**M.I.T Media Laboratory Perceptual Computing Section Technical Report No. 388**
**Appears: Proceedings of the Second International Conference on Automatic Face and Gesture Recognition, October, 1996**

# Recovering the Temporal Structure of Natural Gesture

**Andrew D. Wilson** **Aaron F. Bobick**
**Vision and Modeling Group**

**Justine Cassell**
**Gesture and Narrative Language Group**

MIT Media Laboratory
20 Ames St., Cambridge, MA 02139
(drew, bobick, justine@media.mit.edu)

## Abstract

A method for the recovery of the temporal structure and phases in natural gesture is presented. The work is motivated by recent developments in the theory of natural gesture which have identified several key aspects of gesture important to communication. In particular, gesticulation during conversation can be coarsely characterized as periods of bi-phasic or tri-phasic gesture separated by a rest state. We first present an automatic procedure for hypothesizing plausible rest state configurations of a speaker; the method uses the repetition of subsequences to indicate potential rest states. Second, we develop a state-based parsing algorithm used to both select among candidate rest states and to parse an incoming video stream into bi-phasic and multi-phasic gestures. We present results from examples of story-telling speakers.

## 1 Introduction

The traditional paradigm for hand gesture recognition involves the construction of a model for each gesture to be recognized. This usually proceeds by collecting a number of examples of the gesture, computing the "mean gesture" and quantifying the variance seen in the examples. The hope is that this description will generalize to the actual test data. Examples of this approach include [9, 1, 13, 4, 10].

This typical pattern recognition approach may be well suited to the recognition of stylized or literal gesture, such as the gestures made by a user navigating aeronautical data in a virtual reality system by contorting their hands. These actions are less gestures than particular literal movements. Others examples are the emblematic gestures substituting for simple linguistic constructs: the ubiquitous OK sign or "giving someone the finger." These situations lend themselves to the construction of sophisticated models capable of representing the variations between people; in the case of the VR-controller, one might even alter the gesture vocabulary to make the recognition more robust.

However, as an approach to *natural gesture* understanding, this methodology seems inappropriate. By "natural gesture" we mean the types of gestures spontaneously generated by a person telling a story, speaking in public, or holding a conversation. The reasons for this skepticism are clear. First, the particular configurations and motions observed in natural gesture are inherently speaker dependent, influenced by cultural, educational, and situational factors [5]. An approach employing fixed, physical descriptions of gesture might find no cross-speaker invariances.

Second, and more important, is that the literal representation of the gesture assumes that the spatial configuration is in fact the most significant aspect of the signal to be extracted. Given that we are observing a sequence, it is plausible that more abstract *temporal* properties are the important elements of a gesture.

In this paper we develop a method for the detection of the important temporal structure — the *gestural phases* — in natural gesture. We begin by briefly relating some recent developments in the theory of natural gesture which have identified several key *temporal* aspects of gesture important to communication. We next present an automatic procedure for hypothesizing plausible rest state configurations of a speaker; the method uses the repetition of subsequences to indicate potential rest states. Lastly, we develop a state-based parsing algorithm used to both select among candidate rest states and to parse an incoming video stream into bi-phasic and multi-phasic gestures. We present results from two extended examples of story-telling speakers.

## 2 Gesture in Communication

Recent research in the field of natural gesture generation and parsing has identified four basic types of gesture generated during discourse [6, 3]. Three of these are considered to have meaning in a dialog: *iconic*, where the motion or configuration of the hands physically match the object or situation of narration; *deictic*, a pointing gesture; *metaphoric*, where the motion or shape of the hands is somehow suggestive of the situation. The fourth gesture type, *beats*, is generated to show emphasis or to repair mis-spoken segments.

Characteristic of these gesture types are particular temporal signatures. For example, each is typically bracketed by the hands being in a "rest state." Beat gestures — the simplest — consist only of a small baton-like movement away from the rest state and then back again; these gestures may be termed "bi-phasic." The iconic, metaphoric, and deictic gestures are executed by first "transitioning" from the rest phase into gesture space (the space in front of the speaker), then executing a smaller movement (the "stroke"), remaining at that configuration for a short duration, and then transitioning back to the rest state. Thus, these gestures may be termed "tri-phasic." What it means for a movement of the hands to be a "natural gesture" is defined, at least in part, by these temporal characteristics. The bi-phasic and tri-phasic distinction is introduced in [2]. The distinction between beats and representational gestures (iconic and metaphoric) is also discussed in [11].

In this paper we employ the above descriptions to derive a parsing mechanism sensitive to the temporal structure of natural gesture. Our initial goal is to find possible instances of bi- and tri-phasic gestures in a video sequence of someone telling a story.

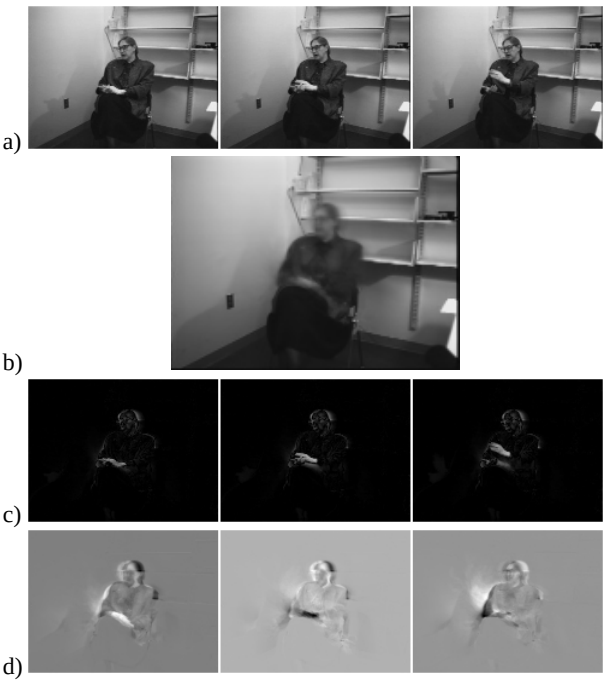


Figure 1: Three consecutive frames of the sequence used to illustrate this paper are shown in (a). (c) is the result of computing at each pixel the absolute value of the difference between the images in (a) and the mean image (b) computed from all frames of the sequence. (d) The first 3 eigenvectors of the image sequence.

The motivation is that the tri-phasic gestures encode meaning and need to be segmented from the input gesture stream if they are to be incorporated into any additional interpretation processes.

# 3 Detecting candidate rest states

## 3.1 Gesture data

The data presented in this paper are extracted from video of naive subjects relating a story. The subject was led into a closed room and asked to think of a time in which they believed they were in grave danger. The subject was then asked to tell a story of this event. The subject was instructed to look at the experimenter and not the camera, and was also warned that the experimenter would only provide minimal (nonverbal) feedback. Recording proceeded for as long as it took the subject to recount the story.

To reduce the size of recorded imagery, the video was digitized at low spatial ($120 \times 160$ pixels), temporal (10Hz), and photometric (8-bit gray scale) resolutions. The two sequences used to illustrate the results of this paper are 3min38sec and 4min10sec long, for a total of 4700 frames or 90MB of data. A few frames of the first sequence are shown in Figure 1.

## 3.2 Feature extraction

To analyze and compare subsequences we require a compact representation for the imagery. Because the focus of our analysis is on the temporal characteristics of the sequences we select the rather aggressive approach of representing each frame by a small number of coefficients derived from an eigenvector decomposition of the images [12].

We apply the technique to image sequences by randomly selecting a few hundred frames, computing the eigenvector decomposition of these frames, and then projecting all frames of the image sequence onto the resulting basis set. Next, the basis set vectors are ordered by how much variance each accounts for in the training frames. Because there is not tremendous variation in imagery of a person telling a story, and since it can be shown that two points that are nearby in the original image space are also nearby in the resulting low-dimensional space [7], we only need retain a small number of coefficients for this work. In the experiments reported here, we use only $n = 10$ coefficients to represent each frame; on average the 10 coefficients account for 55% of the variance. These coefficients are the entire representation used for all further processing.

## 3.3 Subsequence distance matrix

Let $\mathbf{x}_i$ be the $n$-vector of the eigenvector projection coefficients representing the $i$th frame of the image sequence. We define $d_{i,j}$ to be the difference between two frames $\mathbf{x}_i$ and $\mathbf{x}_j$ using a distance metric such as the Euclidean norm. Denoting the length $L$ subsequence beginning at frame $i$ and ending with frame $(i+L-1)$ as $\mathbf{x}_i^L$, we can define the difference between two subsequences $\mathbf{x}_i^L$ and $\mathbf{x}_j^L$ as the total Euclidean distance:

$$d_{i,j}^L = \left[ \sum_{k=0}^{L-1} d_{i+k,j+k}^2 \right]^{\frac{1}{2}}$$

By computing $d_{i,j}^L$ for all pairs of $i$, $j$ we can construct a matrix for all the subsequences.

Figure 2 presents the subsequence distance matrix for a central part of one of the test sequences. The diagonal is black, indicating perfect correlation. Black regions off the diagonal indicate points in time where a particular length $L$ subsequence is repeated. For example, beat gestures, which appear in the video as short repeated motions, show up as dark, short parallel lines. Subsequences that are fairly generic (e.g., hands are near the rest position) are likely to have several regions of high similarity. Conversely, motions or configurations that are highly unusual in the sense that they are unlike any other subsequences manifest themselves in the matrix as a row (and corresponding column) in which the mean distance is much greater than the overall mean.

The nature of the distance matrix is sensitive to the subsequence length $L$. If $L$ is small, we may see spurious similarities. If $L$ is too big, then the matching of "atomic" or primitive subsequences may be prevented. For the results reported here we have set $L = 5$ (one half second at 10Hz); we have not systematically experimented with varying $L$.

## 3.4 Selecting candidate rest states

Because rest states start and end each bi-phasic and tri-phasic gesture, and because rest states involve little or no motion for a reasonable duration of time, one expects a subsequence corresponding to a rest state to be repeated often. Furthermore, if one were to reconstruct a sequence using only a small collection of primitive subsequences, then one would want to use the rest state subsequence(s) as one of the primitives since it would well describe the imagery.

Our approach to finding candidate rest states is to use the reconstruction idea and to select a small set of subsequences which when repeated and assembled in the right order would reconstruct the original sequence as best possible. Of course, finding the optimal set of $k$ subsequences for reconstruction is an exponential

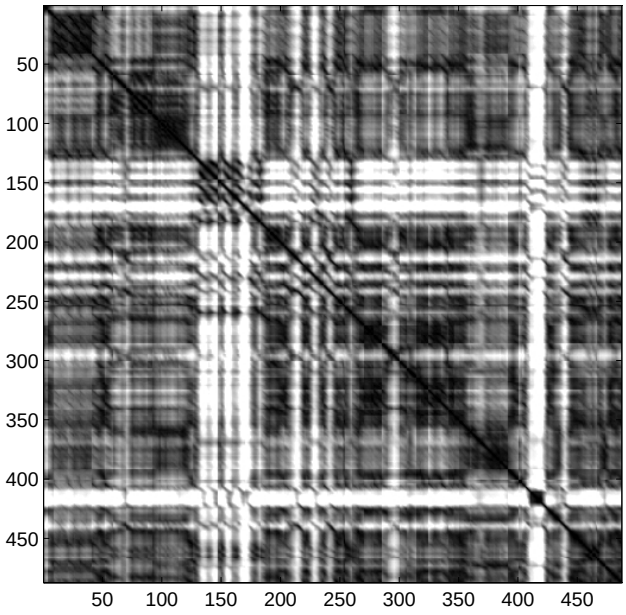


Figure 2: Distance matrix of subsequences of length 5, for a 300 frame section of the original video sequence. Dark parallel diagonal lines indicate repeated motions, possibly by "beat" gestures. An interesting sequence in which the subject repeatedly waves her arm at her side in a circular fashion begins at $i = 130$. The white bar around $i = 415$ indicates atypical movement; the subject is waving both arms high above her head.

problem since the best $k$ does not necessarily contain the best $k-1$ set. However, we expect the reconstruction to be dominated by a few rest states plus many unrelated motions. Therefore we use a "greedy" algorithm to select a set of reconstructing subsequences.

Let $\mathcal{M}$ be the set of all subsequences (call these models). Let $M \subseteq \mathcal{M}$ be a set of subsequences, where each $m \in M$ specifies the length $L$ subsequence beginning at $\mathbf{x}_m$ (frame $m$ in the original sequence). For each $\mathbf{x}_i$ define

$$y_i = \arg\min_{m \in M} d^L_{m,i}$$

That is, the sequence $y_i$ is the best reconstruction of the sequence $\mathbf{x}_i$ given the models $M$. The approximation error at frame $i$ is $e_i = \min_{m \in M} d^L_{m,i}$.

The "greedy" procedure is as follows: given the previously selected models $M$, pick the new subsequence model to add to $M$ such that the decrease in $\sum_i e_i$ is maximized. The algorithm is initialized by choosing the best single subsequence, $M = \{i\}$ where $i = \arg\min_j \sum_k d^L_{j,k}$.

The algorithm can be iterated as many times as there are frames; at that point $\sum_i e_i = 0$. However, each additional decrease in approximation error becomes quite small after a small number of models are included. For the 2200 frame sequence of Figure 1 we select only the first 40 subsequences; an additional 60 subsequence would be required to reduce the error only by one half.

Figure 3 illustrates the top six ranked (length 5) subsequences. Notice the variety of candidate rest states. The last example (6) is one which will later be rejected by our parsing mechanism: although the subsequence can be used to reconstruct a significant part of the original video, it does not have the right temporal properties to be considered a rest state. Figure 4 illustrates the top four candidates from a second example sequence. In this example, notice the radically different rest states recovered.


Figure 3: The top six ranked (length 5) subsequences for reconstruction. This selection illustrates the variety of candidate rest states. The last candidate (6) will be rejected by the temporal parsing.


Figure 4: The top four ranked subsequences for reconstruction for a second subject.

We note that we have begun to develop a "personality sensitive" video coding system based upon this technique. The resulting sequence is a reconstruction using the gesture primitives that well represent a person's style and thus feel less temporally aliased than that produced by standard systems that use a fixed low frame rate.

## 4 Detecting gesture phases

Given candidate rest states, we can now simultaneously evaluate them and parse the gesture stream into bi-phasic and tri-phasic gestures. The approach is to use a Markovian state description, but with the traditional use of transition probabilities replaced with an explicit model of duration.

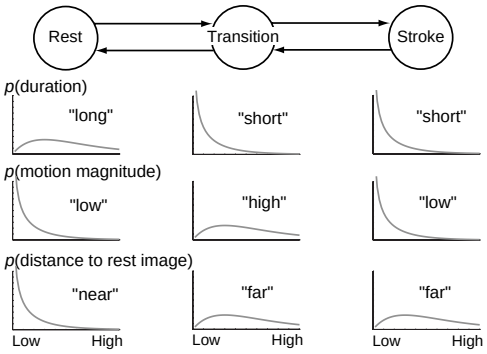


Figure 5: The three state machine describing the possible gestures. Below each state is a description of the gamma-density pdf for the given variables. The transitions are unlabeled because we do not use transition probabilities in generating a parse; rather, the duration models drive the temporal relationships.

## 4.1 Markovian states with duration modeling

Although Hidden Markov Models have been a popular technique for the recognition of gesture (see [14, 9, 10, 13]) we note that in our system *the states are not hidden.* In particular, our analysis of natural gesture types in section 2 identifies rest (R), transition (T), and stroke (S) states. The properties of these states are known and can be characterized by similarity in appearance to a rest state, amount of motion exhibited, and the duration during which the state is maintained. Probabilistic densities for these descriptions can be derived directly from training data.

Furthermore, the temporal structure of gesture can be described *a priori* using these states. Beat gestures correspond to moving from R to T and back to R: <R-T-R>; tri-phasic gestures traverse from R to T to S to T and back to R: <R-T-S-T-R>.[1] The *a priori* knowledge of the structure and properties of the states distinguishes our work from the typical HMM techniques.

Figure 5 graphically depicts a gesture phase finite state machine (FSM) and the associated properties of each state. While the exact form and values of the probability densities are not critical (each are modeled by gamma densities) it is important to understand their qualitative nature. The rest state R is modeled as tending to be "near" the rest state's position in eigen-space (using, say, the Euclidean norm), to have "low" motion as measured by the average traversal in eigen-space of the coefficients used to describe each frame, and of "long" duration. Likewise the T state is "far", "high", and "short" while the S state is "far", "low", and "short."

Given these descriptions, one might be tempted to just cluster and classify the image frames using appearance and velocity as features, and ignore any notion of transition. The difficulty with this is the idea of *duration*, which is well modeled using a Markovian system ([8]) where a modified Viterbi algorithm exists for parsing input streams with respect to duration models. Duration is fundamental to the idea of being a rest, transition, or stroke phase. The property of duration is much more critical to the gesture-parsing than is the probability of a transition occurring between any two states.

[1] We can also define multi-phasic gestures to be tri-phasic gesture which cycles through the T-S-T sequence more than once: <R-T-[S-T]$^+$-R>; this is sometimes seen when tri-phasic gestures are tightly repeated or overlap.

In traditional Markov systems, loopback transitions and their associated probabilities are manipulated in an attempt to alter the duration that a traversal remains in a given state. Formally, a fixed loopback transition probability is equivalent to an exponential density on duration, favoring shorter stays within a state. With such systems it is difficult if not impossible to disallow short durations.

To incorporate duration models and to use the Viterbi algorithm to generate the best possible parse, we adopt the framework of a Markov system, but with *no cost for a transition.* The result is a FSM where only the state-output probabilities and the duration the system remains in each state affect the parse. The effect is that instead of using transition probabilities to drive the temporal structure, we use the duration model. Proposing a traversal from state $i$ to state $j$ at time $t$ requires accepting the cost of ending the duration in the first state and starting that of the next.

## 4.2 Identifying rest states

The verification of rest states is accomplished by selecting a candidate subsequence, defining a gesture-phase-FSM using that candidate to define the rest state location in eigenspace, and then parsing the input data. If the tested subsequence is indeed a rest state, then the parsed input should spend a significant amount of time in the rest state R. If it is not, then most of the parse will oscillate between states T and S.

This verification process was applied to each of 40 candidate subsequences, ordered by the reconstruction method of section 3. Two points are of interest. First, many of the initial candidates (e.g. those ranked 6, 7, and 9) do not satisfy the rest state criteria when considered in a temporal context; their elimination validates the need for the temporal analysis beyond clustering.

Second, many candidate subsequences exhibit good rest state behavior, confirming the idea that there may be several rest states for a given speaker in a given situation. To select a set of rest states adequate to parse the gesture, we again construct a greedy algorithm; here we accumulate rest states according to how many new time-steps are now parsed as rest states if a new candidate is included. For the example of Figure 3 we use 20 rest states.[2] Manual thresholding selected this number. However, for the method of detecting the gesture states detailed in the next section, overestimating the number of rest states is much less of a problem than underestimating.

## 4.3 Results: detecting bi-phasic and multi-phasic gestures

To detect gesture phases, we need to construct a gesture phase FSM with the necessary rest states, and then parse the input sequence. To incorporate multiple rest states, we redefine distance to the rest state feature as the minimum distance to all of the chosen subsequences. To then detect the relevant gestures we simply parse the incoming video stream with respect to the gesture-phase-FSM; the parsing algorithm is a duration-modified Viterbi optimization [8].

Figure 6 illustrates the results for a 100 second long subsequence of one of the two video sequences tested; the other sections

[2] A few of the images of the different rest states are very similar in appearance. Under the eigenspace distance metric, however, they are quite disparate. This is because eigenspace coefficients are sensitive to global changes (e.g. a shift of the body) which should be abstracted for this domain. A more "hand-centered" or "body-centered" image description could substantially reduce the empirically determined number of rest states.

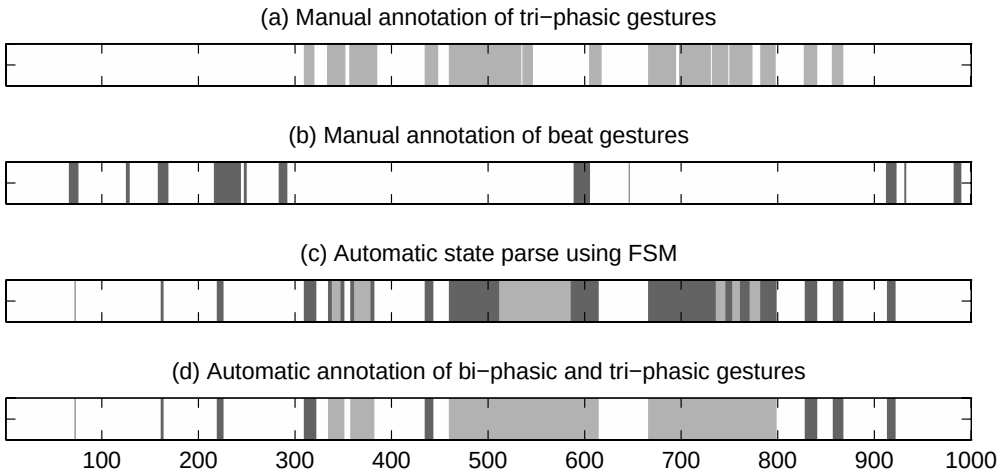


Figure 6: Example results of parsing the gesture video. (a) and (b) Visual encoding of a manual annotation of the presence of gesture. The annotation was produced by an expert in gesture communication who had not seen any of the results before viewing the video. (c) The state parse of our gesture-state-FSM and (d) the automatically derived labelling from the state parse (dark grey indicates bi-phasic beats, light grey tri-phasic gestures).

have similar results. The top two traces indicate the manually annotated labeling of tri-phasic and beat gestures. These labels were generated by the third author before seeing any of the results. The bottom trace depicts the state-based parse of the incoming video. Notice the overall similarity in the detection. The extent of agreement is difficult to measure quantitatively, and perhaps a bit premature as the gesture community still has difficulties agreeing as to what movements are gestures. Our contribution is the demonstration that the temporal structure coupled with an *a priori* state-based description is adequate to recover most of the gestures present.

We also note that we have tested this procedure on the 4min10sec sequence of a different speaker illustrated in Figure 4. Only one parameter of the model needed to be adjusted to generate similar results, and the parsing agrees with the authors' observations of the gesture. As mentioned, this sequence is interesting because of the radically different rest states recovered; a system must be sensitive to multiple rest states if it is to segment the gestures properly. However, we do not yet have independently generated manual annotations with which to compare descriptions.

## 5 Conclusion: Gesture and meaning

The gesture research community has identified fundamental types of natural gesture. In particular the tri-phasic gestures assist in conveying *meaning* in dialog. We have shown how the temporal structure of a video sequence of someone relating a story can be parsed into states that segment many of the tri-phasic gestures. We view this work as an initial step toward incorporating gesture sensitivity into dialog understanding.

We note that there is an immediate application of this technology to the summarization of video. Consider distilling a 3 minute sequence of someone telling a story to just a few frames or a few subsequences accompanying the text. The tri-phasic gestures contain meaning in the mind of the speaker, and are often used to disambiguate sections of narration where words alone do not easily express the idea (in general, gesture is thought to complement speech). The automatic extraction of these gestures should enhance the intelligibility of the summary.

## References

[1] A. F. Bobick and A. D. Wilson. A state-based technique for the summarization and recognition of gesture. *Proc. Int. Conf. Comp. Vis.*, 1995.

[2] J. Cassell. A framework for gesture generation and interpretation. In R. Cipolla and A. Pentland, editors, *Computer vision in human-machine interaction*. Cambridge University Press, in press.

[3] J. Cassell and D. McNeill. Gesture and the poetics of prose. *Poetics Today*, 12(3):375–404, 1991.

[4] Y. Cui and J. Weng. Learning-based hand sign recognition. In *Proc. of the Intl. Workshop on Automatic Face- and Gesture-Recognition*, Zurich, 1995.

[5] A. Kendon. How gestures can become like words. In F. Poyatos, editor, *Cross-cultural perspectives in nonverbal communication*, New York, 1988. C.J. Hogrefe.

[6] D. McNeill. *Hand and Mind: What Gestures Reveal About Thought*. Univ. of Chicago Press, Chicago, 1992.

[7] H. Murase and S. Nayar. Visual learning and recognition of 3-D objects from appearance. *Int. J. of Comp. Vis.*, 14:5–24, 1995.

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

[9] J. Schlenzig, E. Hunter, and R. Jain. Vision based hand gesture interpretation using recursive estimation. In *Proc. of the Twenty-Eighth Asilomar Conf. on Signals, Systems and Comp.*, October 1994.

[10] T. E. Starner and A. Pentland. Visual recognition of American Sign Language using hidden markov models. In *Proc. of the Intl. Workshop on Automatic Face- and Gesture-Recognition*, Zurich, 1995.

[11] K. Tuite. The production of gesture. *Semiotica*, 93-1/2:83–105, 1993.

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

[13] A. D. Wilson and A. F. Bobick. Learning visual behavior for gesture analysis. In *Proc. IEEE Int'l. Symp. on Comp. Vis.*, Coral Gables, Florida, November 1995.

[14] J. Yamato, J. Ohya, and K. Ishii. Recognizing human action in time-sequential images using hidden markov model. *Proc. Comp. Vis. and Pattern Rec.*, pages 379–385, 1992.