Automated Face Identification Using Hidden Markov Models

F. Samaria[a,b] and F. Fallside[a]

[a]Cambridge University Engineering Department, Trumpington Street,
Cambridge CB2 1PZ, United Kingdom, tel.+44 223 332752

[b]Olivetti Research Lab, Old Addenbrookes Site, 24a Trumpington Street,
Cambridge CB2 1QA, United Kingdom, tel.+44 223 343000

Abstract

This paper details work done on automatic face identification. A new approach to the problem was proposed [8] involving the use of Hidden Markov Models. Initial experimental results [9] indicated that left-to-right models with use of structural information yielded better feature extraction than ergodic models. In this paper we illustrate how these hybrid models can be used to extract facial bands and classify face images, showing the benefits of simultaneous use of statistical and structural information. Some experimental results obtained with a simple left-to-right model are presented to support the plausibility of this approach. Successful results were obtained using images with homogeneous background. We conclude indicating present and future directions of research work using these models.

1 INTRODUCTION

Of the five human senses, *sight* seems to be predominant over the others. Vision is an essential activity that human beings perform whenever interacting with their environment in any sort of meaningful way [12]. *Recognition* is the fundamental aim of any visual system. It describes the process of labeling a test object as belonging to a certain object-class. However, if the object is a human face, we are not only interested in recognising that our friend's face is a visual item belonging to the class of faces, but that he or she is indeed our friend. We are required to identify a particular instance of the object within the class. We call this process face identification, where other authors in the literature have often used the term face recognition [1]. The difference between object recognition and object identification is illustrated in figure 1.

Substantial research effort has gone into trying to understand how to build a successful model for face identification, both by psychologists and information scientists. The importance of face identification, other than for the challenge of the task, stems from its numerous potential applications. A successful system could be used for building security, workstation security [2], criminal identification, credit card verification, video-mail retrieval and home video actor-based indexing, where large databases of movies stored on CD-ROMs could be indexed and accessed using actors' faces.

We present a new architecture to describe facial features that uses continuous density Hidden Markov Models (HMMs) [7]. Faces are treated as 2-dimensional objects and the model automatically extracts *statistical* facial features. However, it can at the same time use *structural* information, yielding a *hybrid* model.

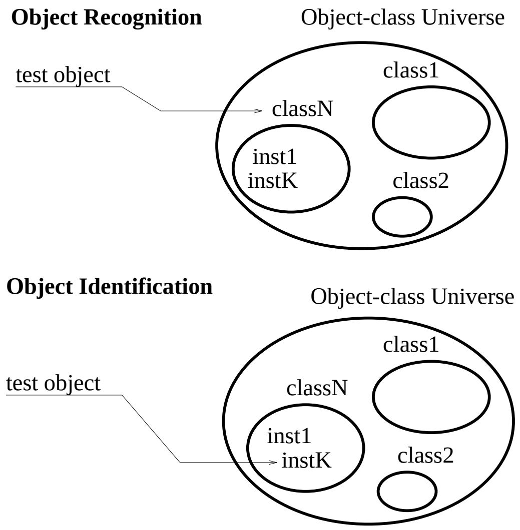

Figure 1: Object Recognition vs Object Identification

The following sections present the proposed approach, a general overview of HMMs and some experimental results. We conclude the paper indicating the strengths of the approach and future areas of research work.

2 PROPOSED APPROACH

The goal of a face identification system is to be able to match a given test face with a specific subject from a database of faces. We propose a method that automatically extracts facial features from a set of training data (database) and successively uses them to identify test images. We achieve this using a particular context-dependent classifier [11] based on HMMs.

2.1 Introducing HMMs

HMMs are a set of statistical models used to characterise the statistical properties of a signal. Rabiner [7] provides an extensive and complete tutorial on HMMs. They are made of two interrelated processes:

(1) an underlying, unobservable Markov chain with a finite number of states, a state transition probability matrix and an initial state probability distribution;

(2) a set of probability density functions associated with each state.

We can formally define the elements of an HMM by specifying the following parameters:

- N is the number of states in the model. If $\mathbf{S}$ is the set of states, then $\mathbf{S} = \{S_1, S_2, \ldots, S_N\}$. The state of the model at time t is given by q_t, $1 \leq t \leq T$ where T is the length of the observation sequence (number of frames).
- M is the number of different observation symbols. If $\mathbf{V}$ is the set of all the possible observation symbols (also called the *code book* of the model), then $\mathbf{V} = \{v_1, v_2, \ldots, v_M\}$.
- $\mathbf{A}$ is the state transition probability matrix, *i.e.* $\mathbf{A} = \{a_{ij}\}$ where:

$$a_{ij} = Pr[q_t = S_j \mid q_{t-1} = S_i], \quad 1 \leq i, j \leq N$$
$$a_{ij} \geq 0$$
$$\sum_{j=1}^{N} a_{ij} = 1, \qquad 1 \leq i \leq N$$

- $\mathbf{B}$ is the observation symbol probability matrix, *i.e.* $\mathbf{B} = \{b_j(k)\}$ where:

$$b_j(k) = Pr[O_t = v_k \mid q_t = S_j] \quad \begin{array}{l} 1 \le j \le N \\ 1 \le k \le M \end{array}$$

  and O_t is the observation symbol at time t.

- π is the initial state distribution, *i.e.* $\pi = \{\pi_i\}$, where:

$$\pi_i = Pr[q_1 = S_i], \quad 1 \le i \le N$$

Using shorthand notation, an HMM is defined as:

$$\lambda = (\mathbf{A}, \mathbf{B}, \pi)$$

2.2 HMMs And Images

HMMs have been extensively and successfully used for speech recognition applications [6]. The one dimensional (1D) nature of speech signals along the time axis is suited to analysis by HMMs, which require a 1D observation sequence. Images are two dimensional (2D), but they can be converted into either 1D temporal sequences or 1D spatial sequences as discussed in [8], where it was concluded that spatial sequences are to be preferred. A spatial sequence may be obtained by sampling the image using a sliding window: this converts the image into a 1D sequence of windowed data where each element of the sequence is a vector made of a certain number of samples (depending on the size of the sliding window). The size of the window and the direction of sampling may vary. Initial results [9] showed that top-to-bottom line block sampling gives psychologically significant features.

2.3 Extracting Facial Bands

Consider a training set made of a collection of face images for each subject. Let us say that we have F distinct subjects and W frontal view images for each of them (*i.e.* W is the size of the training set for each face and $W \times F$ images is the total size of the database). The sampling technique described in figure 2 allows to convert each still image into a sequence of vectors $\mathbf{O} = \mathbf{O}_1\mathbf{O}_2 \ldots \mathbf{O}_T$, which are spatially ordered in a top-to-bottom sense.

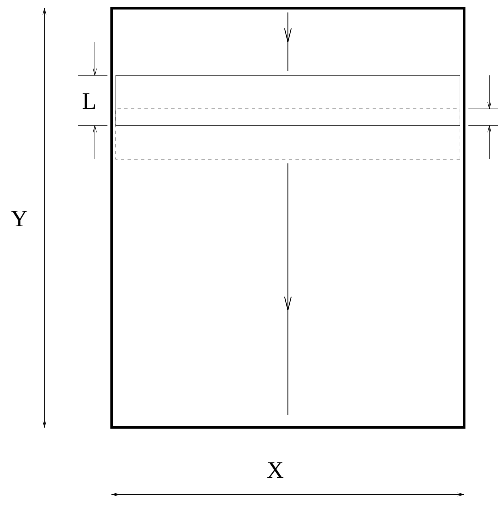

Figure 2: Sampling Technique

Using this technique each image is sampled into a sequence of overlapping line blocks. The overlapping is present to allow the model to capture features independently of their vertical position (a disjoint partitioning of the image could result in features occurring across block boundaries being truncated). Each line block represents one observation vector (of size $X \times L$, where L is the number of lines in each block) and this form of sampling will result into a 1D sequence of observation vectors $\mathbf{O} = \mathbf{O}_1 \mathbf{O}_2 \ldots \mathbf{O}_T$ where T can be calculated as:

$$T = \frac{Y - L}{L - V} + 1$$

and V is the amount of overlap (expressed in number of lines) between successive blocks.

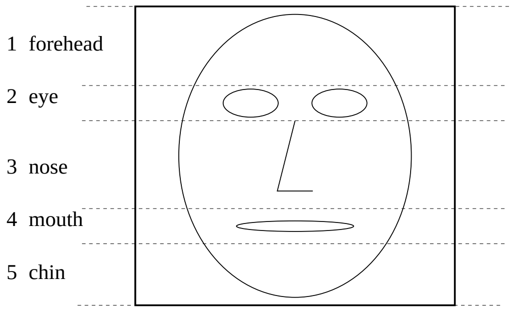

Figure 3: Facial Bands For 5-state HMM

This sequence can be fed to an HMM that will then partition it into a number of states, each of which will represent a facial band. The terms "facial band" here indicate a horizontal block (of size L) which has sufficient statistical significance to result in the HMM assigning a state to represent it. We wish to establish a correspondence between statistical facial bands and feature bands as understood by humans. Assuming that each face is in an upright, frontal position, feature bands would occur in a predictable order, *i.e.* eyes following forehead, nose following eyes, *etc.* (see figure 3).

This form of ordering suggests to use a left-to-right HMM where only transitions between adjacent states in a left-to-right manner will be allowed. Figure 4 shows the corresponding state diagram.

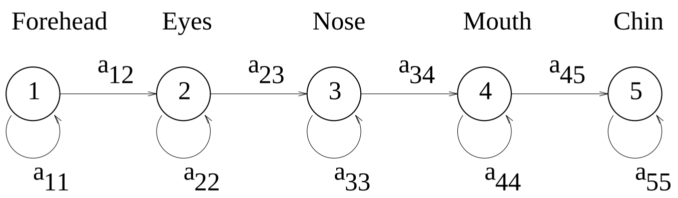

Figure 4: Left-to-right HMM

We can train one HMM per each subject in the database and obtain a left-to-right HMM

$$\lambda^{(k)} = (\mathbf{A}, \mathbf{B}, \pi), 1 \leq k \leq F$$

where F is the total number of different subjects in the database. The elements defining the HMM would have meanings as follows:

A: measures the probability of going from one facial band to another. After training, $\mathbf{A}^{(k)}$ will record the occurrences of transitions from one band to another across the face and the thickness of the various bands (this can be seen by considering the a_{ii} terms which represent the probability of staying in the same facial band).

B: measures the probability of observing a particular feature vector, given that we are looking at a particular facial band. After training, $\mathbf{B}^{(k)}$ will record the feature vector distribution per subject across the various bands.

π: all observations sequences start in the top band and therefore, with the present model, this parameter does not provide any discriminating information ($\pi_1 = 1$ and $\pi_i = 0, \forall i \neq 1$).

3 INITIAL EXPERIMENTAL RESULTS

A simple spatial HMM was built to test some of the ideas presented in the previous sections. A separate HMM was trained per each subject in the database and the resulting models were used to classify unknown images.

3.1 Training The HMM

A database comprising of 20 subjects was used and a total of 5 training images (of size 184×224, 8-bit grey levels) per subject were used to train each HMM. Each image was sampled into an observation sequence using a block of 16 lines, moving vertically 4 lines at a time (*i.e.* with a 12-line overlap). Using the parameters of figure 2, we have $X = 184, Y = 224, L = 16, V = 12$. For each subject, 5 such sequences (one per training image) were generated and used to train a 5-state left-to-right HMM. The reason for choosing the number of states to be 5 was that, by inspection, approximately 5 distinct bands seem to appear in the face image (forehead, eyes, nose, mouth and chin).

Figure 5: Training Data For Subject_1

The training and testing were carried out using the *HTK: Hidden Markov Model Toolkit V1.3* written at the Cambridge University Engineering Department by Steve Young [14]. The training process consisted of the following steps:

1. Create a *prototype* HMM model. The prototype serves the purpose of specifying the number of states in the HMM, the state transitions allowed and the size of the observation sequence vectors.
2. Compute iteratively a set of initial parameter values using the training data. On the first cycle, the data is uniformly segmented and matched with each model state. This is done using *segmental k-means* clustering [7]. On successive cycles, the uniform segmentation is replaced by Viterbi alignment [14].
3. Re-estimate parameters using Baum-Welch re-estimation [7]. The model parameters are adjusted so as to maximise the probability of observing the given training data, given each corresponding model.

Figure 5 shows the training data for one subject in the database. Figure 6 shows the mean vectors for the 5 states found by the HMM for the particular subject after training.

It is interesting to note that we can match some of the states with physical facial features. We can, for example, identify the second state as being the eye band, as predicted in the previous section. This means that the eyes have sufficient statistical significance to induce the HMM to assign a state to represent them. Experiments [10] have shown evidence that one of the salient features for identification used by humans is the eye region of the face. The remaining states extracted by the HMM can also be identified with other facial regions. Some states "leak" into the next one: it is possible, for example, to see part of the eye state lightly appearing in the next band. This is due to those observation vectors (or line blocks) which have, as far as state allocation is concerned, ambiguous statistical properties and give a better statistical match if placed in the next state. They will of course contribute to the overall state mean causing the leaking effect. Previous experiments [9] have shown that left-to-right models perform better feature extraction than ergodic HMMs, which make no use of structural information.

Figure 7 shows some of the other subjects from the face database and figure 8 shows the corresponding facial bands. It is again possible to spot physical facial features in the states extracted by the HMM, even when (like in the case of Subject_6) the training images reproduce the subject in different lighting circumstances and facial expressions.

3.2 Test Phase

The full database comprised of 20 subjects. In order to identify an unknown image, we converted the image into a 1D observation sequence $\mathbf{O}$ using the sampling technique previously illustrated. We matched the observation sequence against each of the 20 models in the database and computed the model likelihoods:

$$Pr[\mathbf{O} \mid \lambda^{(k)}],\ 1 \leq k \leq 20$$

The highest match was chosen and the person corresponding to the chosen model revealed the identity of the subject in the unknown image.

The 100 training images (5 pictures of 20 different subjects) were tested and correctly identified. Figure 9 shows some test data which was not used for training. These images were also correctly identified.

The following table shows the average log-likelihood per frame for the best match (which was correct for all 5 cases) and the second best match. The total number of frames was $T = 53$.

	Best Match	Second Best	Difference
Test_Image_1	-14,380	-14,635	255
Test_Image_2	-13,550	-14,405	855
Test_Image_3	-15,107	-16,068	961
Test_Image_4	-14,435	-15,669	1,234
Test_Image_5	-13,957	-15,685	1,728

Other experiments showed that the first training picture for Subject_6 was identified correctly with average log-likelihood of $-12,965$; an image representing a portion of a painting and an image from the space shuttle (shown in figure 10) scored an average log-likelihood of around $-27,500$ and $-18,750$ respectively. These values suggest that this approach could be used to detect the presence of a face in an image.

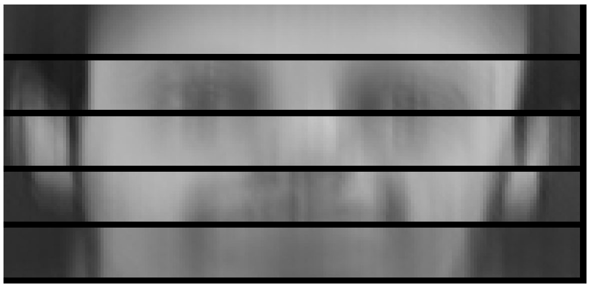

Figure 6: Magnified States For Subject_1

Figure 7: Training Data for Subject_2 To Subject_7

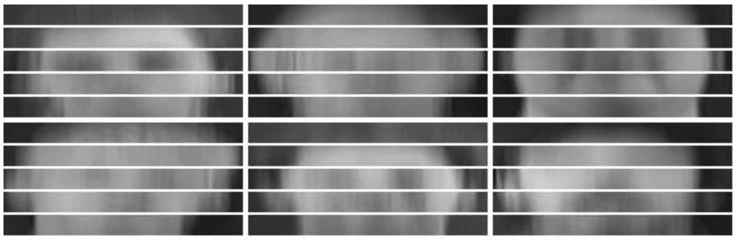

Figure 8: Facial Bands for Subject_2 To Subject_7

Some of the test images differ considerably from the training images and the successful results obtained indicate that the model exhibits a certain robustness to feature variation. Subject_2, for example, is correctly identified even when wearing glasses, despite not wearing them in the training images. Tests were carried out on about 50 more test images for various subjects extracted from the same video sequence as the training data. This guarantees similar lighting and background conditions; the images were again correctly identified. The results obtained for the images in figure 9, however, suggest that similar lighting conditions are not an essential constraint for the model to work successfully. If shown to hold for a larger database and test set, these results would constitute a strong advantage for the HMM based approach over other more traditional approaches. Recent work on human action recognition [13] also reports successful results using an HMM based method applied to images.

4 CONCLUSIONS

The approach proposed in this paper shows that HMMs provide a new method for automatic feature extraction and image classification. The statistical features obtained by the HMM have been shown to correspond to physical features as understood by humans, when structural information is used to build

Figure 9: Some Test Images

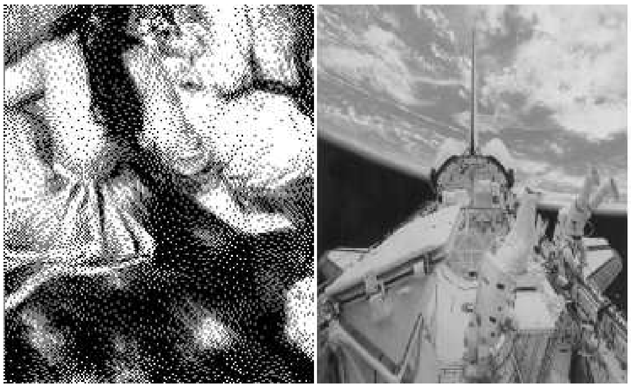

Figure 10: Test Images Not Containing A Face

the model. Other methods for feature extraction, such as Artificial Neural Networks [4], deformable templates [15] or snakes [5], have usually required considerable initial guidance. Successful identification results were obtained with relatively low constraints on the image data. Small orientation changes, non-homogeneous lighting and local feature variation (open/closed eyes, with glasses/without glasses) are dealt with automatically. Template-based models have generally suffered from such variations.

The method presented currently forces the extracted features to have a rectangular shape. It may be more appropriate to modify the shape of the sliding window (into hexagonal, say, covering the original face image with a grid of overlapping hexagons) and verify if better feature extraction could be achieved, thus improving the identification performance. The rectangular shape was originally chosen for ease of implementation.

The classifier at present will always output the best match from the available database. This would obviously result in an error whenever the given test face is not in the database or if the test image is not a face in the first place. A threshold level for the probability value of the best match can be set and it would be of interest to investigate how such a threshold level would have to be adjusted in order to detect the presence of a face in an image. Initial results indicate that the log-likelihood values are sufficiently apart and that faces could be detected using this approach.

Future work will investigate the use of HMMs for automatic face segmentation and the possibility of enhancing the facial bands using standard image processing techniques. Facial bands are effectively blurred images obtained by averaging all those line blocks which exhibit similar statistical properties. Image enhancement techniques [3] applied to facial band images may yield more suitable state mean vectors for classification.

Acknowledgements

This work is supported by a Trinity College Internal Graduate Studentship and an Olivetti Research Lab CASE award. Their support is gratefully acknowledged. We also wish to thank many of our colleagues in

Cambridge: Andy Hopper, Leslie French, Phil Elwell and Martin Brown of Olivetti Research, for useful discussions and for the image capture software; Steve Young and the Speech Group at CUED for the HMM software; Owen Jones and Gabor Megezy of the Pure Mathematics Department, Barney Pell of the Computer Lab, and Steve Hodges, Gopal Chand and Charlie Allen of CUED for the many ideas discussed together.

References

[1] V. Bruce. *Recognising Faces*. Laurence Erlbaum Associates, 1988.

[2] R. Gallery and T.I.P. Trew. An architecture for face classification. *IEE Colloquium on 'Machine Storage and Recognition of Faces', Digest No: 1992/017*, 2:1–5, 1992.

[3] R.C. Gonzalez and P. Wintz. *Digital Image Processing*. Addison-Wesley, second edition, 1987.

[4] R.A. Hutchinson and W.J. Welsh. Comparison of neural networks and conventional techniques for feature location in facial images. *IEE International Conference on Artificial Neural Networks, Conf. Publication Number 313*, 1989.

[5] M. Kass, A. Witkin, and D. Terzopoulos. Snakes: Active contour models. *ICCV*, pages 259–268, 1987.

[6] T.W. Parsons. *Voice and speech processing*. McGraw-Hill, 1986.

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

[8] F. Samaria. *Face Identification using Hidden Markov Models*. 1st Year Report, Cambridge University Engineering Department, 1992.

[9] F. Samaria and F. Fallside. Face identification and feature extraction using hidden markov models. In G. Vernazza, editor, *Image Processing: Theory and Applications*. Elsevier, 1993.

[10] J. Shepherd, G. Davies, and H. Ellis. Studies of cue saliency. In G. Davies, H. Ellis, and J. Shepherd, editors, *Perceiving and remembering faces*, pages 105–131. Academic Press, 1981.

[11] C.W. Therrien. *Decision estimation and classification*. Wiley, 1989.

[12] H. Wechsler. *Computational Vision*. Academic Press, San Diego CA, 1990.

[13] J. Yamato, J. Ohya, and K. Ishii. Recognizing human action in time-sequential images using hidden markov model. *Proceedings of CVPR'92*, pages 379–385, 1992.

[14] S.J. Young. *HTK: Hidden Markov Model Toolkit V1.3*. Reference Manual, Cambridge University Engineering Department, 1992.

[15] A.L. Yuille, P.W. Hallinan, and D.S. Cohen. Feature extraction from faces using deformable templates. *International Journal of Computer Vision*, 8,2:99–11, 1992.