Learning Prior and Observation Augmented Density Models for Behaviour Recognition

Michael Walter †, Alexandra Psarrou † and Shaogang Gong ‡

† Harrow School of Computer Science, University of Westminster,
Harrow HA1 3TP, UK {zeoec,psarroa}@wmin.ac.uk

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

Abstract

Recognition of human behaviours requires modeling the underlying spatial and temporal structures of their motion patterns. Such structures are intrinsically probabilistic and therefore should be modelled as stochastic processes. In this paper we introduce a framework to recognise behaviours based on both learning prior and continuous propagation of density models of behaviour patterns. Prior is learned from training sequences using hidden Markov models and density models are augmented by current visual observation.

1 Introduction

The ability to interpret human motion patterns is an essential part of our perception. We use motion perception to communicate with other people through gestures and to understand others' actions and behaviours. In order to recognise a behaviour, it is necessary to model the underlying spatial and in particular temporal structure of its motion patterns. These structures are essentially probabilistic and often rather ambiguous. In general, they can be treated as *temporal trajectories* in a high-dimensional feature space representing closely correlated measurements on visual observations. For example, the spatio-temporal structure of a simple behaviour such as walking towards a telephone can be represented by the trajectory of an observation vector given by the mean position and displacement of the human body. 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.

Given that behaviours can be modelled by structures of probabilistic trajectories in a high-dimensional feature space, behaviour recognition can then be performed by matching these trajectories. Based on this rather general concept, recognition of gestures, for example, can be treated as the problem of matching "holistic static shape" templates in a spatio-temporal feature space [3, 8]. Unfortunately, modelling temporal structures as static templates can be sensitive to noise and ambiguities in observation trajectories. Other factors contributing to the spatio-temporal structure of a behaviour include (1) covariance in observation measurements, (2) nonlinear temporal scaling, and (3) ambiguities in temporal segmentation of a behaviour. To address such problems, the temporal structure of a behaviour can be modelled as a stochastic process under which salient phases of the structure are modelled as states. Predicting state transitions then provides more robust means

BMVC 1999 doi:10.5244/C.13.3

to cope with time scaling and avoids the need for determining the starting and ending points of behaviours. An example of this approach is the use of Markov processes.

Hidden Markov Models (HMMs) are widely used for modelling temporal structures. A HMM can perform nonlinear dynamic time warping, stretching and squashing on temporal structures. HMMs have been applied to speech recognition [9], visual focus of attention [10], learning object movement and behaviour models [4, 7] and more recently gesture recognition [11, 2]. Hidden Markov states are usually selected so as to capture the locations along the observation trajectories where measurements undergo significant change. Prior knowledge in the form of state transition probabilities and conditional observation covariances are estimated from training examples. However, the main disadvantages of HMMs are that (i) they can be used to estimate the probabilities for only one temporal model at a time and (ii) they only give an estimate of the final probability for each model.

More recently the CONditional DENSity propagATION (CONDENSATION) algorithm was proposed [5, 6]. Instead of modelling observation probabilities conditional to a finite set of discrete states, a set of probabilities for different models is continuously propagated over time. For gesture recognition, CONDENSATION has been adopted by Black and Jepson [1]. Unfortunately, the model does not consider measurement covariance therefore is sensitive to noise. It also does not use any prior knowledge 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 is needed to be initialised and then propagated over time. State predictions are simply previous states plus arbitrary Gaussian noise and the observation variances are fixed. This is computationally expensive.

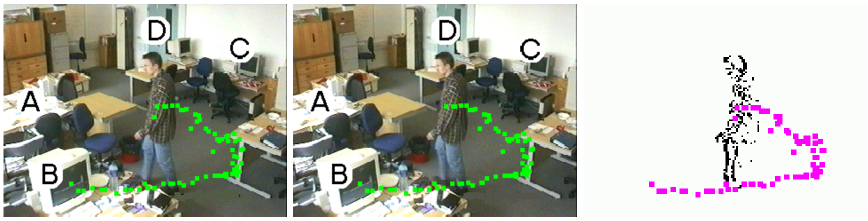

Figure 1: A behaviour of a walking person from station B to station A in an office with overlaid trajectory. Trajectories are extracted using temporal image differencing.

In this work, we introduce a method for learning both prior knowledge and a model that can be used for recognising structures of behaviour in a state space. We illustrate the method through the recognition of behaviours associated with people walking between different areas of interest, referred to as stations, in an office environment. Figure 1 illustrates an example. In this scenario, we define four stations of interest A, B, C and D and the behaviours consist of walking from one station to another. In the rest of this paper, we first introduce the concept of modelling temporal structures by statistical dynamic systems using a first-order Markov process. In Section 3 we show how this approach can be extended to (i) learn prior knowledge on both state distributions and observation covariances and, (ii) perform automatic state selection and segmentation using temporal clustering. In Section 4 we show how to continuously propagate state densities via hidden Markov states both under the constraint of the learned prior and also subject to augmentation by current visual observation. Experiments on the recognition of behaviours using this model are described in Section 5 before we conclude in Section 6.

2 Propagating First-order Markov Processes

Human behaviours are temporal. Markov processes can be used to describe statistical dynamic systems with temporal history by a sequence of characteristic states, capturing locations where the system undergoes significant changes, e.g. changes in speed or direction of a movement. The states are mutually connected and the system can undergo a change of state at discrete times. The transitions are based on a set of probabilities associated with each state and its history. State transitions at time (t) depend on the previous state at time ($t-1$) and their previous history. Observation probabilities $p(o_t|s_t)$ are assigned to each state, containing information about the likelihood of an observation o_t occurred in a specific state s_t.

If the temporal structure of a behaviour can be modelled by a dynamic system described by a Markov process, we now outline an algorithm that aims to both estimate the current model probability $p_t(s_t) \equiv p(s_t|O_t)$ and propagate multiple hypothesis of different models simultaneously over time. First, let us assume that the dynamics of behaviours can be largely regarded as first-order Markov processes, although this is not true for the structures of all behaviours. In other words, a state (s_t) at time (t) will only depend on its previous state at time ($t-1$) and is independent of its former history $S_{t-1} = (s_1, s_2, \ldots, s_{t-1})$, i.e.

$$p(s_t|s_{t-1}) = p(s_t|S_{t-1}) \tag{1}$$

Furthermore if we assume that a conditional, multi-modal observation probability $p(o_t|s_t)$ is independent of its observation history $O_{t-1} = (o_{t-1}, o_{t-2}, \ldots, o_1)$ and is equal to $p(o_t|s_t, O_{t-1})$, the conditional observation probability given the history, we can then propagate the model (posterior) probability $p(s_t|O_t)$ based on Bayes' rule as follows:

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

where $p(s_t|O_{t-1})$ is the "predicted" prior, $p(o_t|s_t)$ the conditional observation density and k_t a normalisation factor. The prior $p(s_t|O_{t-1})$ for accumulated observation history up to time ($t-1$) can be regarded as a prediction taken from the posterior at the previous time step $p(s_{t-1}|O_{t-1})$ and the state transition probability $p(s_t|s_{t-1})$:

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

Such an algorithm can be implemented using factored sampling and is known as CONDENSATION [5]. For prediction based on Equation (3), the posterior $p(s_{t-1}|O_{t-1})$ is approximated by a fixed number of state density samples. The prediction is more accurate as the number of samples increases. Samples are taken from the posterior in proportion to the state probability. This can be done by constructing a cumulative probability distribution and uniformly sampling thereafter. Some states especially those with high weights may be selected several times, thus leading to identical samples whilst others with lower probability are likely to be less frequently sampled if at all. New samples are predicted according to their state transition probabilities $p(s_t|s_{t-1})$ and a weight is assigned to all samples in proportion to the value of the observation density $p(o_t|s_t)$. The weighted sample set then serves as a representation for the posterior $p(s_t|O_t)$, and is suitable for sampling.

3 Learning Prior using HMMs and EM

It is clear that without using any prior knowledge on the structures of behaviours, the state densities $p(s_t|s_{t-1})$ can only be poorly "guessed" as the previous estimation plus arbitrary noise. As a result full estimation of the prior $p(s_t|O_{t-1})$ can only be obtained through propagation of a very large number (thousands) of samples [1]. This is both expensive and sensitive to outliers.

A potential solution to this problem can be derived from Hidden Markov Models (HMMs). A HMM is defined by a number of discrete states $q \in \{q_1, q_2, \dots q_N\}$, with probabilistic transitions between states and observation probabilities for each state. A model can be fully described by a set of probabilistic parameters $\lambda = (A, B, \pi)$ where:(1) $A \in \{a_{ij}\}$ is a set of state transition probabilities, describing the probability $p(q_{t+1} = j \,|\, q_t = i)$ of being in a state q_i at time t and q_j at time $t+1$ where $\sum_{j=1}^{N} a_{ij} = 1$; (2) $B \in \{b_j(o) = p(o_t|q_t = j)\}$ is the observation density distributions at all states. The observation density $b_j(o)$ can be discrete or continuous, e.g. a Gaussian mixture $b(o) = \sum_{k=1}^{K} c_k \, \mathcal{G}(o, \mu_k, \Sigma_k)$ with mixture coefficient c_k, mean μ_k and covariance Σ_k for the kth mixture in a given state; (3) $\pi \in \{\pi_1, \pi_2, \dots \pi_N\}$ is the initial probabilities of being in state i at time $t = 1$ where $\sum_{i=1}^{N} \pi_i = 1$. A HMM is illustrated in Figure 2.

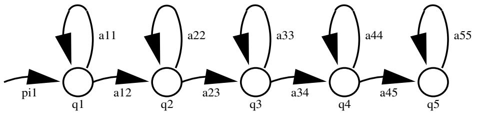

Figure 2: Example of a simple 5-state forward chained left right Hidden Markov model. Transitions are only allowed to the same or immediate next state to the right.

By training HMMs on a set of observed trajectories of behaviours, prior knowledge on both the state propagation and conditional observation density can be learned by assigning the hidden Markov state transition probabilities $p(q_t = j \,|\, q_{t-1} = i)$ to the CONDENSATION state propagation densities of

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

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

$$p(o_t \,|\, s_t) \;=\; p(o_t \,|\, q_t = j, \lambda) \;=\; b_j(o_t) \tag{5}$$

The observation covariance given by the density function at each hidden Markov state enables the model to cope with measurement covariance and noise. This defines a sample as a vector consisting of the current Markov state q_t for a given model λ.

Learning the prior involves (1) the automatic hidden state segmentation through temporal clustering, (2) the estimation of hidden state transition distribution and (3) conditional observation density distribution at each hidden state. This can be achieved using the Baum-Welch method, an iterative method, which tries to maximise the likelihood $P(O|\lambda)$ for a given model $\lambda = (A, B, \pi)$. 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 associated with each state can be performed as follows:

1. Initialise $\pi = \{1, 0, \dots 0\}$ and the state transition matrix A as

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

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

For a first-order HMM, the average time $\hat{t}$ in a state is given by

$$\hat{t} = \sum_{n=1}^{\infty} n\, a_{ii}^{n-1}\,(1 - a_{ii}) = \frac{1}{1 - a_{ii}} \tag{6}$$

and can be estimated as the ratio between the mean trajectory duration $\hat{T}$ of a behaviour in the training set and the number of states N, $\hat{t} = \frac{\hat{T}}{N}$.

2. 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 illustrates the iterative process of automatic clustering the hidden states of a walking behaviour going from station *C* to *B* in our office environment. In this example four training sequences are used. The number of hidden states is set to 12, with conditional observation density distribution set to 1.

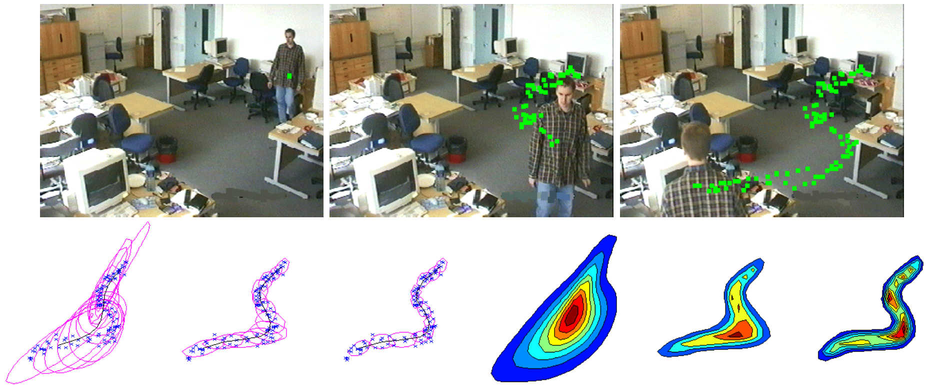

Figure 3: Learning the spatio-temporal structure of a walking behaviour going from station *C* to *B* using HMM and EM clustering. The process of iteration is shown in the bottom row from left to right. The first three images show the clustering on positions. The last three images show the corresponding density distributions over the entire structure based on the clustered hidden states and their distributions in space and time.

4 Observation Augmented Density Propagation

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, $p(s_t|s_{t-1}, o_t) = p(s_t|s_{t-1}, O_t)$. Assuming observations are independent over time and future observations have no effect on past states $p(s_{t-1}|O_t) = p(s_{t-1}|O_{t-1})$, the prediction process of Equation (3) can then be replaced by

$$\begin{aligned} p(s_t|O_t) &= \int_{s_{t-1}} p(s_t|s_{t-1}, o_t)\; p(s_{t-1}|O_{t-1}) \\ &= \int_{s_{t-1}} k_t p(o_t|s_t) p(s_t|s_{t-1}) p(s_{t-1}|o_{t-1}) \end{aligned} \tag{7}$$

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

$$\begin{aligned} p(s_t|s_{t-1}, o_t) &= \frac{p(o_t, s_t|s_{t-1})}{p(o_t|s_{t-1})} \\ &= \frac{p(o_t|s_t, s_{t-1}) p(s_t|s_{t-1})}{p(o_t|s_{t-1})} = \frac{p(o_t|s_t) p(s_t|s_{t-1})}{p(o_t|s_{t-1})} \end{aligned} \tag{8}$$

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

$$p(s_t|s_{t-1}, o_t) = p(q_t = j|q_{t-1} = i, o_t) = \frac{a^{\lambda}_{ij} b^{\lambda}_{j}(o_t)}{\sum_{n=1}^{N} a^{\lambda}_{in} b^{\lambda}_{n}(o_t)} \tag{9}$$

The observation augmented prediction unifies the processes of 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 is aimed to both improve the recognition rate and reduce the number of samples used for propagation.

5 Experiments

To illustrate our approach, we defined four stations within an office environment as shown in Figure 1. Eight behaviours were selected and a database containing 120 sequences was built. Each behaviour was performed five times by three different subjects and captured at 12 frames per second. Features were extracted using temporal image differencing and stored in a vector $o = (x, y, dx, dy)$ containing the center of mass (x, y), relative to the initial starting position and the displacement (dx, dy) of the moving person between two consecutive frames. Examples of the eight behaviours and how they overlap can be seen in Figures 4 and 5.

We built HMMs using four example trajectories from each behaviour and learned prior knowledge on the state propagation and conditional observation density. The same examples were also used to compute mean trajectories and variances for each of the behaviours that were used as models for the CONDENSATION algorithm. The remaining 11 trajectories for each behaviour were used for recognition.

Figure 6 shows the probability likelihoods for all eight behaviours shown in Figure 4 obtained by matching the behaviour models to novel trajectories using (i) observation augmented density propagation based on observation augmented prior (top two rows), (ii) non-augmented density propagation based on prior only (middle two rows) and (iii) the CONDENSATION algorithm (bottom two rows). For each algorithm we show the

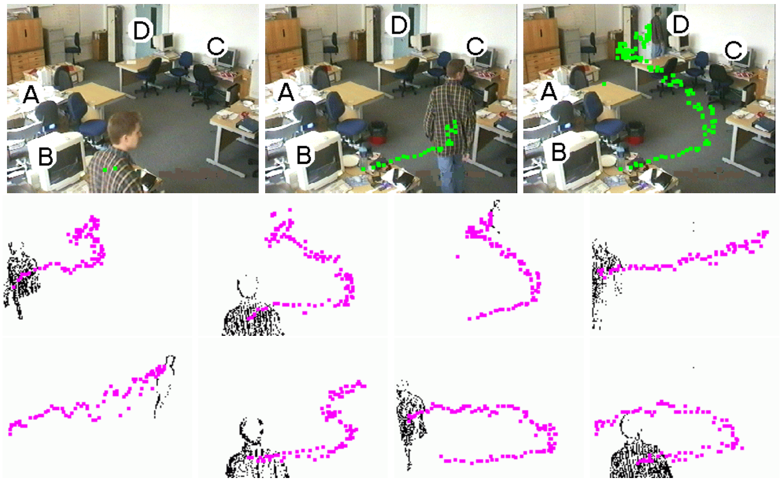

Figure 4: Example trajectory of a behaviour (top). Examples for the eight typical behaviours that the subjects perform in an office. From right to left and top to bottom: (*D to A*), (*D to B*), (*B to D*), (*C to A*), (*A to C*), (*C to B*), (*B to A*) and (*A to B*).

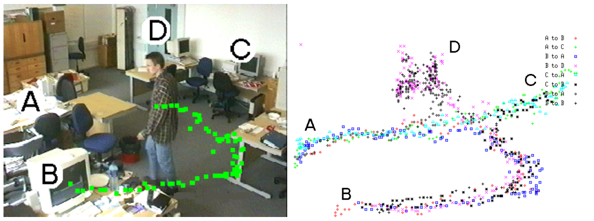

Figure 5: The office environment (left) and over-laid sample trajectories of the walking behaviours performed within such an environment (right).

model probability estimation for each behaviour during the recognition process and the estimated final probability for the recognised behaviour. Figure 6 shows that the observation augmented propagation algorithm was able to recognise all behaviours whereas non-augmented propagation algorithm was not able no recognise behaviour *(B to A)*. In addition the final probabilities estimated for the recognised behaviours by the non-augmented propagation algorithm, are lower to that estimated by the observation augmented algorithm. The CONDENSATION algorithm mis-classified behaviour *(D to B)* for *(D to A)* and *(C to A)* and was not able to recognise behaviours *(A to C)* and *(A to B)*. In general, the probability estimated for a recognised behaviour by the CONDENSATION algorithm was much lower to that estimated by both the observation-augmented and non-augmented propagation algorithms.

The block diagrams in Figure 7 show the recognition rate (black area) and the misclassification rate (dark grey area) of each of the three algorithms for each behaviour, taking into account all tested 88 novel trajectories. The observation augmented density propagation algorithm (hmm.aug) recognised most of the trajectories for all behaviours but it misclassified some of the *(B to D)* behaviours as *(B to A)*. The non-augmented density propagation algorithm (hmm.non) misclassified some of the *(A to B)* and *(A to C)* behaviours, whereas the CONDENSATION algorithm misclassified some of the *(A to B)*, *(A to C)*, *(B to A)* and *(B to D)* behaviours and failed to recognise all *(D to B)* behaviours.

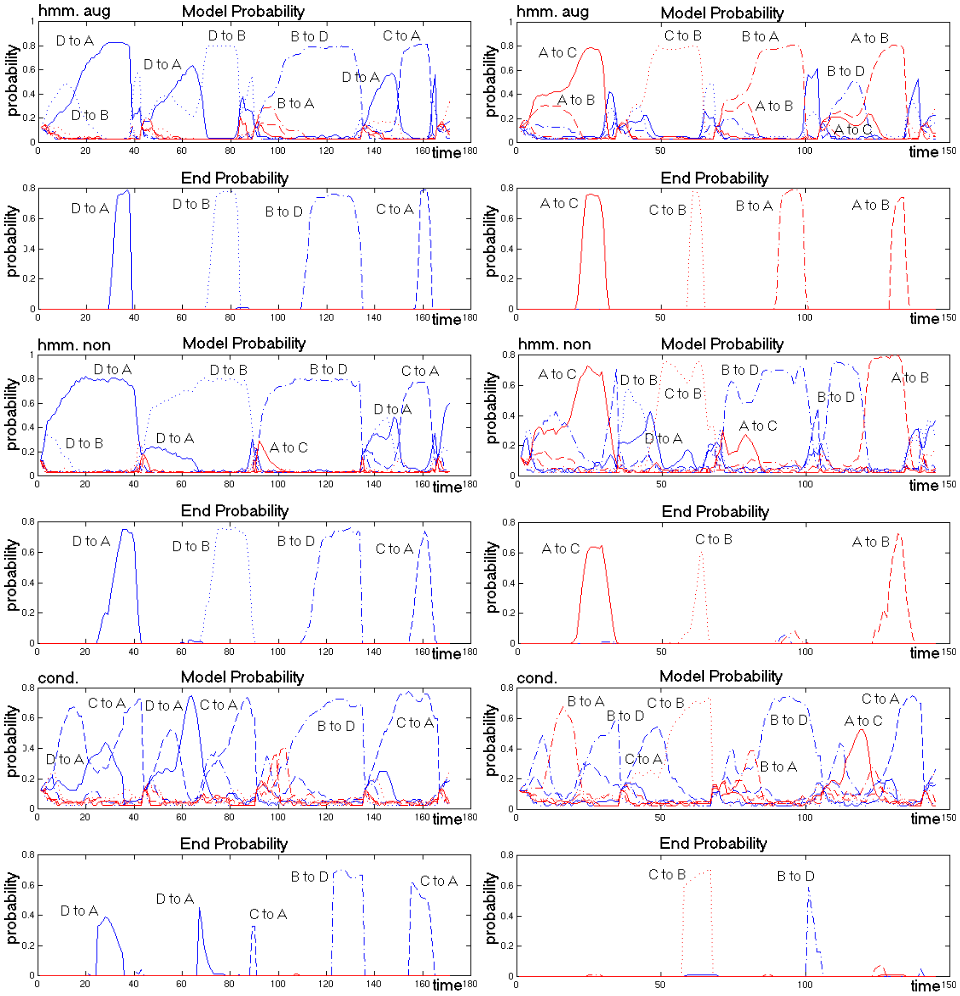

Figure 6: Behaviour likelihoods estimated over time and final probability estimation for the walking sequences (*D to A*), (*D to B*),(*B to D*),(*C to A*),(*A to C*),(*C to B*),(*B to A*),(*A to B*),using observation augmented density propagation (top two rows), non augmented density propagation using prior only (middle two rows) and the CONDENSATION algorithm (bottom two rows). The number of samples used for these experiments was 80.

Figure 8 (left) shows the recognition rate for all behaviours with respect to the number of samples propagated. Using only 160 samples the results illustrate that estimating prior knowledge and incorporating it to our behaviour models increases the overall probability estimation by 60% compared to the probability estimation given by the CONDENSATION algorithm. Further using observation augmented propagation of density functions increases the overall probability estimation by 100% compared to the estimation given by the CONDENSATION algorithm. Compared to the non-augmented propagation algorithm the probability estimation is increased by 25%.

It is also significant that the observation augmented propagation algorithm achieves a 64% recognition rate using only 40 samples (Figure 8,left) compared to the 38% recognition rate achieved by the non-augmented algorithm and 29% rate achieved by the CON-

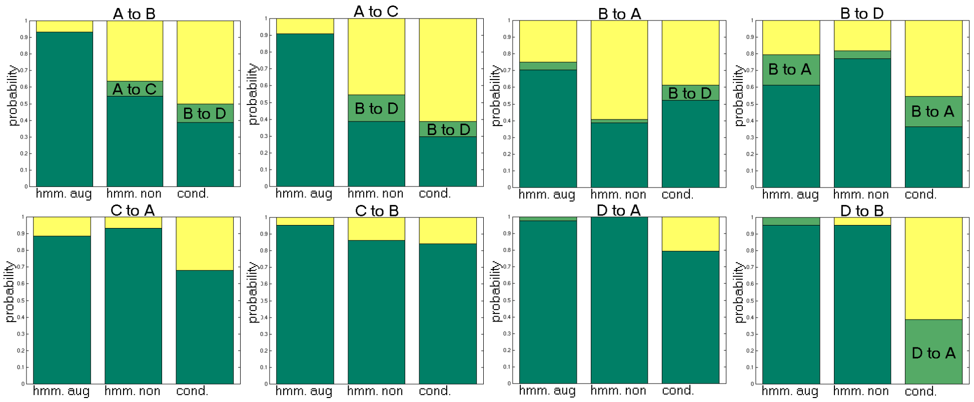

Figure 7: Recognition (black area) and misclassification (dark grey area) rate for all novel trajectories of the data set using (1) observation augmented density propagation, (2) non-augmented propagation based on prior only, (3) the CONDENSATION algorithm.

DENSATION algorithm using the same number of samples. The recognition rate of the observation augmented propagation algorithm can only be matched by the non-augmented algorithm when 640 samples are used. Using 640 sample, the observation augmented propagation algorithm gives a recognition rate over 70%.

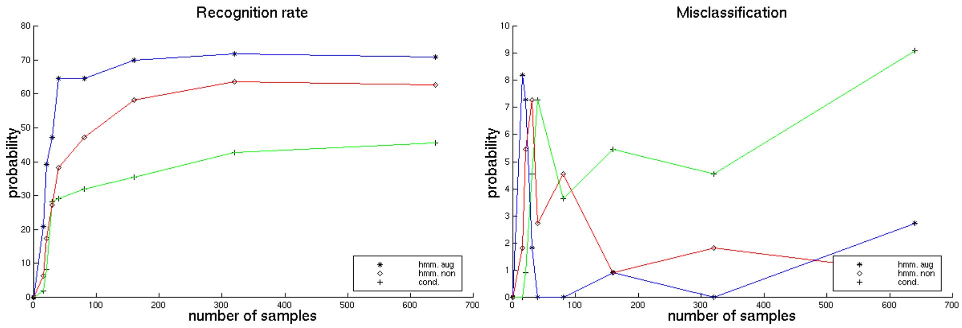

Figure 8: Recognition rate (left) and misclassification rate (right) with respect to the number of samples used for the observation augmented, non augmented and CONDENSATION algorithm.

Figure 8 (right) shows the misclassification rate with respect to the number of samples for the three algorithms. The graph illustrates that the misclassification rate is much higher for the CONDENSATION algorithm relative to both observation augmented and non-augmented algorithms.

	40	80	160	320	640
aug	8617s	7642s	9402s	14019s	24953s
non-aug	7014s	7799s	9522s	13640s	28866s
COND	6029s	6673s	8214s	11650s	20815s

Table 1: Recognition time in seconds using the observation augmented, non augmented and CONDENSATION algorithm implemented in non-optimised MATLAB running on a SGI Octane.

The different algorithms do not have significant differences in the computational costs as is shown in Table 1. Whereas the cost for the observation augmented, non-augmented and the CONDENSATION algorithm is similar, both the observation augmented and non-augmented algorithms achieve a higher recognition rate.

6 Conclusions

We described a framework to model and recognise temporal structures of human behaviours. The method is based on first-order Markov model descriptions and continuous propagation of density distributions. Prior knowledge is learned from training examples using methods adapted from Hidden Markov Models. Recognition was made more robust by taking the current observation into account to temporarily adapt the learned prior.

From the experiments we have shown that we can achieve a high recognition rate (64%) with using only a very small number of samples (40). This rate compares very favourably with 29% rate achieved by the CONDENSATION algorithm and the 38% rate achieved by the non-augmented algorithm using the same number of samples. Using 640 samples the observation augmented algorithm was able to reach a recognition rate of 72% compared to the 60% rate of the non-augmented algorithm and 43% rate of the CONDENSATION algorithm. It is significant that such performance improvement is achieved with less computational cost since both the observation augmented and non-augmented algorithms require a smaller number of samples.

References

[1] M.J. Black and A.D. Jepson. Recognizing temporal trajectories using the condensation algorithm. In *IEEE Conference on Face & Gesture Recognition*, pages 16–21, 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] S. Gong and H. Buxton. On the expectations of moving objects. In *ECAI*, pages 781–786, Vienna, Austria, August 1992.

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

[6] 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.

[7] N. Johnson. *Learning object behaviour models*. PhD thesis, School of Computer Studies, University of Leeds, England, September 1998.

[8] S. McKenna and S. Gong. Gesture recognition for visually mediated interaction using probabilistic event trajectories. In *BMVC*, volume 2, pages 498–508, Southampton, UK, 1998.

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

[10] R.D. Rimey and C.M. Brown. Controlling eye movements with hidden markov models. *Int. J. on Computer Vision*, 7(1):47–66, November 1991.

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