Learning Visual Behavior for Gesture Analysis

Andrew D. Wilson   Aaron F. Bobick
Perceptual Computing Group
MIT Media Laboratory
Cambridge, MA 02139
drew, bobick@media.mit.edu

Abstract

A state-based method for learning visual behavior from image sequences is presented. The technique is novel for its incorporation of multiple representations into the Hidden Markov Model framework. Independent representations of the instantaneous visual input at each state of the Markov model are estimated concurrently with the learning of the temporal characteristics. Measures of the degree to which each representation describes the input are combined to determine an input's overall membership to a state. We exploit two constraints allowing application of the technique to view-based gesture recognition: gestures are modal *in the space of possible human motion, and gestures are* viewpoint-dependent. *The recovery of the visual behavior of a number of simple gestures with a small number of low resolution image sequences is shown.*

1 From human motion to gesture

For all the degrees of freedom available to the human body, we seem to habitually use a only small class of motions that they permit. Even athletes, which as a group use their bodies in ways that most people do not, aspire to repeat motions flawlessly, spending hours upon hours practicing the same motion. In the space of motions allowed by the body's degrees of freedom, there is a subspace that most of us use. For example, if the body is modeled by a series of joints and angles between them, there would be many combinations of joint angles that we would never see.

Gesture is one interesting subspace of human motion. For the purposes of this paper, we define gesture to be motions of the body that are intended to communicate to another agent. Therefore, the gesturer and the recipient of the gesture must share a knowledge of gesture to communicate effectively. By way of simplification, this means the gesturer's movement must be one of a predefined set. We do not mean that a given gesture has a fixed geometry; a "wave" might be a single gesture in which the hand can be at any height between the chest and the top of the head. Rather, for now we are assuming that there is a set of gestures and each gesture defines a range of motions that are to be interpreted as being examples of that gesture. Thus gestures are *modal* in the space of possible human motion.

Due to the variability of human movement, the behavior of the gesture must be described without regards to precise geometry or precise temporal information. We take *visual behavior* to mean the sequence of visual events that makes a complete action or gesture. We assume that two gestures that have the same visual behavior are in fact the same gesture, thus ignoring the delicate problem of relating the form and meaning of natural gesture [6].

1.1 View-based approach

The machine understanding of human movement and gesture brings new possibilities to computer-human interaction. Such interest has inspired research into the recovery of the the complete 3-dimensional pose of the body or hand using a 3-dimensional physical model (e.g. [12, 13]). The presumption behind such work is that a complete kinematic model of the body will be required for useful inferences.

We claim that gestures are embedded within communication. As such, the gesturer typically orients the movements towards the recipient of the gesture. Visual gestures are therefore *viewpoint-dependent.* And the task of gesture understanding is particularly suited to a view-based, multiple model approach in which only a small subspace of human motions is represented.

In related work, Darrell and Pentland [3] use dynamic time warping and normalized correlation to match the interpolated responses of several learned image templates. Murase and Nayar [8] parameterize multiple eigenspaces over pose and illumination angle for object recognition. Polana and Nelson [10] match low level templates of spatiotemporal motion to recognize periodic human motions in image sequences. Cui and Weng [2] use learned decision boundaries to recognize sequences of vector-quantized images of hands. Starner and Pentland [15] extract the position and dominant orientation of both hands for the recognition of simple American Sign Language.

2 Representation of gesture

2.1 Multiple models for gesture

We claim that gestures are modal in the space of human motion. But how should a system model human motion to capture the constraints present in the gestures? There may be no single set of features that makes explicit the relationships that hold for a given gesture. In the case of hand gestures, for example, the spatial configuration of the hand may be important (as in a point gesture, when the observer must notice a particular pose of the hand), or alternatively, the gross motion of the hand may be important (as in a friendly wave across the quad). Quek [11] has observed that it is rare for both the pose and the position of the hand to simultaneously change in a meaningful way during a gesture.

Rather than use one model that is only partially effective, the approach here is to allow for multiple models. By *model* we mean a systematic way to describe a set of existing sensor data and a method to measure how well it describes new sensor data. Different models may interpret the same sensor data in different ways or they may take data from different sensors, in which case *sensor fusion* is the goal. The use of multiple models in a visual classification task is discussed in [9].

One goal is to develop an approach that can exploit multiple models simultaneously, where the type of models might be quite distinct. Model types useful for characterizing images in an image sequence might include eigenvector decomposition of sets of images [16], orientation histograms [4], peak temporal frequencies [10], tracked position of objects in the frame, and optic flow field summaries.

2.2 State-based descriptions

In previous work [1] we defined gesture to be a sequence of states in a configuration space. States were defined on some input space (say the joint angles returned by a DataGlove) and were designed to capture the constraints present in a series of training examples. Membership in a state was governed by probabilistic functions that attempted to capture the natural variability of motion in terms of the variances of these functions.

The temporal aspect of a gesture was incorporated by requiring that the states be defined along a prototype derived from training examples. Once defined, these states would be used to determine if a particular trajectory through the input space was an example of the learned gesture: the trajectory had to sequentially pass through each state attaining sufficient memberships in sequence. The actual time course was not important as long as the sequence was appropriate.

In the work presented here, we continue to consider a gesture as a sequence of states. At each point in time, the observed visual input reflects the current state and perhaps the transition to the next state. This state-based description is easily extended to accommodate multiple models for the representation of different gestures or even different phases of the same gesture. The basic idea is that the different models need to approximate the (small) subspace associated with a particular state. Membership in a state is determined by how well the state models can represent the current observation.

2.3 Learning visual behavior

In this paper we develop a technique for learning visual behaviors that 1) incorporates the notion of multiple models; 2) makes explicit the idea that a given phase of a gesture is constrained to be within some small subspace of possible human motions; and 3) represents time in a more probabilistic manner than defined by a prototype approach. In the remaining sections we first derive a state model and membership function based upon *residual*, or how well a given model can represent the current sensor input. We then embed this residual-based technique within a Hidden Markov Model framework; the HMM's represent the temporal aspect of the gestures in a probabilistic manner and provide an implicit form of dynamic time warping for the recognition of gesture. Finally, we demonstrate the technique on several examples of gesture and discuss possible recognition and coding applications.

3 Modeling gestures

3.1 Model instances and memberships

Suppose we have a set of observations $O = O_1, O_2, \ldots, O_T$ and a collection of states numbered $j = 1 \ldots N$. Assume that for each observation O_t we are given a degree of belief $\gamma_t(j)$ that O_t belongs to a state j; we require that $\sum_{j=1}^{N} \gamma_t(j) = 1$. We can interpret $\gamma_t(j)$ as the *membership* of O_t to state j. In this section we use the combination of multiple models to describe the set of observations belonging to a single state; in the next section we consider modeling the time course of states.

A number of *model types* A, B, etc. are selected *a priori* to describe the observations. One instance of each model type is associated with a state. Let us denote the set of model instances at state j as $\mathcal{M}_j = \{A_j, B_j, \ldots\}$. Each *model instance* is defined by the set of parameters that limit the model type to match some set of observations. For example, eigenvector decomposition of images may be a model type. An instance of the model type would comprise a particular mean image and set of eigenvectors (eigenimages) that are computed from a set of examples.

The parameters of $\mathcal{M}_j$ are computed from the set of example observations with the highest membership $\gamma_t(j)$. This may be accomplished by weighting each of the examples in the computation of the parameters, or simply by selecting some fixed number of the top observations, ranked by $\gamma_t(j)$. In the examples presented, the latter approach is taken.

For each model instance $m \in \mathcal{M}_j$ and an observation $\mathbf{x}$ we can compute a distance $d_m(\mathbf{x})$ which measures the degree to which the model instance m is unable to match $\mathbf{x}$. In this sense, $d_m(\mathbf{x})$ is a *reconstruction error* or *residual.* If we think of the parameters of $\mathcal{M}_j$ as limiting the model to a subspace of the samples, then we may also call $d_m(\mathbf{x})$ a *distance to model subspace.* The distances to each model instance may be combined to give

$$\mathbf{d}_j(\mathbf{x}) = \langle d_{A_j}(\mathbf{x}), d_{B_j}(\mathbf{x}), \ldots \rangle.$$

This quantity is similar to the "distance from feature space" derived in [7].

Next we consider the observation probability distribution $b_j(\mathbf{x})$ which describes the probability of measuring a particular residual for an observation when that observation is really generated by state j. The probability $b_j(\mathbf{x})$ may be estimated from the observations, each weighted by $\gamma_t(j)$. We estimate b_j as a normal[1] joint distribution on $\mathbf{d}_j$: $b_j(\mathbf{x}) = \mathcal{N}[\mathbf{d}_j(\mathbf{x}), \boldsymbol{\mu}_j, \Sigma_j]$, with

$$\boldsymbol{\mu}_j = \sum_{t=1}^{T} \frac{\gamma_t(j)}{\sum\limits_{t=1}^{T} \gamma_t(j)} \mathbf{d}_j(O_t),$$

and

$$\Sigma_j = \sum_{t=1}^{T} \frac{\gamma_t(j)}{\sum\limits_{t=1}^{T} \gamma_t(j)} (\mathbf{d}_j(O_t) - \boldsymbol{\mu}_j)(\mathbf{d}_j(O_t) - \boldsymbol{\mu}_j)^T.$$

The probability $b_j(\mathbf{x})$ may then be used to compute a new $\gamma_t(j)$. In the next section, Hidden Markov Models are presented as a technique for computing $\gamma_t(j)$ that treats the observations as a sequence rather than a set.

Having updated $\gamma_t(j)$, the estimation of the model instances $\mathcal{M}_j$ described above is iterated. In this way the memberships $\gamma_t(j)$ and the model instances are tuned to define the states in a way that best represents the observations. Summarizing, we list the following requirements of model instances:

- Model instance parameters are computed using the observations and their membership to the state, $\gamma_t(j)$.
- Each model instance delineates some subspace of the space of observations.
- The distances $\mathbf{d}_j(\mathbf{x})$ must be lowest for observations with a high degree of membership $\gamma_t(j)$.

3.2 HMM's with multiple independent model subspaces

Following a trend in speech recognition, vision researchers have applied the Hidden Markov Model technique to gesture recognition. Yamato et al. [17] compute a simple region-based statistic from each frame of image sequences of tennis swings. Sequences of the vector-quantized features are then identified by a trained HMM. Schlenzig et al. [14] use a rotation invariant representation of binary images and a neural net to quantize the image to a hand pose token before using an HMM. Starner and Pentland [15] apply continuous HMM's with the orientation and position of both hands wearing colored gloves.

HMM's are attractive because they put dynamic time warping on a probabilistic foundation and produce a Markov model of discrete states that codes the temporal structure of the observations [5]. Training an HMM involves inferring a first-order Markov model from a set of possibly continuous observation vectors. Each state is associated with the observation probability distribution $b_j(\mathbf{x})$. The probability of making a transition from a state i to a state j in one time step is denoted as $\mathbf{A}_{ij}$. The relationship between the discrete states and $b_j(\mathbf{x})$ is depicted in Figure 1(a).

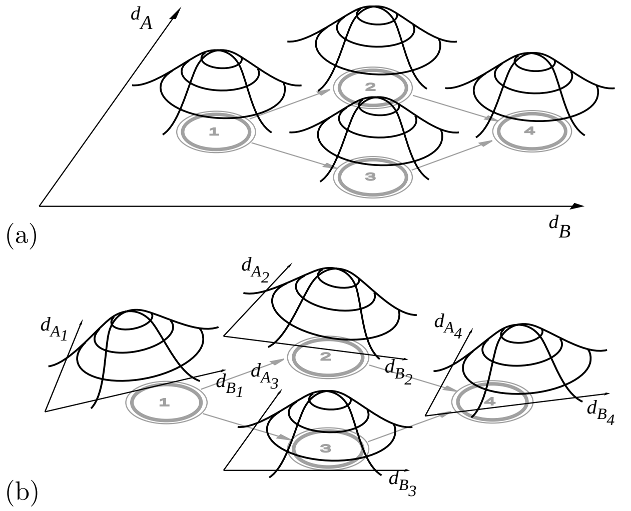

Figure 1: Each state of the Markov model (gray) is associated with a unimodal observation pdf. (a) In the conventional HMM framework all observation distributions reside in the same space of measurements from model A and B. (b) In the multiple independent model subspace HMM, each state is associated with an independent space of measurements from model A_j and B_j.

Treating O as a sequence, we may now interpret $\gamma_t(j)$ from Section 3.1 as the probability of being in state j at time t given the observation sequence O and the HMM. Computed by the standard "forward-backward" procedure, $\gamma_t(j)$ is used by the Baum-Welch algorithm to iteratively adjust $b_j(\mathbf{x})$ and $\mathbf{A}_{ij}$ until the probability of the HMM generating the observations is maximized.

Training the HMM with multiple independent model subspaces proceeds by interleaving iterations of the Baum-Welch algorithm (giving an updated $\gamma_t(j)$ to reflect the updated transition matrix $\mathbf{A}$) with reestimating the parameters of the model instances. In this way the representation encoded at each state is trained concurrently with the transition model. The relationship between each discrete state of the Markov model and the *multiple independent model subspaces* is depicted in Figure 1(b).

An advantage to this approach is that the representation at each state is designed to match the particular temporal model, while the temporal model is designed for the particular choice of representations as well. Additionally, by having multiple independent models, we do not rely on any one particular model instance to fit all observations for all states.

3.3 HMM topology

Before training the HMM, initial transition probabilities $\mathbf{A}_{ij}$ must be provided. The topology of the resulting Markov model is constrained by initially set-

[1] Since there are no negative distances, the gamma distribution may be more appropriate.

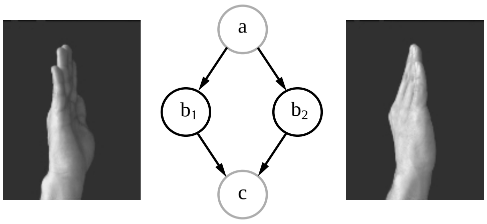

Figure 2: A state-based description of gesture must encode the relevant perceptual states. These images of an upright open hand share the same conceptual description, but have very different perceptual descriptions due to a slight change in viewing angle.

ting some $\mathbf{A}_{ij}$ to zero. To ease training, the topology is commonly constrained to be simple (e.g. causal).

The topology of the Markov model has the capability of encoding the temporal structure of the gesture. We choose not to restrict the topology of the Markov model initially and instead recover the topology through training (though the *number* of states is assumed). The reasons for doing so are twofold. First, by not providing a strict initial model we may recover interesting temporal structures that would otherwise escape notice, such as symmetry in time. In such cases the structure of the recovered transition matrix contributes to our understanding of the gesture.

Second, by providing a strict initial model we make implicit assumptions about the distribution of the sensor outputs (e.g., unimodal along the gesture in the case of a strictly linear Markov model). These assumptions may be unwarranted: while a simple gesture may seem to us a simple sequence of *conceptual states*, the sensors may see the movement as a complicated tangle of *perceptual states*. This may occur, for example, when the sensors used do not embody the same invariances as our own visual system. Figure 2 illustrates a single conceptual state (the upright hand) generating grossly different observations. If a single $b_j(\mathbf{x})$ cannot encode both observations equally well, then additional Markov states are required to span the single conceptual state. The addition of these states require the flexibility of the Markov model to deviate from strictly causal topologies.

3.4 Algorithm

To recover the temporal structure of the gesture and to train the representations at each state to suit the temporal model, we initialize the iterative algorithm sketched above with a uniform transition matrix and a random membership for each observation (Figure 3).

In the conventional HMM framework, the Baum-Welch algorithm is guaranteed to converge [5]. By interleaving the estimation of model parameters with the Baum-Welch algorithm, the proof of convergence may be invalidated. However, convergence has not been a problem with the examples tried thus far.

```
Initialization:
  set $\mathbf{A}_{ij} = \frac{1}{N}$ for all $i$, $j$
  initialize $\gamma_t(j)$ randomly, $\sum_{j=1}^{N} \gamma_t(j) = 1$

Algorithm:
  repeat until parameters of $\mathcal{M}_j$ do not change:
    for each state $j$ (parallelizable):
      estimate parameters to models $m \in \mathcal{M}_j$
      compute $\mathbf{d}_j(\mathbf{x})$ for all $\mathbf{x} \in O$
      estimate $b_j(\mathbf{x}) = \mathcal{N}[\mathbf{d}_j(\mathbf{x}), \boldsymbol{\mu}_j, \Sigma_j]$
    end
    update $A$, $\gamma_t(j)$ by Baum-Welch algorithm
  end
```

Figure 3: Training algorithm

4 Examples

4.1 Single model

For our first example, we use the eigenvector decomposition of the image as the single model type. In this case, the parameters associated with a model instance are simply a number of the top eigenimages that account for most of the variance of the training images (as indicated by the eigenvalues) and the mean training image. The training images for a model instance at state j are selected by ranking all samples by $\gamma_t(j)$ and selecting some number of the top samples. Given a model instance E_j and a sample image $\mathbf{x}$, $\mathbf{d}_j(\mathbf{x}) = \langle d_{E_j}(\mathbf{x}) \rangle$ is simply the reconstruction residual of the image using the precomputed eigenvectors at state j.

The input consists of 32 image sequences of a waving hand, each about 25 frames (60 by 80 pixels, grayscale) in length. The top 50 $\gamma_t(j)$-ranked sample images were used, and the number of eigenvectors was chosen to account for 70% of the variance of the selected sample images. Given the limited data, the Markov model was allowed 4 states.

The recovered Markov model, the mean image at each state, and a plot of $\gamma_t(j)$ and d_{E_j} for one training sequence are shown in Figure 4. The recovered Markov model permits the symmetry shown by the plot of $\gamma_t(j)$ over an observation sequence. Some other observation sequences differ in the extent of the wave motion; in these cases the state representing the hand at its lowest or highest position in the frame is not used.

The plot of $\gamma_t(j)$ reveals the time warping of the observation sequence to the Markov model. For example, the hand must decelerate to stop at the top of the wave, and then accelerate to continue. This is shown by the longer duration of membership to the first (top) state shown in the figure.

4.2 Position and configuration

The second example describes the position and configuration of a waving, pointing hand. In each frame of the training sequences, a 50 by 50 pixel image of the hand was tracked and clipped from a larger image with a cluttered background. Foreground segmentation was accomplished using the known background. The con-

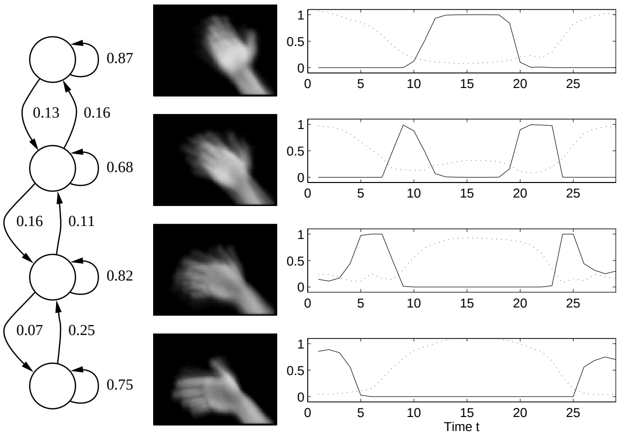

Figure 4: The recovered Markov model for all training sequences at left shows the symmetry of the gesture. The $\gamma_t(j)$-weighted mean image for each state is shown in the middle. On the right is a plot of $\gamma_t(j)$ (solid line) and d_{E_j} (dotted line) for each state for one training sequence. d_{E_j} was scaled to fit. The exact shape of the plots varies in response to the variance and length of the sequence.

figuration C of the hand is modeled by the eigenvector decomposition of the 50 by 50 images. The position P of the hand is modeled by the location of the tracked hand within the larger image; at each state, P is estimated as the $\gamma_t(j)$-weighted mean location.

The recovered Markov model is similar to that of the waving hand in the previous example. The mean images and $\mathbf{d}_j = \langle d_{C_j}, d_{P_j} \rangle$ for each state are shown in Figure 5.

The variance of each feature indicates the importance of the feature in describing the gesture. In this example both the position and configuration of the hand was relevant in describing the gesture. Had the location of the hand varied greatly in the training set, the high variance of the position representation would have indicated that position was not important in describing the gesture.

4.3 Two camera views

The final example shows how models for two camera views may be combined. Two views may be useful in describing the simultaneous movement of multiple objects (e.g. two hands, or a face and hands), or in describing gestures that exhibit movement in depth. Fifteen examples of a "pushing" gesture (see Figure 6(a)) were taken from both a side and front view. Eigenvector decomposition was used to model each view; $\mathbf{d}_j = \langle d_{E_j^{\text{front}}}, d_{E_j^{\text{side}}} \rangle$. The mean image for each view and plots of $\mathbf{d}_j$ are shown in Figure 6(b).

Note that in both views foreground segmentation is unnecessary since the camera is stationary. The mean image corresponds to the static parts of the image; the eigenvector decomposition subtracts the mean image at each state.

5 Conclusion and future work

The examples in Section 4 demonstrate that with a practical number of low resolution image sequences and weak initial conditions, the algorithm is able to recover the visual behavior of the gesture, including interesting temporal structure. Future implementations may be useful in novel wireless computer-human interfaces, real time "object-based" coding from video, and studying the relevant features of human gesture.

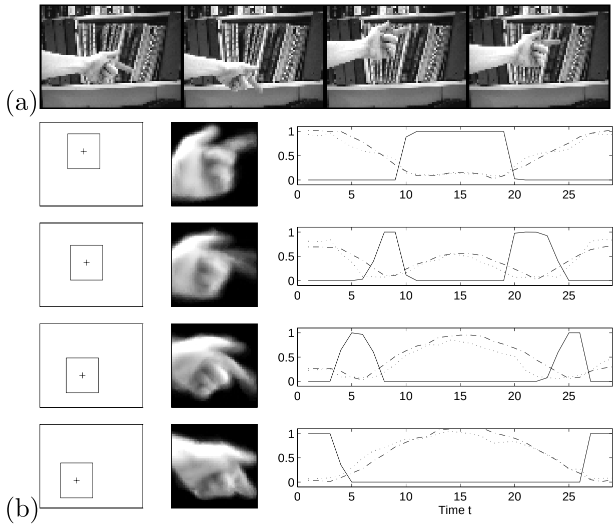

Figure 5: (a) Four representative frames (ordered left to right) are shown from one training sequence. (b) The $\gamma_t(j)$-weighted mean location of the tracked hand in the larger image is shown on the left. The mean image for each state is shown in the middle. On the right is a plot of $\gamma_t(j)$ (solid line), d_{C_j} (dotted line), and d_{P_j} (dash-dotted line) for each state for one training sequence. (d_{C_j} and d_{P_j} were scaled to fit.)

5.1 Real time coding

In HMM applications, recognition may be performed by the "forward-backward" algorithm, which returns the probability of an observation sequence given a particular HMM. The algorithm requires computing the observation probability at each state, at each time step. When there is a large number of states to consider or the cost of computing $b_j(\mathbf{x})$ is high, the forward-backward algorithm may not scale to a reasonable number of states, especially in a real time recognition or coding application.

In our framework the cost of computing the probability of a new observation is quite high, since d_m must be computed for all model instances $m \in \mathcal{M}_j$. If we were to use the eigenvector decomposition of the image as our model with 4 HMM's with 25 states each, keeping the top 10 eigenvectors at each state, then 1,000 image correlations would be required at every time step.

The development of a robust coding method of low complexity is an important step in a real time implementation of the framework. One possibility is to use a *beam search* in which $b_j(\mathbf{x})$ is computed for a small subset of the available states, chosen to maximize the probability of being in the states given the past observations. This strategy is suboptimal in the sense that it will miss paths through the model that are locally of

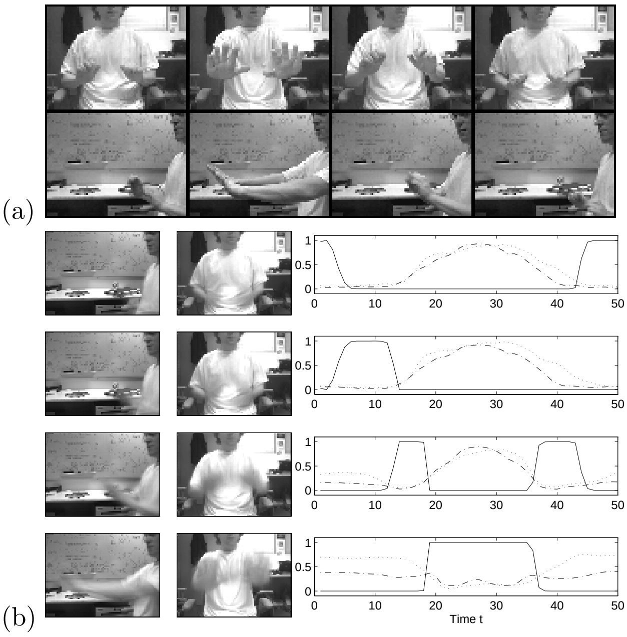

Figure 6: (a) Four representative frames (ordered left to right) are shown from each view of one training "push" gesture. (b) The mean images for both the side view and front view at each state are shown on the left. Plots of $\gamma_t(j)$, $d_{E_j^{\text{front}}}$ (dotted line) and $d_{E_j^{\text{side}}}$ (dash-dotted line) are from one training sequence.

low probability but globally of high probability. It remains to be seen whether this limitation has practical merit.

5.2 Model type selection

Because the observation distribution is a joint distribution on the distances to multiple independent model subspaces, the set of model types at each state does not have to be the same across all states. One topic of future work involves the automatic selection of model types at each state. For example, one state may be characterized by motion sufficiently, while in another both motion and shape are required to characterize the observations.

Automatic selection of models is desirable for a number of reasons. First, a particular model type may be unable to characterize the training examples at a state, in which case erroneous model instances should be removed from the set of models at the state. Secondly, if the appropriate set of features is not known beforehand, a good approach is to use all available models and let the system select the useful ones, possibly subject to a cost constraint. Lastly, the selection of different subsets of model types at each state allows the characterization of behaviors in which the set of relevant features changes over time. Automatic selection of models may provide insight into the degree to which this happens in the case of human gesture.

Acknowledgments

Special thanks to Tom Minka, Kris Popat, and Baback Moghaddam for valuable discussions on multiple models and eigenspace representations.

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

[3] T.J. Darrell and A.P. Pentland. Space-time gestures. *Proc. Comp. Vis. and Pattern Rec.*, pages 335–340, 1993.

[4] W. T. Freeman and M. Roth. Orientation histograms for hand gesture recognition. In *Proc. of the Intl. Workshop on Automatic Face- and Gesture-Recognition*, Zurich, 1995.

[5] X.D. Huang, Y. Ariki, and M. A. Jack. *Hidden Markov Models for Speech Recognition.* Edinburgh University Press, 1990.

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

[7] B. Moghaddam and A. Pentland. Probabalistic visual learning for object detection. In *Proc. Int. Conf. Comp. Vis.*, Cambridge, MA, 1995.

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

[9] R. W. Picard and T. P. Minka. Vision texture for annotation. *Journal of Multimedia Systems*, 3:3–14, 1995.

[10] R. Polana and R. Nelson. Low level recognition of human motion. In *Proc. of the Workshop on Motion of Non-Rigid and Articulated Objects*, pages 77–82, Austin, Texas, Nov. 1994.

[11] F. Quek. Hand gesture interface for human-machine interaction. In *Proc. of Virtual Reality Systems*, volume Fall, 1993.

[12] J. M. Rehg and T. Kanade. Visual tracking of high DOF articulated structures: an application to human hand tracking. *Proc. European Conf. Comp. Vis.*, 2:35–46, 1994.

[13] K. Rohr. Towards model-based recognition of human movements in image sequences. *Comp. Vis., Graph., and Img. Proc.*, 59(1):94–115, 1994.

[14] 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.

[15] 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.

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

[17] 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.