the proposed method is detailed. In particular, the coding procedure is described in Section 4.1, and the classification strategy is detailed in Section 4.2.

4.1. The coding strategy

The strategy used to obtain the data sequence from an object image consists of three steps. First, the image is converted from the color format to the grey level format. This is important to assess the capability of the proposed approach in capturing the geometry of the object, rather than the color. In the second step, a sequence of sub-images of fixed dimension is obtained by sliding over the object image, in a raster scan fashion, a square window of fixed size, with a predefined overlap. In this way we could capture relevant information about the local geometry of the object to be encoded: the sequence of subsequent windows summarizes the local object structure. The procedure for scanning the image is visualized in Fig. 1. Finally, the third step consists in applying the wavelet transform to each gathered sub-image. The proposed algorithm calculates the coefficients representing the image with a normalized two-dimensional Haar basis, sorting these coefficients in order of decreasing magnitude. Subsequently, the first M coefficients (i.e., the coefficients with higher magnitude) are retained, performing a lossy image (sub-image) compression. As for image compression, the retained coefficients represent the more significative information. Hence, we use them to recognize the objects. In particular, the number

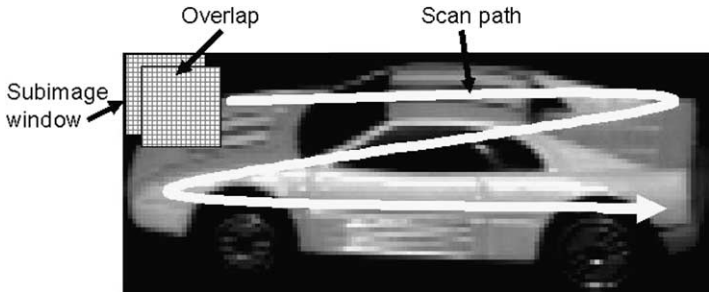

Fig. 1. Sampling scheme to generate the sequence of sub-images.

of retained coefficients determines the dimensionality of the observation vector (i.e. the *local descriptor*), while its length is determined by the number of sub-images gathered. By applying this step to all the sub-images of the sequence, we finally get the actual sequence observation. Its dimensionality will be $M \times T$, where M is the number of the wavelet coefficients retained, and T is the number of sub-images gathered in the sample scanning operation.

4.2. The recognition strategy

The standard way to use HMM is to train one model for each class. The subsequent classification is performed using standard maximum likelihood classification rule, i.e., assigning an unknown item to the class whose model shows the highest likelihood. A different rule has been used in this paper: instead of training one model for each class, i.e., using all the object views to train a model, we train *one* model for *each* object view. The classification step is performed by assigning an unknown object view to the class of the model showing the maximum likelihood. Notice that this method may be considered as a nearest neighbor (NN) classifier, with the proximity measure defined by the likelihood function. The use of HMM to compute distance between sequences is not new in the literature, as it has already been used in the context of HMM-based clustering of sequences (Smyth, 1997; Panuccio et al., 2002). This classification scheme seems very suitable for the object recognition problem: to identify an object given an aspect, we look for the view most similar to it, which, probably, is represented by one of the near views of the same object. In our case, after training one model for each object aspect, an unknown view is assigned to the class of the most similar view, where similarity is computed by using the likelihood of the HMM. A great advantage of this scheme is that if a new view of the object is added, we should not re-train the class model.

Regarding the HMM training, particular attention was devoted to the model selection and the initialization issues. These issues have been addressed by using a method recently introduced in (Bicego et al., 2003b). The proposed technique is able to deal with drawbacks of standard general purpose methods, like those based on the *Bayesian inference criterion* (BIC) (Schwarz, 1978), i.e., computational requirements, and sensitivity to initialization of the training procedure. The basic idea is to perform a ''decreasing'' learning, starting each training session from an informative situation derived from the previous training phase. More specifically, the procedure consists in starting the model training using a large number of states, run the estimation algorithm, and, after convergence, evaluate the chosen model selection criterion for that model. In this case the BIC criterion was used. Then, the importance of each model state is determined, using the stationary distribution of the Markov Chain associated to the HMM. Finally, the ''least probable'' state is pruned, and this configuration is taken as initial situation from which to start again the training procedure. In this way, each training session is started from a ''nearly good'' estimate. This permits to obtain better estimates for the model, increasing the efficacy of the proposed approach. Moreover, by starting from a good situation, the number of iterations required by the training algorithm to converge is reduced, resulting in a less computationally demanding procedure. In our approach we used Gaussian HMMs, where the emission probability of each state is modelled using the Gaussian function. Learning is finally performed using standard Baum Welch procedure, stopping the procedure after likelihood convergence.

5. Experimental results

The proposed approach has been thoroughly tested on the images of the COIL-20 database (Nene et al., 1996). This public database has been largely used in object recognition literature. It contains 20 objects: for each object 72 views are gathered, with a separation of 5°. The objects of the data set are presented in Fig. 2. As an example, the 72 views of the first object are displayed in Fig. 3.

We tested our approach by varying the free parameters of the techniques, i.e., the window size, the overlap ratio, and the number of the retained

Fig. 2. The objects contained in the COIL-20 dataset.

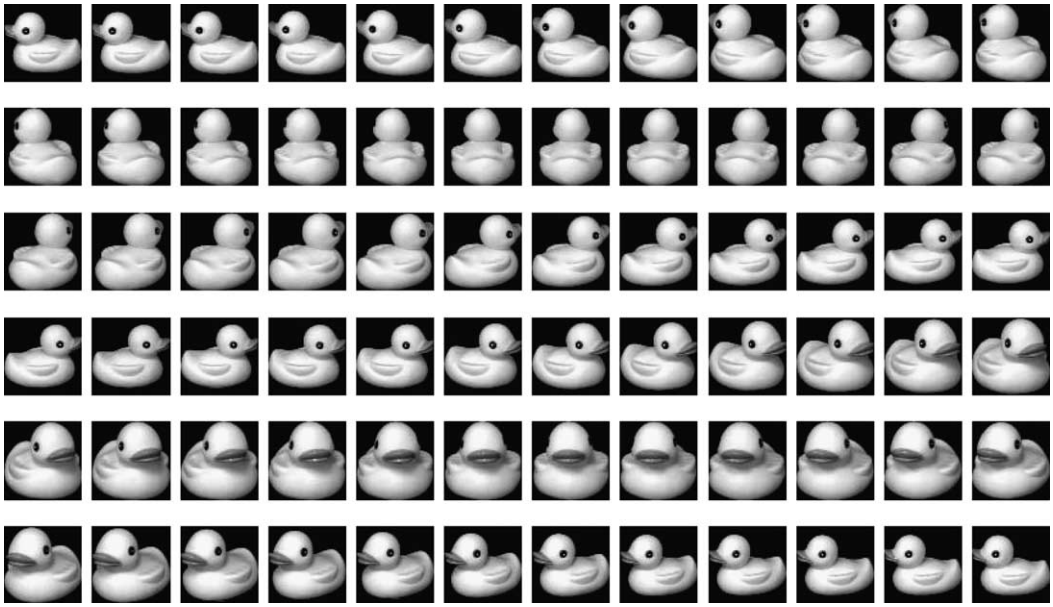

Fig. 3. The 72 views of the first object in the COIL-20 dataset.

coefficients. The classification accuracy was computed using the cross-validation holdout scheme (Theodoridis and Koutroumbas, 1999): the dataset is divided in two mutually disjoint sets, one used for the training phase and one used for the testing. The fact that the system is tested on a set different from the one used for training is a crucial factor for assessing the capability of the system in generalizing to different views. In particular, odd views are used to build the classifier and even views are used for testing. Classification accuracies are presented in Table 1(a) and (b), for windows sizes equal to 8 and 16, respectively.

From these tables, one can notice that results are really satisfactory: for one parameters' configurations the system is able to perfectly recognize all the objects. In particular, for a window size equal to 8 and an overlapping ratio of 75%, the system performances are very high, almost perfect. Moreover, the system seems to better perform using a small window. Probably this is due to the fact that a longer sequence is better modelled by the HMM, which has at disposal more data usable for a more accurate model selection estimation. From this table it is worth noticing that the system is very effective also using few wavelet coefficients. Actually, this is reasonable as, using few coefficients, the signal is quite rough, even if sufficiently informative to discriminate between objects, and only the general behavior is understood. This

Table 1
Classification accuracies for different overlapping ratios and number of the considered coefficients

Number of coefficients	Overlap ratio	Classification accuracies (%)
(a)		
4	0.5	98.47
5	0.5	98.33
6	0.5	98.33
7	0.5	97.78
4	0.75	100
5	0.75	98.61
6	0.75	98.33
7	0.75	97.22
(b)		
4	0.5	96.67
5	0.5	97.5
6	0.5	96.53
7	0.5	96.81
4	0.75	97.92
5	0.75	97.64
6	0.75	98.61
7	0.75	97.92

The sampling image size was fixed to (a) 8×8 pixels and (b) 16×16 pixels. For these experiments, odd views are used to build the classifier and even views are used for testing.

permits the system to generalize better, (i.e., not focusing on the specific view) by trying to capture the general behavior.

5.1. Occlusions

The proposed approach has also been tested in case of occlusion. Occlusion is one of the most severe problem in the object recognition context, and is the principal cause of failure of several approaches. The object occlusion has been simulated in two ways. The first way, which we called "random occlusion", follows the approach proposed in (Pontil and Verri, 1998): a fixed size windows of noisy pixels has been randomly generated, and located over the object. The position of this location is random, this fact increasing the statistical robustness of the evaluation. In the second way, which is more realistic and more difficult, the occluding window represents a patch extracted from another object of the dataset, simulating a real occlusion (an actual object occluding the analyzed view). In this case, for each view, the occluding patch was extracted from a randomly chosen view of a randomly chosen object, increasing even further the statistical robustness of the evaluation. We called this approach "patch occlusion". The dimension of the occluding windows has varied from 16 to 40: some examples of occluded objects with both approaches are presented in Fig. 4 (first and second rows, respectively), for different occluding levels (dimensions of the occluding window).

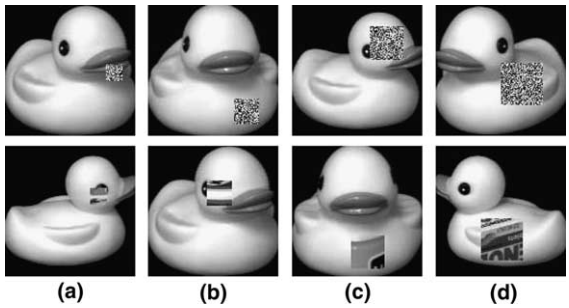

Fig. 4. Examples of occluded objects for different occlusion levels: (a) 16×16; (b) 24×24; (c) 32×32; (d) 40×40. The examples shown in the first row are obtained with the "random occlusion", while in the second the "patch occlusion" method was applied.

In this occlusion experiments, only odd views were used to build the classifier, using the occluded versions of the even ones for testing. For each object, 10 views have been randomly selected and occluded. The results are then computed using the best configuration of parameters derived from the previous analysis, i.e. using a scanning window of size 8×8 pixels, with overlap of 75% and retaining four coefficients. Accuracies for both types of occlusion are presented in Table 2. These results are really satisfactory: even if the objects are largely occluded, the system is able to recognize them. Clearly, the system performs worse when objects

Table 2
Classification accuracies for the occluded objects, for different occlusion levels

Occlusion level	Random occlusion (%)	Patch occlusion (%)
16	99.00	98.00
24	98.50	98.00
32	97.00	92.00
40	92.50	90.00

are occluded using patches from actual objects, still remaining at really satisfactory levels. The ability of HMM-based methods to recognize occluded shapes has been already demonstrated in the case of 2D shapes (Bicego and Murino, 2004).

5.2. Distortions

This section investigates the robustness of the proposed approach to image distortion, which could happen when changing the len of the video camera. The testing images have been perturbed using a radial distortion.[2] As in the case of occlusions, odd views were used for training, while the distorted versions of the even ones were employed for testing. For each object, 10 views have been randomly selected and affected by radial distortion. Some examples of distorted images are presented in Fig. 5, for different values of the distortion parameter K. In the figure, in order to have a direct visualization of the effect of the perturbation, also the differences between the original images and the distorted ones are displayed. Classification accuracies are proposed in Table 3: the system is very robust, also in case of highly perturbing distortion, but, obviously, the increase of the perturbation level decreases the accuracy of the system, which however remains at satisfactory levels.

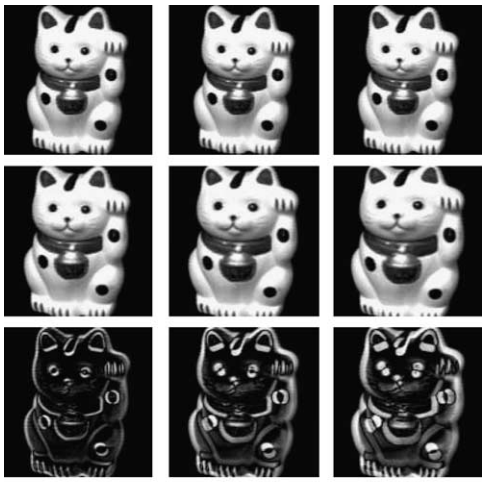

Fig. 5. Effect of distortion for different levels of distortion: (first row) original image; (second row) distorted one; (third row) difference between them.

Table 3
Accuracy on the distorsion experiment, for different values of the parameter K

Parameter K	Accuracy (%)
0.10	97.5
0.20	97.0
0.30	93.5

5.3. Number of training views

In order to get a better insight into the method, we have performed a further analysis on the COIL database in order to assess the system performances when trained with a reduced number of views. Therefore, we have performed an experiment using, in the training phase, a decreasing number of views for each object. Results are displayed in Table 4 in function of the number of training views. As before, we used the best parameters determined in the previous analysis. We also show the separation degree, that is the angle between two consecutive views. Results are really interesting, showing that the method is very robust. Actually, even if we use only one view every 45° (only eight views in total), the system is able to recognize the objects with an accuracy larger than 90%. This finding has a great practical importance, demonstrating that the proposed system can be

Table 4
Classification accuracies while decreasing the number of views used for training

Number of views used for training	Separating degree	Classification accuracy (%)
72	5	100.0
36	10	100.0
18	20	96.46
8	45	91.94
4	90	76.04
2	180	67.22

[2] We used the `Qlensdistort.m` function of the 'Q' Software package, a MATLAB image processing toolbox downloadable from `http://www.cs.dartmouth.edu/~farid/tutorials`.

trained using very few aspects, so reducing the computational and storage requirements.

5.4. Discussion

This final section contains some further discussions about the method presented in this paper. The main goal of the paper is to investigate HMM capabilities in 3D objects recognition: in particular, the method is mainly focused on the classification part, i.e. on the recognition of the object *given* the segmented object; the problem of accurately segmenting the object from the background is out of the scope of this paper. With respect to the classification issue, the proposed approach has several appealing characteristics, such as the scalability (the possibility of adding new views or new objects to the recognition framework without the need of retraining the whole system), and the simplicity of application, given by the nearest neighbor rule in the likelihood space. Moreover the experimental evaluation has shown that the system is particularly robust to perturbations, like radial distortion or occlusions, and it is scalable, since the number of views required for training could be reduced without a significant loss of accuracy.

6. Conclusions

In this paper, a new method for appearance-based 3D object recognition has been proposed, based on the Hidden Markov Model approach. The view of an object is visited in a raster-scan fashion to obtain a sequence of partially overlapped sub-images. For each sub-image, wavelet coefficients are computed, the most significant are retained, and, finally, arranged to compose a feature vector. The sequences of vectors (one for each view) are subsequently modelled using HMMs, paying particular attention to the initialization and the model selection issues. Classification is carried out by using a nearest neighbor rule, where distance is computed using the HMM likelihood function. A thorough experimental evaluation has shown that the proposed approach is very promising for classifying 3D objects from partial, non-dense, set of views. Furthermore, the proposed method remains quite accurate even in case of heavily occluded or distorted objects.

An interesting extension of the method could go toward the investigation of the use of the 3D information obtained by acquiring the objects with a 3D scanner. In particular, we are investigating the integration of appearance and range data for recognition.

References

Arica, N., Yarman-Vural, F., 2000. A shape descriptor based on circular Hidden Markov Model. In: IEEE Proc. Int. Conf. on Pattern Recognition, vol. 1, pp. 924–927.

Arman, F., Aggarwal, J.K., 1993. Model-based object recognition in dense-range images: A review. ACM Comput. Surveys 25, 5–43.

Barla, A., Franceschi, E., Odone, F., Verri, A., 2002. Image kernels. In: Proc. Int. Workshop on Pattern Recognition with Support Vector Machines. ICPR 2002, pp. 83–95.

Batlle, J., Casals, A., Freixenet, J., Marti, J., 2000. A review on strategies for recognizing natural objects in colour images of outdoor scenes. Image Vision Comput. 18 (6–7), 515–530.

Baum, L., 1970. An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes. Inequality 3, 1–8.

Baum, L., Petrie, T., Soules, G., Weiss, N., 1970. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. Ann. Math. Statist. 41 (1), 164–171.

Belongie, S., Malik, J., Puzicha, J., 2002. Shape matching and object recognition using shape contexts. IEEE Trans. Pattern Anal. Mach. Intell. 24 (24), 509–522.

Bicego, M., Murino, V., 2004. Investigating Hidden Markov Models' capabilities in 2D shape classification. IEEE Trans. Pattern Anal. Mach. Intell. 26 (2), 281–286.

Bicego, M., Castellani, U., Murino, V., 2003a. Using Hidden Markov Models and wavelets for face recognition. In: Proc. Int. Conf. on Image Analysis and Processing, pp. 52–56.

Bicego, M., Murino, V., Figueiredo, M., 2003b. A sequential pruning strategy for the selection of the number of states in Hidden Markov Models. Pattern Recognition Lett. 24 (9–10), 1395–1407.

Bischof, H., Wildenauer, H., Leonardis, A., 2001. Illumination insensitive eigenspaces. In: IEEE Int. Conf. on Computer Vision, ICCV 2001, pp. 233–238.

Brand, M., 1999. An entropic estimator for structure discovery. In: Kearns, M., Solla, S., Cohn, D. (Eds.), Advances in Neural Information Processing Systems, vol. 11. MIT Press, pp. 723–729.

Bulthoff, H., Edelman, S., Tarr, M., 1994. How are three-dimensional objects represented in the brain. Tech. rep. Artificial Intelligence Lab, Massachussets Institute of Technology.

Cai, J., Liu, Z.-Q., 2001. Hidden Markov Models with spectral features for 2D shape recognition. IEEE Trans. Pattern Anal. Mach. Intell. 23 (12), 1454–1458.

Campbell, R., Flynn, P., 2001. A survey of free-form object representation and recognition techniques. Computer Vision and Image Understanding 81, 166–210.

Caputo, B., Bouattour, S., Niemann, H., 2002. Robust appearance-based object recognition using a fully connected Markov random field. In: Proc. Int. Conf. on Pattern Recognition, ICPR 2002, pp. 11–15.

Cyr, C.M., Kimia, B.B., 2001. 3D object recognition using shape similarity-based aspect graph. In: Proc. Int. Conf. on Computer Vision. Vancouver, Canada, pp. 254–261.

de Verdiere, V.C., Crowley, J.L., 1998. Visual recognition using local appearance. In: Proc. Eur. Conf. on Computer Vision, pp. 640–654.

De Vore, R., Jawerth, B., Lucier, B., 1992. Image compression through wavelet transform coding. IEEE Trans. Inform. Theory 38 (2), 719–746.

Duda, R., Hart, P., Stork, D., 2001. Pattern Classification, second ed. John Wiley and Sons.

Fred, A., Marques, J., Jorge, P., 1997. Hidden Markov Models vs. syntactic modeling in object recognition. In: IEEE Proc. Int. Conf. on Image Processing, vol. 1, pp. 893–896.

Garcia, C., Zikos, G., Tziritas, G., 1998. A wavelet-based framework for face recognition. In: Proc. Eur. Conf. on Computer Vision, ECCV 1998, pp. 1–7.

Hagedoorn, M., 2000. Pattern Matching using Similarity Measures. Ph.D. Thesis. Universiteit Utrecht.

Ham, Y., Park, R.-H., 1999. 3D object recognition in range images using hidden Markov models and neural networks. Pattern Recognition 32 (5), 729–742.

He, Y., Kundu, A., 1991. 2-D shape classification using Hidden Markov Model. IEEE Trans. Pattern Anal. Mach. Intell. 13 (11), 1172–1184.

Huttenlocher, D., Klanderman, G., Rucklidge, W., 1993. Comparing images using the Hausdorff distance. IEEE Trans. Pattern Anal. Mach. Intell. 15 (9), 850–863.

Krüger, N., Peters, G., 1997. Object recognition with banana wavelets. In: Proc. Eur. Symp. on Artificial Neural Networks, pp. 61–66.

Leonardis, A., Bischof, H., 2000. Robust recognition using eigenimages. Computer Vision and Image Understanding 78 (1), 99–118.

Loncaric, S., 1998. A survey of shape analysis techniques. Pattern Recognition 31 (8), 983–1001.

Moghaddam, B., Guillamet, D., Vitria, J., 2003. Local appearance-based models using high-order statistics of image features. In: Proc. Int. Conf. on Computer Vision and Pattern Recognition, pp. 729–735.

Murase, H., Nayar, S., 1995. Visual learning and recognition of 3D objects from appearance. Internat. J. Comput. Vision 14 (1), 5–24.

Nene, S.A., Nayar, S.K., Murase, H., 1996. Columbia Object Image Library (COIL-20). Tech. Rep. CUCS-005-96, Columbia University.

Panuccio, A., Bicego, M., Murino, V., 2002. A Hidden Markov Model-based approach to sequential data clustering. In: Caelli, T., Amin, A., Duin, R., Kamel, M., de Ridder, D. (Eds.), Structural, Syntactic and Statistical Pattern Recognition, LNCS 2396. Springer, pp. 734–742.

Pontil, M., Verri, A., 1998. Support vector machines for 3-D object recognition. IEEE Trans. Pattern Anal. Mach. Intell. 20, 637–646.

Pope, A., 1994. Model-based object recognition—a survey of recent research. Tech. Rep. 9404. UBC Department of Computer Science.

Pope, A., Lowe, D., 2000. Probabilistic models of appearance for 3-d object recognition. Internat. J. Comput. Vision 40 (2), 149–167.

Prokop, R., Reeves, A., 1992. A survey of moment-based techniques for unoccluded object representation and recognition. CVGIP—Graphical Models Image Process. 54 (5), 438–460.

Rabiner, L., 1989. A tutorial on Hidden Markov Models and selected applications in speech recognition. Proc. IEEE 77 (2), 257–286.

Reinhold, M., Paulus, D., Niemann, H., 2001. Improved appearance-based 3d object recognition using wavelets. In: Proc. Vision, Modelling and Visualization, VMV 2001, pp. 473–480.

Roobaert, D., Zillich, M., Eklundh, J., 2001. A pure learning approach to background invariant object recognition using pedagogical support vector learning. In: Proc. Int. Conf. on Computer Vision and Pattern Recognition, CVPR 2001, pp. 351–357.

Saha, S., 2000. Image compression—from DCT to wavelets: A review. ACM Crossroads Magazine.

Schneiderman, H., Kanade, T., 1998. Probabilistic modeling of local appearance and spatial relationship for object recognition. In: Proc. Int. Conf. on Computer Vision and Pattern Recognition, pp. 45–51.

Schwarz, G., 1978. Estimating the dimension of a model. Ann. Statist. 6 (2), 461–464.

Smyth, P., 1997. Clustering sequences with hidden Markov models. In: Mozer, M., Jordan, M., Petsche, T. (Eds.), Advances in Neural Information Processing Systems, vol. 9. MIT Press, p. 648.

Stolcke, A., Omohundro, S., 1993. Hidden Markov Model induction by Bayesian model merging. In: Hanson, S., Cowan, J., Giles, C. (Eds.), Advances in Neural Information Processing Systems, vol. 5. Morgan Kaufmann, San Mateo, CA, pp. 11–18.

Theodoridis, S., Koutroumbas, K., 1999. Pattern Recognition. Academic Press.

Trazegnies, C., Urdiales, C., Bandera, A., Sandoval, F., 2003. 3D object recognition based on curvature information of planar views. Pattern Recognition 36 (11), 2571–2584.

Ullman, S., Basri, R., 1991. Recognition by linear combination of models. IEEE Trans. Pattern Anal. Mach. Intell. 13, 991–1006.

Wu, X., Bhanu, B., 1995. Gabor wavelets for 3-d object recognition. In: Proc. Int. Conf. on Computer Vision, p. 537.