Face Segmentation For Identification Using Hidden Markov Models

Ferdinando Samaria
(fs@eng.cam.ac.uk)
Cambridge University Engineering Department, Trumpington St.
Cambridge CB2 1PZ, United Kingdom, tel.+44 223 332752

Olivetti Research Ltd, Old Addenbrookes Site, 24a Trumpington St.
Cambridge CB2 1QA, United Kingdom, tel.+44 223 343000

Abstract

This paper details work done on face processing using a novel approach involving Hidden Markov Models. Experimental results from earlier work [14] indicated that left-to-right models with use of structural information yield better feature extraction than ergodic models. This paper illustrates how these hybrid models can be used to extract facial bands and automatically segment a face image into meaningful regions, showing the benefits of simultaneous use of statistical and structural information. It is shown how the segmented data can be used to identify different subjects. Successful segmentation and identification of face images was obtained, even when facial details (with/without glasses, smiling/non-smiling, open/closed eyes) were varied. Some experiments with a simple left-to-right model are presented to support the plausibility of this approach. Finally, present and future directions of research work using these models are indicated.

1 INTRODUCTION

Most of our social behaviour is dependent on the correct identification of the people surrounding us. Humans are generally able to infer information regarding sex, age and expression and use it reliably to identify faces. We are able to identify faces even when they are distorted (as in a caricature) or coarsely quantised, when they have occluded details and sometimes even when they have been inverted [2]. This process is here called face identification [13], where other authors in the literature have often used the term face recognition [1].

Faces play a fundamental role in social interactions and substantial research effort has gone into trying to understand how to build a successful model for face identification, both by psychologists and information scientists. Apart from its relevance to research into artificial intelligence, the importance of face identification stems from its numerous potential applications. A successful system could be used for building and workstation security [3], criminal identification, credit card verification and video-mail retrieval [5].

This paper presents a new architecture to describe facial features that uses continuous density Hidden Markov Models (HMMs) [11]. Face images are automatically segmented into horizontal regions, each of which is represented in the model by a "facial band" [13]. Faces are treated as two-dimensional objects and the segmentation is performed by extracting *statistical* facial features. However, by making simultaneous use of *structural* information (yielding a *hybrid* model), these statistical facial features are made to correspond to features as understood by humans.

The following sections present the proposed approach with a general overview of HMMs and some experimental results. The paper concludes with an indication of the strengths of the approach and areas of future research work.

2 HMM BASED APPROACH

A new method is proposed to automatically extract facial features from a set of training data (database) and successively use them to identify test images. This is achieved using a particular context-dependent classifier [16] based on HMMs.

2.1 Introducing HMMs

HMMs are a set of statistical models used to characterise the statistical properties of a signal. Rabiner [11] provides an extensive and complete tutorial on HMMs. The elements of an HMM can be formally defined by specifying the following parameters:

- $N = |S|$ is the number of states in the model, where $S = \{s_1, s_2, \ldots, s_N\}$ is the set of possible states s_i. The state of the model at time t is given by $q_t \in S$, $1 \leq t \leq T$ where T is the length of the observation sequence (number of frames).
- $M = |V|$ is the number of the different observation symbols, where $V = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_M\}$ is the set of all the possible observation symbols $\mathbf{v}_i$ (also called the *code book* of the model). The observation symbol at time t is given by $\mathbf{o}_t \in V$, $1 \leq t \leq T$.
- $\mathbf{A} = \{a_{ij}\}$ is the state transition probability matrix, where:

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

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

$$b_j(k) = Pr[\mathbf{o}_t = \mathbf{v}_k \mid q_t = s_j] \quad \begin{array}{l} 1 \leq j \leq N \\ 1 \leq k \leq M \end{array}$$

- $\pi = \{\pi_1, \pi_2, \ldots, \pi_N\}$ is the initial state distribution, where:

$$\pi_i = Pr[q_1 = s_i], \quad 1 \leq i \leq N$$

Using shorthand notation, an HMM is defined as:

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

2.2 Preparing Face Images For HMM Analysis

HMMs provide a way of modeling the statistical properties of one-dimensional (1D) signals. The 1D nature of certain signals, for example speech [10], is suited to analysis by HMM: the probability of an observation sequence $\mathbf{O}$ given a model λ is computed using the forward-backward algorithm, which reduces the order of computation from $2TN^T$ to N^2T [11]. Images, however, are two-dimensional (2D). No direct equivalent of 1D HMMs exists for 2D signals, where a fully connected 2D HMM would lead to an NP-complete problem [9]. Attempts have been made to use multi-model representations that give pseudo 2D structures [8]. In this paper, traditional 1D HMMs will be used. This poses the problem of extracting a meaningful 1D sequence from a 2D image. One solution is to consider either a temporal or a spatial sequence: these two methods were discussed in [12], where it was concluded that spatial sequences of windowed data give more meaningful models. Initial results [14] showed that a left-to-right HMM with top-to-bottom line block sampling gives physiologically significant features. The following section details how these models can be built.

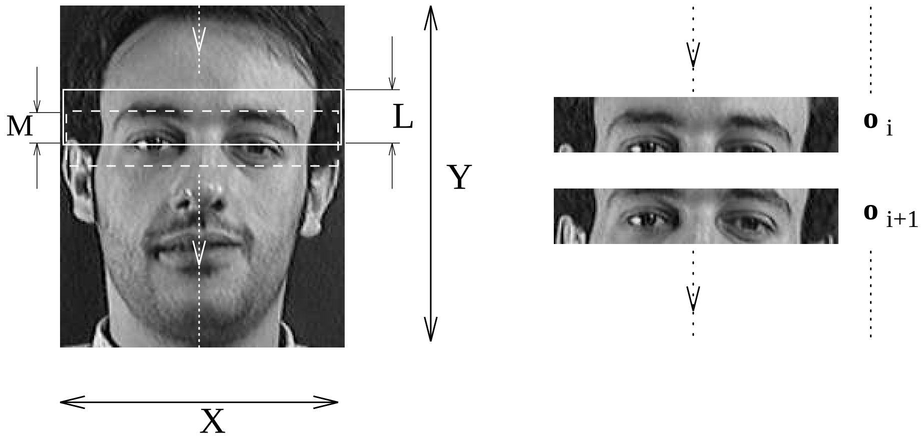

Figure 1: Sampling Technique

2.3 Training The Models

A training set of different face images is collected for each subject. The sampling technique described in figure 1 converts each image into a sequence of $X \times L$ column vectors $\mathbf{O} = (\mathbf{o}_1, \mathbf{o}_2, \ldots, \mathbf{o}_T)$, which are spatially ordered in a top-to-bottom sense[1]. Each vector $\mathbf{o}_i$ represents the intensity level values of all the pixels contained in the corresponding window. Using this technique each image is sampled into a sequence of overlapping line blocks. The overlapping allows the

[1] The number of observation vectors can be calculated as $T = \frac{Y-L}{L-M} + 1$

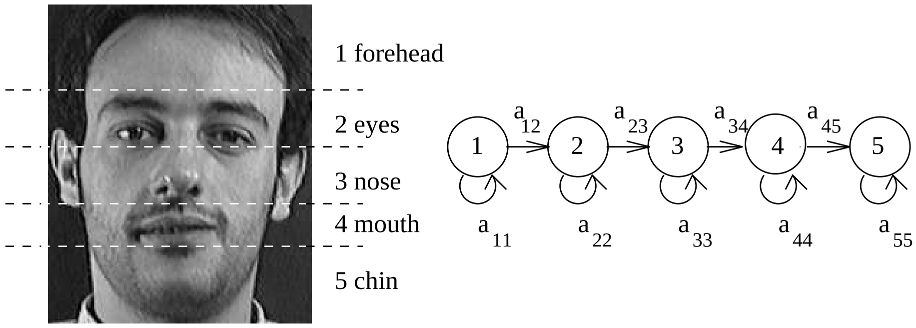

Figure 2: Facial Regions For 5-state left-to-right HMM

model to capture significant features independently of their vertical position (a disjoint partitioning of the image could result in features occurring across block boundaries being truncated). Assuming that each face is in an upright, frontal position, feature regions would occur in a predictable order, for example eyes following forehead, nose following eyes, and so on. This ordering suggests the use of a left-to-right HMM (see figure 2 for a 5-state model), where only transitions between adjacent states in a left-to-right manner will be allowed. The HMM will then segment the image into statistically similar regions, each of which will be represented by a facial band [13]. Each facial band corresponds to one state in the model and is represented by the mean of that state. One model is trained for each subject in the database, yielding a left-to-right HMM:

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

where F is the total number of different subjects in the database. The parameters of the trained HMM have meanings as follows:

$\mathbf{A}^{(k)}$: 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).

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

$\pi^{(k)}$: all observations sequences start in the top band and therefore, with the present model, this parameter does not provide any useful discriminating information ($\pi_1^{(k)} = 1$ and $\pi_i^{(k)} = 0, \ 1 < i \leq N, \ 1 \leq k \leq F$).

3 EXPERIMENTAL RESULTS

To test some of the ideas presented in the previous sections, a simple spatial left-to-right HMM was built. Five images of 20 different subjects were collected and a separate HMM trained for each subject. The 20 resulting HMMs were used to

Figure 3: Training Data

classify unknown images. Each image was of size 184×224 (8-bit grey levels) and was sampled into a 1D observation sequence using a line block of size 16, moving down 4 lines at a time (*i.e.* with a 12-line overlap). For such model the parameters of figure 1 have the following values: $X = 184, Y = 224, L = 16, M = 12$. For each subject, five such sequences (one per training image) were generated and used to train a 5-state left-to-right HMM. The number of states of the HMM was set to five because, by inspection, approximately five distinct regions seem to appear in the face image (see figure 2).

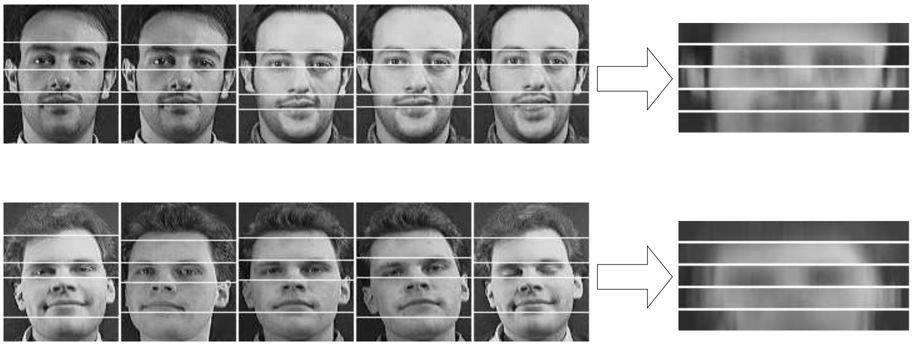

Figure 4: Segmented Training Data and Magnified State Mean Vectors

3.1 Training, Feature Extraction And Segmentation

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 [20]. After creating a prototype model (specifying number of states, observation vector size and allowed transitions), a set of initial parameters was iteratively computed from the training data, using at first uniform segmentation and successively Viterbi alignment. Parameters were then re-estimated using Baum-Welch re-estimation [11] and adjusted so as to maximise the probability of observing the training data, given each corresponding model.

Figure 3 shows the training data for two subjects in the database. An HMM was trained for each of them and figure 4 shows how the data was segmented, together with the mean vectors for the 5 states found by the HMM.

The training data is segmented as predicted in figure 2. The training images for the same subject differ from each other, nonetheless the segmented regions are found to be consistent across the training ensemble. The states for the two subjects correspond to physical facial features. For example, the eye band can be identified in both cases. There is evidence that one of the salient features for identification used by humans is the eye region of the face [15]. In some of the remaining states extracted by the HMM, a leaking effect is visible where some states appear lightly in the next one. Because of the overlap, the first observation vector to fall into a new state contains M lines in common with the previous state. The contribution to the state mean of the common lines will cause the leaking effect.

Figure 5 shows the segmented training data for 4 other subjects. The segmented regions are again consistent with one another, even when facial details differ greatly (as in the case of the subject with and without glasses).

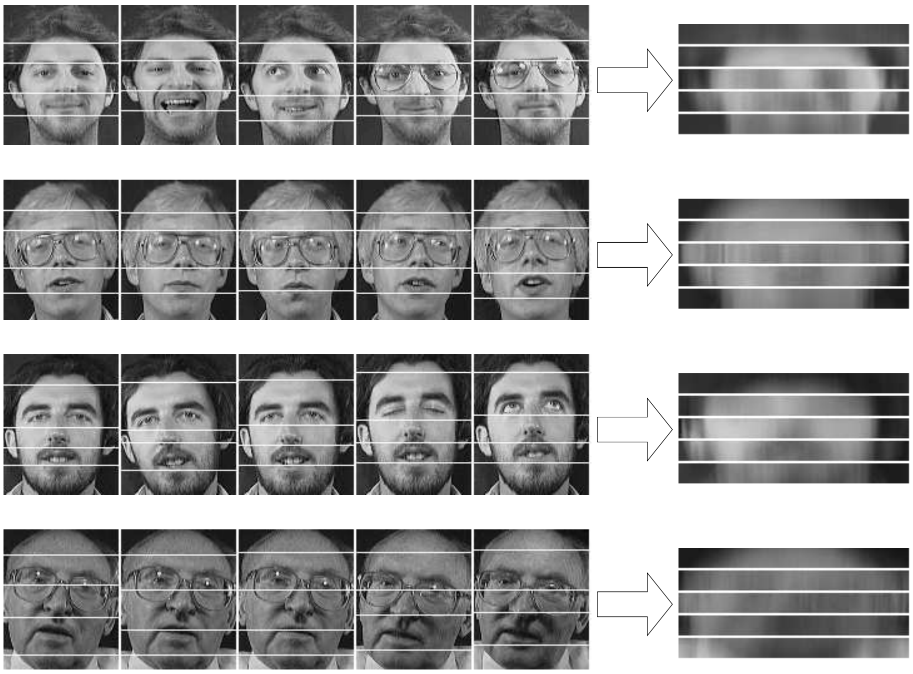

Figure 5: Segmented Training Data For Other Subjects

3.2 Using The Models For Identification

The full database comprised of 20 distinct subjects. In order to identify a test image, each was converted into a 1D observation sequence **O** using the sampling technique previously illustrated. The observation sequence was matched against each of the 20 models in the database and the model likelihoods were computed:

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

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

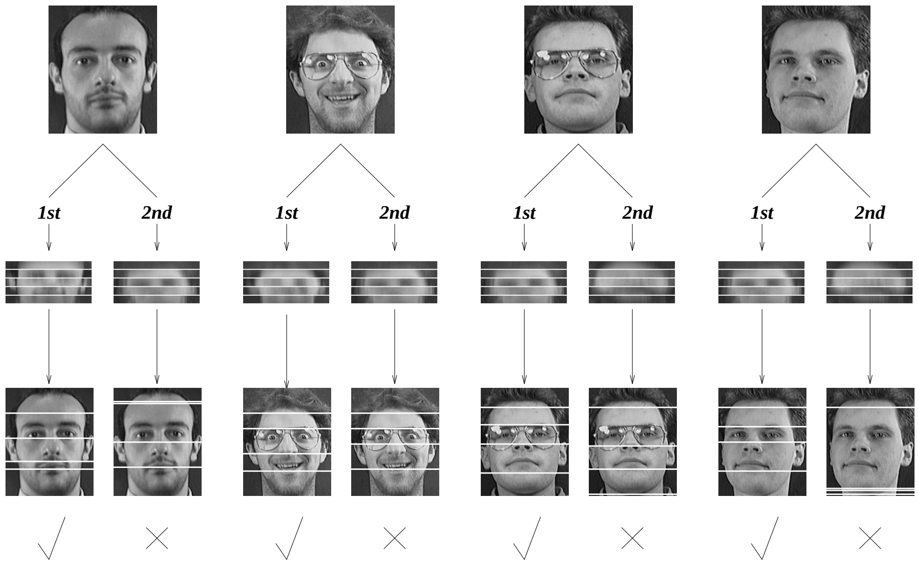

Figure 6: Test Results For Unseen Images

The 100 training images (5 pictures of 20 different subjects) were tested and correctly identified. The average log-likelihood per frame[2] for the correct matches varied between $-12,600$ and $-13,700$. Other experiments were carried out on images that had not been seen during the training phase. Figure 6 shows four such test images and the segmentation obtained with the best and second best match. The average log-likelihood per frame for the best match (which was correct for all 4 cases) varied between $-13,500$ and $-15,000$ [13]. Other experiments showed that various images not containing a face had scores below $-18,500$ [13].

The hand-drawn head depicted in figure 7 had a best match score of approximately $-15,500$. The segmentation results reflect quite accurately the model states. The eye band is located correctly and the score indicates that this method could be used to test the presence of a face-shaped object in an image.

Some of the test images in figure 6 differ considerably from the training images and the successful results obtained indicate that the model exhibits a certain robustness to feature variation. The segmentation obtained for these images is also consistent with that of the training data, even when facial details differ substantially. These results suggest that similar lighting conditions are not an essential

[2]The total number of frames was $T = 53$

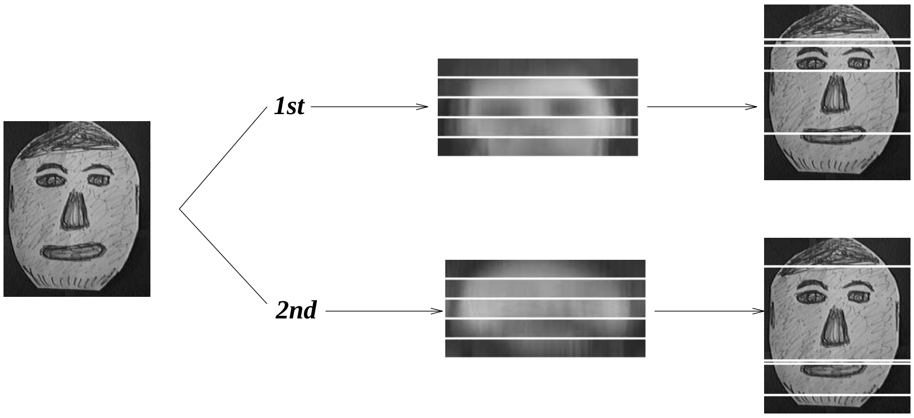

Figure 7: Segmentation Of Hand-drawn Face

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 [19] 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 face segmentation 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 the model. Other methods for feature extraction, such as Artificial Neural Networks [6], deformable templates [21] or snakes [7], 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 (with/without glasses, smiling/non-smiling, open/closed eyes) are dealt with automatically. Template-based models have generally suffered from such variations. It was also shown that robust face segmentation was obtained even for images substantially different from the training images. Experiments have started to compare this approach with other purely statistical methods: initial results based on a system as described in [17] with only one training image per subject indicate that a better identification performance is achieved with the HMM based approach. The classifier always outputs the best matching model from the database. The initial results investigated in this paper seem to indicate that the log-likelihood values are sufficiently discriminate to enable the presence of a face in the image to be detected.

Further vertical segmentation of each horizontal band will be investigated. This approach may yield more accurate feature location: for example, segmenting the eye band into 5 further vertical regions might locate the eyes. Initial experiments with very limited training data have already shown encouraging results. Model likelihood can be obtained using a pseudo 2D structure as described in [8]. This kind of representation makes more use of 2D information and will be compared with the purely 1D approach presented in this paper.

Future work will investigate 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 [4] applied to facial band images may yield more suitable state mean vectors for segmentation and classification.

At present the image analysis is carried out in the space domain. There is evidence [18] that frequency or frequency/space representations may yield better data separability and approaches will be investigated to process the face bands in the frequency domain.

Acknowledgements

This work is supported by a Trinity College Internal Graduate Studentship and an Olivetti Research Ltd CASE award. Their support is gratefully acknowledged. Thanks also to many colleagues in Cambridge: Andy Hopper, Frazer Bennett and Andy Harter 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 Department of Pure Mathematics and Mathematical Statistics, Barney Pell of the Computer Lab, and Steve Hodges and Tat-Jen Cham of CUED for the many ideas discussed together. Lastly, I wish to remember the late Prof. Frank Fallside, whose ingenuity illuminated my work on so many occasions.

References

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

[2] R. Diamond and S. Carey. Why faces are and are not special: an effect of expertise. *Journal of Experimantal Psychology: General*, 115,2:107–117, 1986.

[3] 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.

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

[5] A. Hopper. Digital video on computer workstations. *Proceedings of Eurographics'92*, 1992.

[6] 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.

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

[8] S. Kuo and O.E. Agazzi. Machine vision for keyword spotting using pseudo 2d hidden markov models. *Proceedings of ICASSP'93*, V:81–84, 1993.

[9] E. Levin and R. Pieraccini. Dynamic planar warping for optical character recognition. *Proceedings of ICASSP'92*, III:149–152, 1992.

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

[11] 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.

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

[13] F. Samaria and F. Fallside. Automated face identification using hidden markov models. In *Proceedings of the International Conference on Advanced Mechatronics.* The Japan Society of mechanical Engineers, 1993.

[14] 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.

[15] 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.

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

[17] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3,1:71–86, 1991.

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

[19] 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.

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

[21] 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.