Recognition of Head Gestures Using Hidden Markov Models

Carlos Morimoto Yaser Yacoob Larry Davis
Computer Vision Laboratory, Center for Automation Research
University of Maryland, College Park, MD 20742
yaser@cfar.umd.edu

Abstract

This paper explores the use of Hidden Markov Models (HMMs) for the recognition of head gestures. A gesture corresponds to a particular pattern of head movement. The facial plane is tracked using a parameterized model and the temporal sequence of three image rotation parameters are used to describe four gestures. A dynamic vector quantization scheme was implemented to transform the parameters into suitable input data for the HMMs. Each model was trained by the iterative Baum-Welch procedure using 28 sequences taken from 5 persons. Experimental results from a different data set (33 new sequences from 6 other persons) demonstrate the effectiveness of this approach.

1. Introduction

The analysis of human gestures, postures and expressions can facilitate better human-computer interaction. Unfortunately, isolating the roles of individual body-parts and their actions turns out to be a difficult task [5]. In this paper we focus on the recognition of a subset of head gestures.

Automatic face segmentation, tracking and recognition of face expressions have recently been reported [3, 4, 10]. We present in this paper an approach using Hidden Markov models (HMMs) to recognize head gestures. Hidden Markov models are an extension of the theory of Markov chains where the observation of a certain output is a probabilistic function of the state. A good introduction of this theory with several application examples for speech recognition is presented by Rabiner [8]. HMMs were recently used by [9] in a real-time system for recognition of a subset of American Sign Language hand gestures.

A head gesture corresponds to a particular pattern of head movement. We use the system developed by Black and Yacoob [3] to track the face and estimate its image motion parameters. The face is treated as a rigid body moving in 3D space, and its motion is characterized using 8 parameters. The parameters are coded by a vector quantization scheme to obtain a single data stream suitable for the HMMs. An approach for dynamic vector quantization to segment a sequence of gestures and to set thresholds for detecting the three rotations is applied.

Each gesture is modeled by an HMM. The probability distributions of each model are obtained by training (iterative refinement of the probabilities) in order to maximize the response of the model to the corresponding gesture. During testing, the input data is presented to all trained HMMs and the one with maximal response to the input pattern determines the output of the system.

This paper is organized as follows. Section 2 briefly introduces the theory of HMMs, and Section 3 describes the implementation of the head gesture recognition system using HMMs. Preliminary experimental results are presented in Section 4 and Section 5 concludes this paper.

2. Hidden Markov Models

Hidden Markov Models are an extension of the theory of discrete Markov chains. Rabiner [8] presents a tutorial on HMMs with applications to speech recognition. The special case of a discrete, first order, Markov chain is completely described by a set of states, a set of state transition probabilities and the specification of the initial state. This kind of stochastic process could be called an *observable* Markov model since the output of the process is the set of states at each instant of time, where each state corresponds to a physical (observable)

[0] The support of the Defense Advanced Research Projects Agency (ARPA Order No. #C635), the Office of Naval Research (contract N000149510521), and of the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) is gratefully acknowledged

event.

This concept of Markov model can be extended to include the case where the observation is a probabilistic function of the state, so that the state is not directly observable, i.e., *hidden*. Formally, an HMM is defined by the following [8]:

1. A set of states $S = \{s_1, s_2, ..., s_N\}$ of the model, where N is the number of states of the model. We will denote the state at time t as q_t.

2. A set of distinct observation symbols $V = \{v_1, v_2, ..., v_M\}$, where M is the number of observation symbols of the model.

3. A set of state transition probabilities $A = \{a_{ij}\}$, where $a_{ij} = P[q_{t+1} = s_j | q_t = s_i], 1 \leq i, j \leq N$;

4. A set of observation symbol probability distribution on state j, $B = \{b_j(k)\}$, where $b_j(k) = P[v_k | q_t = s_j]$, $1 \leq j \leq N$ and $1 \leq k \leq M$;

5. The initial state distribution $\pi = \pi_i$ where $\pi_i = P[q_1 = S_i], 1 \leq i \leq N$.

Observation sequences $O = O_1 O_2 ... O_T$ (where each observation O_i is one of the symbols from V) can be directly obtained from the HMM specification. For convenience, we use the compact notation $\lambda = (A, B, \pi)$ to indicate the complete parameter set of the model.

In order to be useful in real-world applications, there are three basic problems that must be solved:

P1 -*The evaluation problem:* given a model λ and a sequence of observations O, how to compute the probability that the observed sequence was produced by the model ($P[O|\lambda]$). The solution of this problem is necessary to determine which model, among several, is the most likely to generate the sequence O. This problem can be efficiently solved by the forward-backward procedure presented in [8].

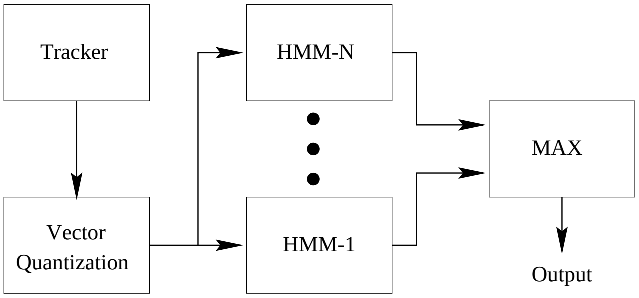
P2 -*The determination of the optimal observation sequence:* given O and λ, how to choose a state sequence $Q = Q_1 Q_2 ... Q_T$ which is "optimal" in some useful way (i.e., that best explains the observations). The choice of optimality criteria is a strong function of the intended use for the uncovered state sequence. Typically, it is used to learn about the structure of the model, to find optimal state sequences, or to get average statistics of individual states, etc. The most widely used criterion for solving this problem is to maximize $P[Q|O, \lambda]$ to determine the single best state sequence. The Viterbi algorithm [6], based on dynamic programming methods, is used to find such sequence.

P3 -*The training problem:* given λ and O, how to adjust the model parameters to maximize $P[O|\lambda]$. Although there is no known way to analytically solve this problem, a model λ can be chosen such that $P[O|\lambda]$ is locally maximized using the iterative Baum-Welch method [2].

The first task is to build individual gesture models by using the solution of the training problem P3 to optimally estimate their parameters. Refinements of the model can be made by using the solution of P2, that helps us understand the meaning of the states of the model. Finally, once the HMMs have been studied and optimized, recognition of a gesture is performed by using the solution to P1. A gesture sequence is presented to all HMMs and the one with maximum likehood determines the output of the system. The next section describes the implementation of our whole system.

3. System Implementation

The block diagram of our head gesture recognition system is presented in Figure 1. We use the approach developed by Black and Yacoob [3] to track the motion of the face. A brief description of the relevant part of this system is given in Section 3.1. The rigid-motion of the face is characterized by 8 parameters. The vector quantization module, described in Section 3.2, transforms these parameters into observation sequences that are used by the HMMs. The HMMs, already trained, use the solution of P1 to score the input observation sequence, and the model with the highest score (i.e., maximum likelihood) determines the output symbol. The description of the head gestures used for the experimentation of the system and their models are given in Section 3.3.



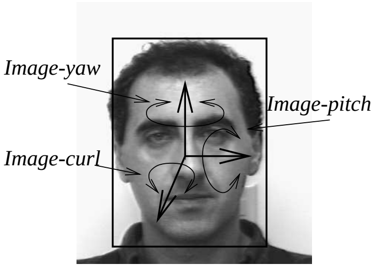
3.1. Tracking the Face Rigid Motion

The face tracking system developed by Black and Yacoob [3] is able to track face rigid and non-rigid motion based on some facial features. The face is treated as a rigid-body moving in 3D space and it is approximated by a plane viewed under perspective projection. This planar model has proven to be a reasonable approximation of the face image motion under a wide range of head movements. The image motion of a rigid planar region of the scene can be described by the following model [1]:

$$u(x,y) = a_0 + a_1 x + a_2 y + b_0 x^2 + b_1 xy$$
$$v(x,y) = a_3 + a_4 x + a_5 y + b_0 xy + b_1 y^2$$

where the a_i and b_j are parameters to be estimated and $u(x,y)$ and $v(x,y)$ are the horizontal and vertical components of the optical flow at the image point $\mathbf{p} = (x,y)$.

For recognizing head motions we are primarily interested in those motion parameters that express properties of the image motion related to 3D head rotations. In our experiments we concentrate on the b_0 and b_1 terms which roughly correspond to "yaw" and "pitch", respectively. We will call these rotations "image-yaw" and "image-pitch" to distinguish these from the related 3D head motions (see Figure 2). The image-curl in Figure 2 is a measure of rotation given by $\textit{image-curl} = a_4 - a_2$. Other measures of facial motion and the tracking of some facial features are described in [3].



3.2. Vector Quantization

Vector quantization [7] is used to code the parameters of interest into observation symbols that can be processed by the HMMs. The face tracking system provides streams for the eight motion parameters, from which the three image rotations are obtained. Figure 3 shows typical results of image rotation sequences. Each column corresponds to three gestures. The dashed lines of the graphs on top denote image-pitch and the dotted lines image-yaw. The graphs on the bottom correspond to image-curl. The first column presents, from left to right, the gestures YES, NO and MAYBE. The second column presents three YES gestures performed at different speeds. There are periods of silence between the presentation of gestures. These silent periods are used for gesture segmentation.

Each of these image rotations can be considered to be positive (+), negative (−), or zero (0), so that a simple thresholding scheme can be used to determine the instantaneous rotation status.

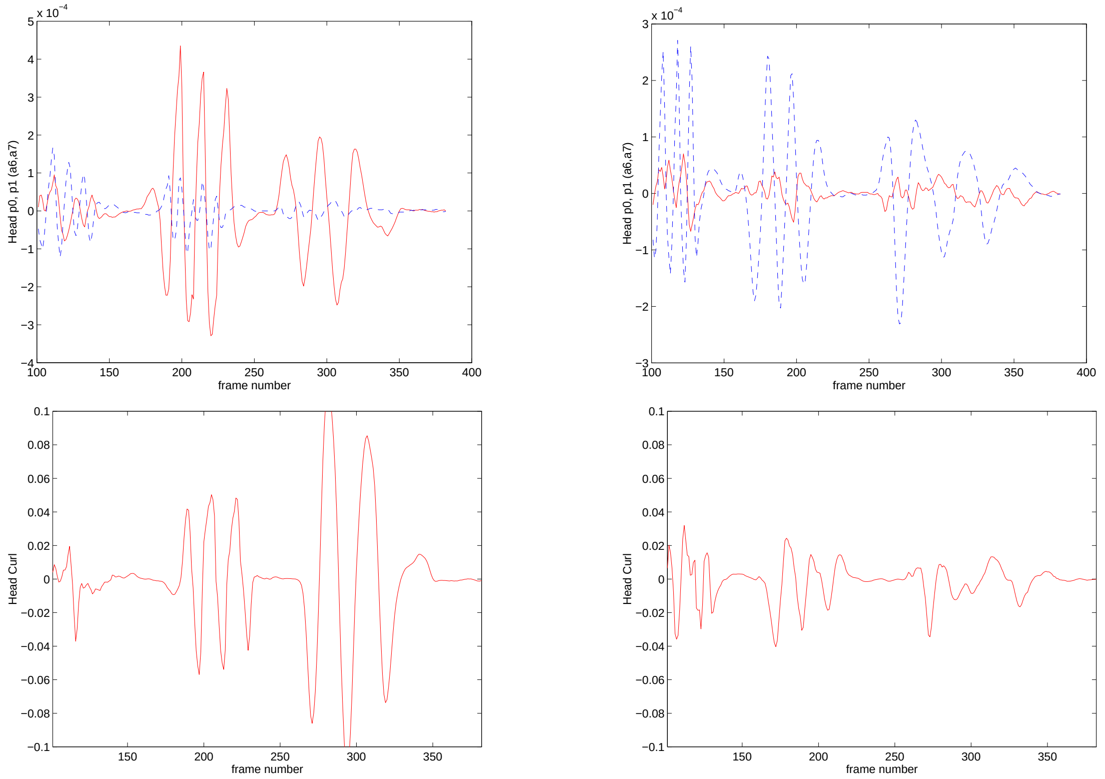
With 3 independent parameters which can assume 3 different values, 27 symbols are necessary to code a single state (e.g., the symbol "0" can be used to code the parameter state (+, +, +), the symbol "1" to code (+,+,-), and so on, until all combinations are coded).

To facilitate the training and recognition operation modes of the HMMs, the number of observation symbols were reduced by mapping the image rotation parameters into the seven symbols: UP, DOWN, LEFT, RIGHT, IN, OUT, and REST. A symbol is selected as output based on the dominant motion which is determined by the signal with highest energy. UP or DOWN are selected when image-pitch is dominant, LEFT or RIGHT when image-yaw is dominant, and IN or OUT when image-curl is dominant. When none of the signals is dominant, the symbol REST is selected.

To improve the performance of the HMMs, we use a dynamic mechanism for selecting an appropriate threshold before the quantization process. First, a low threshold is used to localize gestures in the input sequence. It is assumed that periods of silence (i.e., all image rotations close to zero) separate gestures. After the localization of a gesture in the sequence, the energy of each parameter is computed, and its energy level is used to select an appropriate threshold. The image rotations are then thresholded and mapped to the seven possible observation symbols, depending on the instantaneous highest energy band. Although the system could work by directly considering the overall highest energy band, we choose this approach of selecting the instantaneous highest band in order to model more complex gestures. Actually, if we restrict the system to these four gestures, considering the overall highest energy would increase the robustness of the system, but we would not be able to recognize gestures composed of different rotations.

3.3. Head Gestures

For our preliminary experiments we developed one HMM for each gesture. The task of this system is to map each gesture to the output symbols YES, NO, MAYBE and HELLO. The YES gesture is characterized by periodic cycles in the quantized values of image-pitch. The NO and MAYBE gestures are respectively characterized by cycles in image-yaw and image-curl.



A single nod (one single cycle in image-pitch) characterizes the HELLO gesture.

Most HMMs applied for speech recognition use left-right models. In left-right models, the state transition coefficients have the property $a_{ij} = 0$, for $j < i$, i.e., no transitions are allowed to states whose indices are lower than the current state. Due to the cyclic nature of many head gestures, we adopted ergodic or fully connected HMMs to model the cyclic gestures (YES, NO, and MAYBE gestures) and left-right models to model the sequential one (HELLO gesture). All HMMs have two states and have been trained by the iterative Baum-Welch procedure.

4. Experimental Results

The models were trained using 28 sequences taken from 5 persons. The test data consisted of 33 sequences taken from 6 other persons. Figure 4 shows the number of available gestures of each kind in both training and testing data sets. Due to the simplicity of the models, the parameters of each model were initially "guessed" and then tunned by the Baum-Welch iterative process.

Data	YES	NO	MAYBE	HELLO	Total
Train	18	3	4	3	28
Test	16	6	6	5	33

Figure 5 shows the results for the trained HMMs using test data. Observe that the HELLO gesture is confused most of the time with the YES gesture, because of the definition of the HELLO gesture. Looking at the input sequence data for the HELLO gesture, most subjects present more than a single nod (at least 1.5 cycles), which characterizes the YES gesture. The YES gesture is also confused with the HELLO gesture because some subjects when told to perform the YES gesture, executed only one image-pitch cycle. Other recognition errors occur when the data presents high energy on two or more rotations (ambiguous gesture).

	YES	NO	MAYBE	HELLO
YES	13		2	3
NO		5		
MAYBE	1	1	4	
HELLO	2			2
TOTAL	16	6	6	5

The table in Figure 6 demonstrate the improvement due to the use of a dynamic vector quantization scheme over a simple vector quantization scheme, which uses single thresholds in order to determine the observation symbols. Curiously, the performance of the HMMs improves for all gestures except for the NO gesture. Again, observing the data, one particular data sequence contained high energy in two rotations, image-yaw and image-curl, which caused the confusion in the output.

	YES	NO	MAYBE	HELLO
YES	5			3
NO	11	6	3	1
MAYBE			3	
HELLO				1
TOTAL	16	6	6	5

5. Conclusion

We have described an approach for recognizing human head gestures based on HMMs and a parameterized image motion model. The rigid motion of the face plane is tracked and 8 motion parameters that describe its motion are estimated. These 8 parameters are converted into 3 image-rotation parameters which are coded by a dynamic vector quantization scheme and sent to the HMM of each gesture. The gesture that correspond to the HMM with maximum likelihood is selected as output. Preliminary results with 4 gestures demonstrate the effectiveness of this approach. Further research is under way in order to improve the vector quantization module, and use more complex HMMs (with more states) to model the gestures. Also, we are gathering more data to train and test the system. Since the use of few data sets tend to overfit the models, we expect to obtain a better performance when more data is used to train the HMMs.

The simplicity of the gestures and models suggests the use of deterministic models instead of stochastic ones. They might be a good alternative, but we believe that the vocabulary of the system can be more easily extended when HMMs are used. Since the output of the HMMs are probability values, these could be used as confidence measures, and help disambiguating gestures, which would not easily be supported by pure deterministic systems.

Acknowledgements

We would like to thank the Laboratório de Sistemas Integráveis of the University of São Paulo (LSI - USP), for their kind support in the preparation of this paper.

References

[1] G. Adiv. Determining three-dimensional motion and structure from optical flow generated by several moving objects. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 7(4):384–401, July 1985.

[2] L. Baum, T. Petrie, S. G., and N. Weiss. A maximization technique occuring in the statistical analysis of probabilistic functions of markov chains. *Ann. Math. Stat.*, 41(1):164–171, 1970.

[3] M. Black and Y. Yacoob. Tracking and recognizing rigid and non-rigid facial motions using local parametric models of image motions. In *Proc. International Conference on Computer Vision*, Boston, MA, June 1995.

[4] I. Essa and A. Pentland. Facial expression recognition using a dynamic model and motion energy. In *Proc. International Conference on Computer Vision*, Boston, MA, June 1995.

[5] R. Feldman and B. Rime. *Fundamentals of Non Verbal Behavior*. Cambridge University Press, 1991.

[6] G. Forney. The viterbi algorithm. *Proceedings of the IEEE*, 61:268–278, March 1973.

[7] J. Makhoul, S. Roucos, and H. Gish. Vector quantization in speech coding. *Proceedings of the IEEE*, 73(11):1551–1588, November 1985.

[8] L. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–285, February 1989.

[9] T. Starner and A. Pentland. Visual recognition of american sign language using hidden markov models. In *International Workshop on Automatic Face and Gesture Recognition*, pages 189–194, 1995.

[10] Y. Yacoob and L. Davis. Computing spatio-temporal representations of human faces. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 70–75, 1994.