IEEE INTERNATIONAL CONFERENCE ON COMPUTER VISION, pp. 157-162, CORFU, GREECE, SEPTEMBER 1999

Recognition of Temporal Structures: Learning Prior and Propagating Observation Augmented Densities via Hidden Markov States

Shaogang Gong
Department of Computer Science
Queen Mary and Westfield College
London E1 4NS, UK
sgg@dcs.qmw.ac.uk

Michael Walter and Alexandra Psarrou
Harrow School of Computer Science
University of Westminster
Harrow HA1 3TP, UK
{zeoec,psarroa}@wmin.ac.uk

Abstract

An algorithm is described for modelling and recognising temporal structures of visual activities. The method is based on (1) learning prior probabilistic knowledge using Hidden Markov Models, (2) automatic temporal clustering of hidden Markov states based on Expectation Maximisation and (3) using observation augmented conditional density distributions to reduce the number of samples required for propagation and therefore improve recognition speed and robustness.

1 Introduction

The underlying spatial and in particular temporal structure is important for the modelling of a dynamic visual phenomenon arising from activities such as gestures and human actions. For the purpose of recognition, it is essential to model the temporal structures. Such structures can be extracted by representing the activities as "trajectories" in a high dimensional feature space with its dimension determined by the number of visual measurements. These observations are often highly correlated. For example, a gesture can be represented by the trajectory of an observation vector (x, y, dx, dy) given by the object-centred position and displacement of the body part which performs the gesture. In general, an observation vector can also include among other features the positions and displacements of a set of salient feature points describing the shape of the object of interest, object colour distribution, its 3D pose and configuration [10]. Provided that observations of activities such as gestures can be represented as probabilistic spatio-temporal trajectories in a feature space, gesture recognition can then be performed by simply matching the trajectories as static templates (spatio-temporal structures) in the feature space [3, 9]. However, by and large modelling temporal structures as static templates can be very sensitive to noise and ambiguities in observation trajectories. This is particularly true in gesture recognition because gestures are rather arbitrary in their forms and probabilistic by nature. No person let alone different subjects performs the same gesture in exactly the same way twice. In addition to the "holistic static shape" of a spatio-temporal trajectory, there are other factors contributing to the spatio-temporal structure of a gesture, including (1) covariance in observation measurements due to variations in object-centred spatial position and scale plus measurement noise, (2) nonlinear temporal scaling due to variations in the speed and duration of executing a gesture, and (3) ambiguities in temporal segmentation required for determining the start and end of a gesture.

To cope with such problems, one can introduce a set of finite "hidden temporal states" which aim to capture and explicitly model the salient "phases" of a temporal structure over time. Predicting state transitions then provides a more robust means to cope with time scaling and avoid the need for determining the starting and ending points. To this end, Hidden Markov Models (HMMs) are widely used as a probabilistic framework for modelling temporal structures. HMMs have clear Bayesian semantics and efficient algorithms. They can perform dynamic time warping for structures that have been stretched and squashed in time. HMMs have been successfully applied to speech recognition [11], visual focus of attention [12], learning object movement and behaviour models [5, 8] and more recently gesture recognition [13, 2]. In the case of gesture recognition, the states are usually selected so as to capture the locations along the observation trajectories where measurements undergo significant change. Prior knowledge in the forms of state transition probabilities and conditional observation covariances are estimated from training examples. However, it is generally the case that selecting a finite set of hidden Markov states over time can be both arbitrary and rather over-committing. This is particularly true in the case of gesture recognition. More recently, conditional density propagation (CONDENSATION) algorithm proposed by Isard and Blake [6, 7] sug-

gests a more flexible framework to HMMs. Instead of modelling observation probabilities conditional to a finite set of states, they are continuously propagated over time. For gesture recognition, CONDENSATION has been adopted by Black and Jepson [1]. The model performs fix-sized local linear template matching weighted by the conditional observation densities propagated according to CONDENSATION therefore allowing for a global nonlinear time scaling. Unfortunately, the model does not consider measurements covariance (treated independently) therefore is sensitive to noise. It also does not use any prior knowledge, apart from the accumulated history of the trajectory being recognised, on both the state distribution and the observations of a structure. Consequently, a very large number of density samples (over thousands) with localised uniform distribution are initialised and then propagated over time. State predictions are simply previous states plus arbitrary Gaussian noise. This can lead to local minima and is computationally expensive.

In this work, we introduce a framework for the recognition of temporal structures in state space and illustrate the method through gesture recognition. The model utilises HMMs (1) to learn prior knowledge on both state distributions and observation covariances of temporal structures (2) to perform automatic state selection and segmentation using temporal clustering on training examples and, (3) to continuously propagate observation densities via hidden Markov states under the constraint of the learned prior but also subject to augmentation by the current visual observation.

2 Learning prior on temporal structures

The temporal structures extracted from gestures are probabilistic and ambiguous in nature. Learning prior knowledge from examples plays an important role in modelling such structures and this can be performed using HMMs. A HMM can be seen as the quantisation of the observation feature trajectories into a small set of discrete hidden states. In fact most gestures are temporally ordered therefore only first-order HMMs are usually required. Here the observations are continuous therefore the conditional observation probabilities at each state is given by a density distribution. More precisely, a HMM is defined by a set of hidden states $q \in \{q_1, q_2, \ldots q_N\}$ where N is the number of states. A model can be fully described by probabilistic parameters $\lambda = (A, B, \pi)$ where $A \in \{a_{ij}\}$ are the state transition probabilities, $\pi \in \{\pi_1, \pi_2, \ldots \pi_N\}$ are the initial probabilities of being in state i at time $t = 1$, $B \in \{b_j(o)\}$ are the observation density distributions at all states. At each state j, $b_j(o)$ can be a Gaussian mixture $b(o) = \sum_{k=1}^{K} c_k \, \mathcal{G}(o, \mu_k, \Sigma_k)$ given by mixture coefficient c_k, mean μ_k and covariance Σ_k.

If the prior of a gesture is to be learned using a HMM,

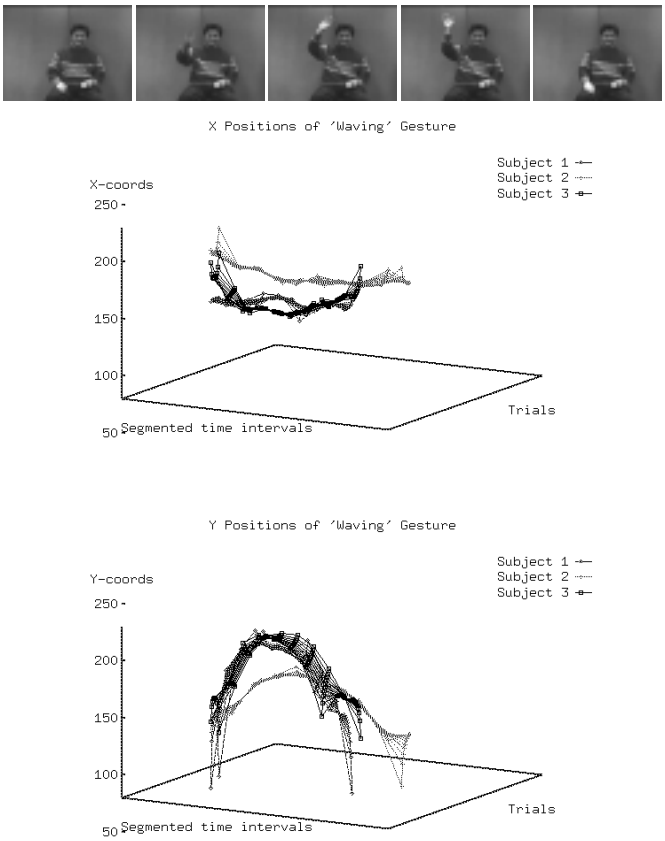

Figure 1. **An example of a waving gesture. The plots show the x-y motion centroid of a collection of waving gestures performed by 3 subjects each repeating 4 times.**

Figure 2. **Gestures of drawing figures.**

let us define the observation vector being $o = [x, y, dx, dy]$, the object-centred position and its displacement of either the mean body movement (Figure 1) or the movement of a particular part of the body such as one hand (Figure 2). Then a gesture state vector can be defined as a specific gesture being in a particular hidden state given by $s = [q_t, l_\lambda]$, where

l_λ and q_t are discrete values of model label and hidden state index at time t respectively. Learning prior then involves (1) automatic hidden state segmentation through temporal clustering, the estimation of (2) hidden state transition distributions and (3) conditional observation density distribution at each hidden state. For state segmentation, one aims to maximise the likelihood $P(O|\lambda)$ for a gesture $\lambda = (A, B, \pi)$, where O denotes an observation training set. This can be achieved using the EM algorithm [4].

Given the number of hidden Markov states to be N, learning the locations of the states (automatic temporal segmentation), their transition probabilities and the conditional observation density distributions of each state can be performed as follows: First, initialise $\pi = \{1, 0, \ldots 0\}$ and the state transition matrix A according to

1. For a first-order HMM, the average time in a state can be estimated as the ratio between the mean duration $\hat{T}$ of the training set and the number of states N, $\hat{t} = \frac{\hat{T}}{N}$.

2. Assign the state transition probabilities according to the average time a gesture remaining in a state

$$\hat{t} = \sum_{n=1}^{\infty} n\, a_{ii}^{n-1}\,(1 - a_{ii}) = \frac{1}{1 - a_{ii}} \qquad (1)$$

3. Initialise the state transition matrix as

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & 0 & 0 \\ 0 & a_{22} & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & a_{N-1N-1} & a_{N-1N} \\ 0 & 0 & \cdots & 0 & 1 \end{bmatrix}$$

where $a_{ii} = 1 - \frac{1}{t}$ and $a_{i\,i+1} = 1 - a_{ii}$.

Second, use the EM algorithm over a training set of M training examples $O = \{O^1, \ldots O^M\}$ to iteratively perform temporal clustering on the states and estimate model probability distributions A, B and π. Figure 3 illustrate the iterative process of automatic clustering the hidden states for a gesture drawing figure "5". In this example, the number of hidden states is set to 15. The mixture on conditional observation density distribution is set to 1.

3 Recognition of temporal structures

For modelling a temporal structure, let us define a state vector at time t as $s_t = [q_t, l_\lambda]$ given by the hidden Markov state q_t of a model l_λ. Notice that this HMM based state vector implicitly encodes the information on both phase ϕ and temporal scaling ρ used by Black and Jepson [1]. In general, at any time t, a temporal structure is fully described by its state history $S_t = \{s_1, s_2, \ldots s_t\}$, its current observation o_t and its observation history over time

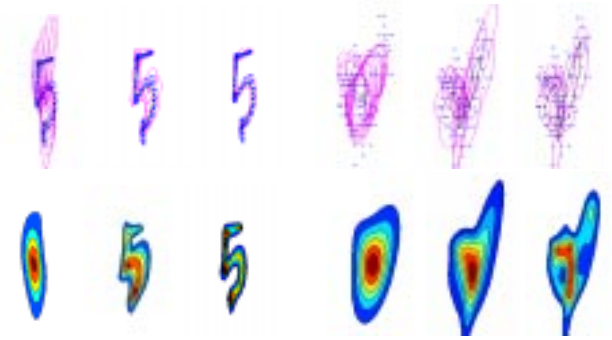

Figure 3. Learning the spatio-temporal structure of drawing figure "5" using HMM and EM clustering. The process of iteration is shown from left to right. The top row shows the clustering on object-centred positions (first 3) and displacements (last 3). The second row shows the corresponding density distributions over the entire structure based on the clustered hidden states and their distributions in space and time.

$O_t = \{o_1, o_2, \ldots, o_t\}$. However, if a temporal structure only forms a first-order Markov chain, a state at time t depends only on its previous state at time $t-1$ and is independent from its history $p(s_t|S_{t-1}) = p(s_t|s_{t-1})$. This is largely true for gestures. Based on the Bayes' rule, gesture recognition at time t can then be performed using the posterior density given observation history

$$p(s_t|O_t) = k_t\, p(o_t|s_t)\, p(s_t|O_{t-1}) \qquad (2)$$

where $p(s_t|O_{t-1})$ is the prior from the accumulated observation history up-to time $t-1$, $p(o_t|s_t)$ is the observation conditional density and k_t is a normalisation factor. The prior density from the observation history $p(s_t|O_{t-1})$ is estimated by multiple conditional density samples s^i through either uniform sampling at time zero or factored (prior weighted) sampling thereafter. The weighted sample set then approximates the prior through prediction

$$p(s_t|O_{t-1}) = \int_{s_{t-1}} p(s_t|s_{t-1})\, p(s_{t-1}|O_{t-1}) \qquad (3)$$

where $p(s_t|s_{t-1})$ is the state propagation density. This is essentially CONDENSATION [6]. However, based on the accumulated history of the current gesture sequence alone without any prior knowledge, the state propagation density $p(s_t|s_{t-1})$ can only be given as the previous estimation plus arbitrary Gaussian noise and consequently, meaningful estimation of the history accumulated prior $p(s_t|O_{t-1})$ can only be obtained through propagating a very large sample set of conditional densities (thousands) over time [1]. As a result, the prediction can be both expensive and sensitive to observation noise.

In order to reduce the required number of samples for the propagation and also to cope with noise and variance

in observation, prior on gesture temporal structures learned from training examples should be used. Let us assign the state propagation densities $p(s_t|s_{t-1})$ to the hidden Markov state transition probabilities of

$$p(s_t \mid s_{t-1}) \;=\; p(q_t = j \mid q_{t-1} = i, l_\lambda) \;=\; a_{ij} \qquad (4)$$

and the observation conditional density $p(o_t \mid s_t)$ as the Markov observation densities given by the prior on the measurement covariance and mean at each hidden state

$$p(o_t \mid s_t) \;=\; p(o_t \mid q_t = j, l_\lambda) \;=\; b_j(o_t) \qquad (5)$$

The observation covariance given by the density function at each hidden Markov state $b_j(o_t)$ enables the model to cope with measurement variance and noise. It also encodes scaling α used by [1].

4 Observation augmented densities

Recognition based on prior can be made more robust if current observation is also taken into account before prediction. Let us consider the state propagation density $p(s_t|s_{t-1})$ in Equation (3) to be augmented by the current observation, therefore $p(s_t|s_{t-1}, o_t)$. Assuming observations are independent over time and future observations have no effect on past states, the prediction process of Equation (3) can then be replaced by

$$\begin{aligned} p(s_t|O_t) \;&=\; \int_{s_{t-1}} p(s_t|s_{t-1}, o_t)\; p(s_{t-1}|O_{t-1}) \\ &=\; \int_{s_{t-1}} k_t p(o_t|s_t) p(s_t|s_{t-1}) p(s_{t-1}|o_{t-1}) \end{aligned} \qquad (6)$$

where $k_t = \frac{1}{p(o_t|s_{t-1})}$ and

$$\begin{aligned} p(s_t|s_{t-1}, o_t) \;&=\; \frac{p(s_t, o_t|s_{t-1})}{p(o_t|s_{t-1})} = \frac{p(s_t|s_{t-1}) p(o_t|s_t, s_{t-1})}{p(o_t|s_{t-1})} \\ &=\; \frac{p(s_t|s_{t-1}) p(o_t|s_t)}{p(o_t|s_{t-1})} \end{aligned} \qquad (7)$$

Given that the observation and state transitions are constrained by the underlying HMM, the state transition density is then given by

$$\begin{aligned} p(s_t|s_{t-1}, o_t) \;&=\; p(q_t = j|q_{t-1} = i, o_t) \\ &=\; \frac{a_{ij}^{l_\lambda} b_j^{l_\lambda}(o_t)}{\sum_{n=1}^{N} a_{in}^{l_\lambda} b_n^{l_\lambda}(o_t)} \end{aligned} \qquad (8)$$

The observation augmented prediction unifies innovation and prediction in CONDENSATION given by Equations (2) and (3). Without augmentation, CONDENSATION performs a "blind" prediction based on observation history alone. Augmented prediction takes the current observation into account and adapts the prior to perform a "guided" search in prediction. This improves the recognition rate and reduces the number of samples required for propagation.

5 Experiments

In order to illustrate our approach we have used two sets of gestures. Set A represents alphanumeric symbols similar to that defined in [1] and they are the numerals "2", "3", "5" and letter "l". The gestures of set B represent natural gestures that are defined within the context of an application in visually mediated interaction and they are (1) pointing left, (2) pointing right, (3) waving high up and (4) waving low down [9]. In set B the object-centred position and displacement (x, y, dx, dy) of a gesture in time t is determined using moment features estimated from image motion as described in [9]. In set A, in addition to image motion the skin colour of the hand is also used for extracting the observation vector (x, y, dx, dy). As a result the gestures in set A are less noisy than that of set B. A database of image sequences was collected and for the purpose of these experiments we build HMMs using four examples of each alphanumeric gesture from set A and six examples of each subject performing gestures from set B. Each sequence in sets A and B has on average 40 frames captured at 12 Hz.

On robustness: Tables 1 and 2 show the results in the form of confusion matrices. Using 40 density samples and 10% to initialise the recognition of gestures from set A, observation augmented propagation of density functions recognised all alphanumeric gestures correctly. When using only prior knowledge learned by HMMs without observation augmentation, 100% of symbol "5" and 75% of symbols "2", "3" and "l" were recognised. Using the CONDENSATION algorithm with observation vector (dx, dy), only 25% of gestures "2" and 50% of symbol "3" were recognised, 25% of gesture "5" were misclassified as "2" and 75% of gesture "5" were not recognised at all. Some gestures are positional dependent (object-centred), e.g. between waving high and waving low. Modelling observation on positions becomes necessary. The CONDENSATION algorithm performed worse when (x, y, dx, dy) was used as the observation vector with only gesture "l" and 50% of gesture "3" recognised. Similar relative performance can be seen in Table 2 that describes the results of applying the four algorithms in the waving and pointing gestures. Using 160 density samples and 10% for initialisation, observation augmented propagation of density functions only misclassified 25% of "waving high" gesture for "pointing right", and 33% of "waving low" gesture for "pointing right". Using only the prior knowledge learned through HMMs, there are 33% of "waving high" and 41% of "waving low" gestures misclassified as "pointing right". However, using the CONDENSATION algorithm with observation vectors (dx, dy) or (x, y, dx, dy) the system misclassified all "waving high" gestures and misclassified 32% of "waving low" gestures for either "pointing right" or "waving high" and 16% of "pointing right" gestures for "waving low". It is important

	aug. (x, y, dx, dy)				**non-aug.** (x, y, dx, dy)			
%	**2**	**3**	**5**	**l**	**2**	**3**	**5**	**l**
2	100	0	0	0	75	0	0	0
3	0	100	0	0	0	75	0	0
5	0	0	100	0	0	0	100	0
l	0	0	0	100	0	0	0	75
Er	0	0	0	0	25	25	0	25

	cond. (dx, dy)				**cond.** (x, y, dx, dy)			
%	**2**	**3**	**5**	**l**	**2**	**3**	**5**	**l**
2	25	0	25	0	0	0	0	0
3	0	50	0	0	0	50	0	0
5	0	0	0	0	0	0	0	0
l	0	0	0	100	0	0	0	100
Er	75	50	75	0	100	50	100	0

Table 1. Confusion matrices (CMs) for test sequences of set A using 40 density samples. Good recognition performances give diagonal CMs of high values.

to point out that the CONDENSATION based recognition has to be based on a set of carefully chosen noise parameters for prediction due to the lack of modelling prior on both state transitions and observation covariance. These parameters are sensitive to observation changes and rather *ad hoc*.

	aug. (x, y, dx, dy)				**non-aug.** (x, y, dx, dy)			
%	PL	PR	WH	WL	PL	PR	WH	WL
PL	83	0	0	0	90	0	0	0
PR	0	75	25	33	0	83	33	41
WH	0	0	75	0	0	0	67	0
WL	0	0	0	59	0	0	0	33
Er.	17	25	0	8	10	17	0	26

	cond. (dx, dy)				**cond.** (x, y, dx, dy)			
%	PL	PR	WH	WL	PL	PR	WH	WL
PL	91	0	58	0	100	0	33	0
PR	0	59	0	16	0	68	9	16
WH	0	9	0	16	0	0	0	16
WL	0	16	42	59	0	16	42	59
Er	9	16	0	9	0	16	16	9

Table 2. Confusion matrices for test sequences of set B (PL: point left, PR: point right, WH: wave high, WL: wave low) using 160 density samples.

Figure 4. Frames 20, 50, 85, 100, 135 and 160 from a test sequence in which a novel subject points left, waves and then points right.

On continuous, multiple gestures: Figure 4 shows an example test sequence in which a novel subject points left, waves low, waves high and points right continuously. Figure 5 shows the gesture likelihoods computed by matching the gesture models to this sequence using (1) observation augmented propagation of densities using prior (two top rows), (2) non-augmented propagation of densities using prior (two middle rows), (3) a CONDENSATION-based algorithm (two bottom rows). For each algorithm we have shown the model probability estimation for each gesture and the final probability estimation. The results illustrate that the CONDENSATION-based algorithm confused gestures "waving high" and "pointing right", the non-augmented algorithm was not able to classify the "waving high" and "pointing right" gestures. The observation augmented algorithm classified all gestures.

On recognition rate: Figure 6 (top) shows the recognition rate of all the alphanumeric gestures using augmented, non-augmented, and the CONDENSATION algorithm with window size $w = 1$ and $w = 5$. Using the observation augmented algorithm a 100% recognition rate was achieved using only 40 samples, compared to the 150 samples required for the non augmented algorithm. In contrast the CONDENSATION algorithm only achieved 80% recognition rate when 360 samples were used. It is important to note that the CONDENSATION algorithm performed worse when window size $w = 5$ is used.

Similar results can be seen in Figure 6 (bottom) for the pointing and waving gestures. A 75% recognition rate was achieved using the observation augmented algorithm using only 80 samples. A similar recognition rate was achieved without augmentation using 160 samples. The CONDENSATION algorithm was only able to achieve a 50% recognition rate even when 360 samples were used.

6 Conclusion

In this work, we introduced a new framework to model and recognise temporal structures of human activities such as gestures based on conditional density propagation via learned prior knowledge on the structures. Prior knowledge is learned from training examples using methods adapted from both Hidden Markov Models and Expectation Maximisation based clustering.

From these experiments, we have shown that using learned prior we can achieve a recognition rate that is about 25% higher compared to that achieved using methods without prior based on CONDENSATION. It is significant that such performance improvement is achieved with less computational cost since it only requires a smaller number of samples. Introducing online observation augmented density propagation further allows us to use only 25% of the number of samples used with the non augmented algorithm and 10% of the number of samples used with the CONDENSATION algorithm without any loss of recognition rate.

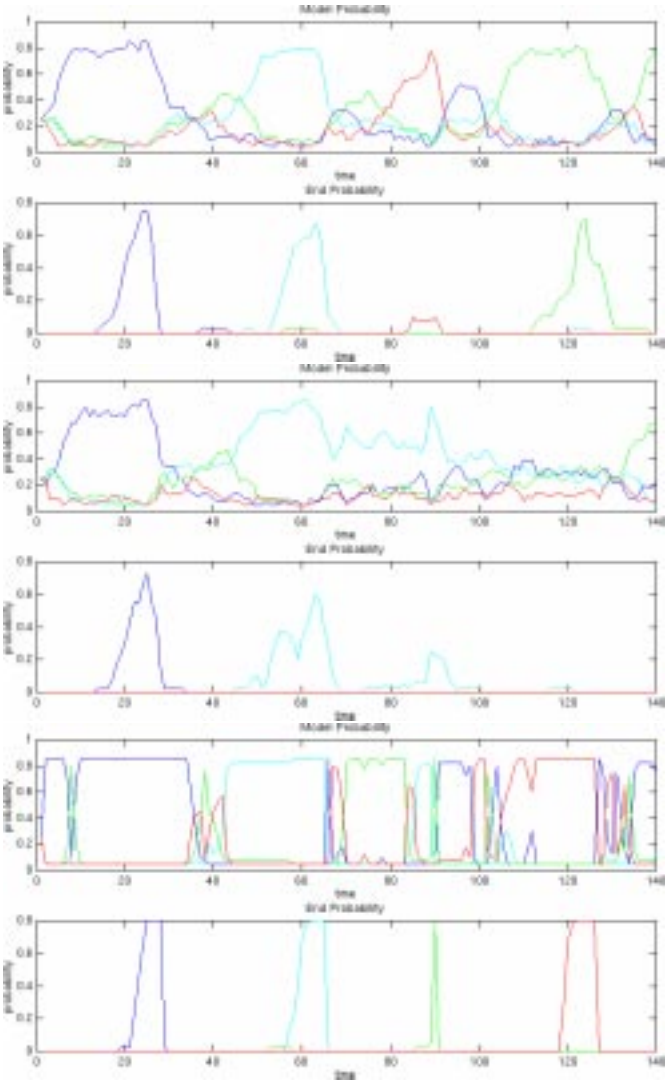

Figure 5. Gesture likelihoods estimated from the waving sequence using observation augmented propagation of density functions (top two rows), propagation of density functions using only prior (middle two rows), the CONDENSATION algorithm (bottom two rows). The number of density samples used for propagation is 40.

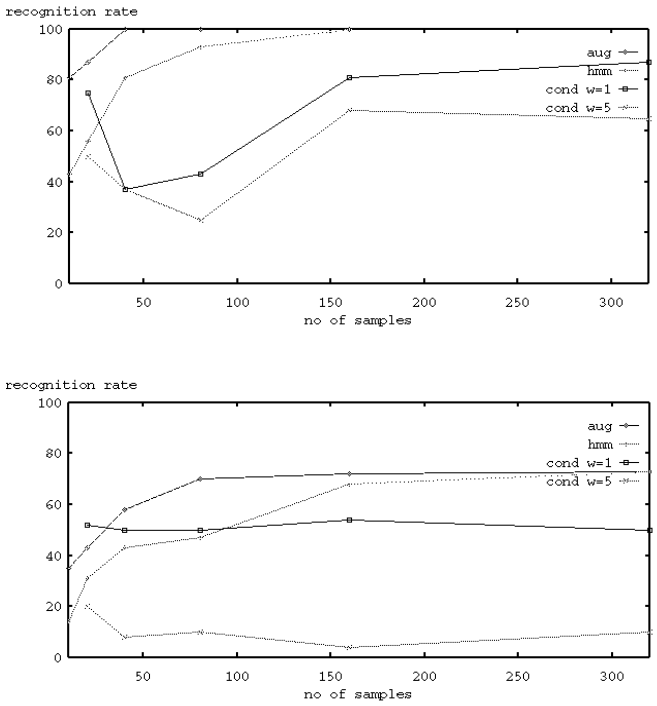

Figure 6. Recognition rate on sets A (top) and B (bottom) using observation augmented propagation of density functions, propagation of density functions using HMM prior and the CONDENSATION algorithm.

References

[1] M. Black and A. Jepson. Recognizing temporal trajectories using the condensation algorithm. In *IEEE Conf. on Face & Gesture Recognition*, pages 16–21, Nara, Japan, 1998.

[2] A. Bobick and A. Wilson. A state-based approach to the representation and recognition of gesture. *IEEE PAMI*, 19(12):1325–1338, December 1997.

[3] J. Davis and A. Bobick. The representation and recognition of action using temporal templates. In *CVPR*, pages 928–934, Puerto Rico, June 1997.

[4] A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the em algorithm. *J. Roy. Stat. Soc.*, 39(1):1–38, 1977.

[5] S. Gong and H. Buxton. On the expectations of moving objects. In *ECAI*, pages 781–786, Vienna, Austria, 1992.

[6] M. Isard and A. Blake. Contour tracking by stochastic propagation of conditional density. In *ECCV*, pages 343–357, Cambridge, UK, April 1996.

[7] M. Isard and A. Blake. Icondensation: Unifying low-level and high-level tracking in a stochastic framework. In *ECCV*, pages 893–909, Freiburg, Germany, June 1998.

[8] N. Johnson. *Learning object behaviour models*. PhD thesis, University of Leeds, England, September 1998.

[9] S. McKenna and S. Gong. Gesture recognition for visually mediated interaction using probabilistic event trajectories. In *BMVC*, pages 498–508, Southampton, UK, 1998.

[10] E.-J. Ong and S. Gong. A dynamic 3d human model from multiple views. In *BMVC*, Nottingham, UK, 1999.

[11] L. Rabiner and B. Juang. *Fundamentals of Speech Recognition*. Prentice Hall, New Jersey, USA, 1993.

[12] R. Rimey and C. Brown. Controlling eye movements with hidden markov models. *IJCV*, 7(1):47–66, November 1991.

[13] T. Starner and A. Pentland. Real-time american sign language recognition from video using hidden markov models. Technical Report 375, MIT Media Lab, 1995.