ELSEVIER

Image and Vision Computing 20 (2002) 349–358

image AND vision COMPUTING

www.elsevier.com/locate/imavis

Recognition of human gestures and behaviour based on motion trajectories

Alexandra Psarrou[a,*], Shaogang Gong[b], Michael Walter[a]

[a]Harrow School of Computer Science, University of Westminster, Harrow HA1 3TP, UK
[b]Department of Computer Science, Queen Mary, University of London, London E1 4NS, UK

Received 16 October 2000; accepted 18 December 2001

Abstract

Human activities are characterised by the spatio-temporal structure of their motion patterns. Such structures can be represented as temporal trajectories in a high-dimensional feature space of closely correlated measurements of visual observations. Models of such temporal structures need to account for the probabilistic and uncertain nature of motion patterns, their non-linear temporal scaling and ambiguities in temporal segmentation. In this paper, we address such problems by introducing a statistical dynamic framework to model and recognise human activities based on learning prior and continuous propagation of density models of behaviour patterns. Prior is learned from example sequences using hidden Markov models and density models are augmented by current visual observations.

Keywords: Gesture recognition; Behaviour recognition; Hidden Markov models; Condensation; Motion-based recognition; Temporal modelling

1. Introduction

The ability to interpret human gestures and behaviour constitutes an essential part of our perception. It reveals the intention, emotional state or even the identity of the people surrounding us and mediates our communication. Such human activities are characterised by the spatio-temporal structure of their motion patterns and can be modelled as temporal trajectories in a high-dimensional feature space representing closely correlated measurements of visual observations. For example, the spatio-temporal structure of a simple behaviour such as walking towards a door could be represented by the trajectory of an observation vector given by the mean position and displacement of the human body (Fig. 1). 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 or the photometric characteristics of the object of interest [5,9,15,17,18,21]. Such models can find numerous applications in visually mediated interaction [14], automated visual surveillance [4] and content-based video indexing.

Given that human activities can be modelled by structures of high-dimensional temporal trajectories, gesture and behaviour recognition can then be performed by measuring the similarity or the distance between such trajectories. Based on this general concept, recognition of human activities can be treated as the problem of matching 'holistic shape' templates in a spatio-temporal feature space [6,14]. However, modelling temporal structures as static templates can be sensitive to noise and ambiguities. Other characteristics intrinsic to the nature of motion patterns and therefore to the spatio-temporal structure of human activities include:

1. Covariance in observation measurements.
2. The temporal window of an activity cannot be constrained. An activity therefore needs to be recognised based on accumulated information and non-linear temporal scaling.
3. The occurrence of an activity in time may change resulting in ambiguities in temporal segmentation.

To address such problems we adopt a statistical modelling approach that can account for the variability in duration and temporal segmentation of the training samples. In this paper, we introduce a method for learning both prior knowledge and a model for recognising structures of human activities in a state space by continuous propagation of density models of behaviour patterns. We illustrate the method through (i) the recognition of *walking* behaviours associated with people walking between different areas of interest in an office environment; (ii) the recognition of *communicative*

* Corresponding author.
E-mail address: psarroa@wmin.ac.uk (A. Psarrou).

0262-8856/02/$ - see front matter
PII: S0262-8856(02)00007-0

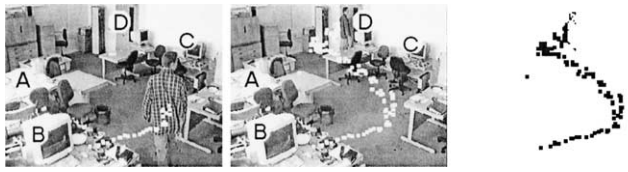

Fig. 1. A behaviour of a walking person from station B to station D in an office with overlaid trajectory. Trajectories are extracted using temporal filtering.

gestures defined within the context of visually mediated interaction [14] and (iii) the recognition of *symbol* gestures representing alphanumeric characters. Fig. 1 illustrates an example of a walking behaviour. 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. Figs. 8 and 10 illustrate examples of gestures defined within the context of visually mediated interaction and gestures representing alphanumeric symbols, respectively.

We begin by stating our motivation and reviewing previous work in this area. In Section 3 we introduce the concept of modelling temporal structures by statistical dynamic systems using first-order Markov processes. In Section 4 we show how this approach can be extended to (i) learn prior knowledge on both state distributions and observation covariances, (ii) perform automatic state selection and segmentation using temporal clustering and, (iii) 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 and gestures using this model are described in Section 5 before we conclude in Section 6.

2. Motivation and prior work

The most common solution to represent temporal information, such as human activities, has been to give it a spatial representation. However, a better approach to address the issues of uncertainty, non-linear temporal scaling and temporal segmentation intrinsic to high-dimensional temporal structures is to represent time implicitly. That is to represent time by the effect it has on processing and not as an additional dimension of the input vector.

One method is the use of procedures based on dynamic programming such as Dynamic Time Warping (DTW) or Hidden Markov Models (HMMs). Example of DTW in gesture recognition is the work of Darell and Pentland [5] and Bo-bick and Wilson [3]. Darell and Pentland applied DTW to match normalised image template correlation scores against learned spatio-temporal hand gesture models whereas Bobick and Wilson represented gesture templates as an ordered sequence of 'fuzzy' states in a configuration space and employed a DTW parsing algorithm for recognition. However, although DTW has been successful in small tasks, its main limitations are that it needs a large number of templates in order to model a range of variations and it cannot handle undefined patterns.

Temporal structures can also be modelled using HMMs as stochastic processes under which salient phases of the structure are represented as states and prior knowledge on both state distributions and observation covariances is learned from training examples. Predicting state transitions then provides more robust means to cope with time scaling and avoids the need for determining the starting and ending points of behaviours [8,16,19]. In HMMs, one of the first applications in behaviour recognition was that of Yamato et al. who used HMMs to recognise sequences of tennis strokes based on quantised time-sequential binary images [21]. HMMs with continuous observation distributions have been applied by Starner and Pentland to model American Sign Language from relatively low resolution hand tracking [20]. HMMs have also been used by Ivanov and Bobick in the recognition of atomic primitives of activities, as for example a 'square' or the movements of a music conductor [13]. However, the main disadvantages of HMMs are that (i) they can be used to estimate the probabilities for only one model at a time and (ii) they can only give an estimate of the final probability for each model.

An alternative approach is the use of artificial neural networks that assume dynamic behaviour and are responsive to time-varying information. Examples are recurrent networks [7,10,18] whose ability to store internal states and implement complex dynamics provide a natural framework for both recognition and prediction of temporal sequences. However, the main disadvantage of such networks is their inability to converge when trained with high-dimensional structures.

More recently the *conditional density propagation* (*condensation*) algorithm was proposed by Isard and Blake [11,12]. 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 [1,2]. The model performs fix-sized local linear template matching weighted by the conditional observation densities propagated according to condensation thus allowing for a global non-linear time scaling. However, the model does not use any prior knowledge on both state transitions and measurement covariance. State predictions are simply previous states plus arbitrary Gaussian noise. Consequently, a very large number of density samples (over

thousands) with localised uniform distribution is needed to be initialised and then propagated over time. This is computationally expensive.

Here we introduce a framework to recognise gestures and 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. We begin by describing the spatio-temporal trajectory exhibited by a gesture or behaviour as a first-order Markov process under which salient phases or states of the movement are explicitly modelled over time. Under this framework, temporal changes are treated as state vector transformations according to the probabilities associated with each state.

3. First-order Markov processes

Human activities are temporally ordered. We would like therefore to recognise gestures or behaviours based on a finite sequence of ordered observations $\mathcal{O}_T = \{\mathbf{o}_1, \ldots, \mathbf{o}_t, \ldots, \mathbf{o}_T\}$ where $\mathbf{o}_t$ denotes the observation vector $\mathbf{o}$ at time t. Furthermore, we would like to predict the observations as this will constrain matching in the next time frame, making tracking efficient and reliable. Markov processes can be used to describe statistical dynamic systems with temporal history by a sequence of characteristic states, capturing *landmark* locations where the system undergoes significant changes. For example, changes in speed or direction of a movement.

Let us assume that the temporal structure of a behaviour or gesture can be modelled by a dynamic system described by a first-order Markov process. In this case the conditional probability of state $\mathbf{q}_t$ given $\mathbf{q}_{t-1}$ is independent of its former history $\mathcal{Q}_{t-2} = \{\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_{t-2}\}$, i.e.

$$p(\mathbf{q}_t|\mathbf{q}_{t-1}) = p(\mathbf{q}_t|\mathcal{Q}_{t-1}) \tag{1}$$

Furthermore, we assume that a conditional, multi-modal observation probability $p(\mathbf{o}_t|\mathbf{q}_t)$ is independent of its observation history $\mathcal{O}_{t-1} = \{\mathbf{o}_1, \mathbf{o}_2, \ldots, \mathbf{o}_{t-1}\}$ and is therefore equal to the conditional observation probability given the history, i.e.

$$p(\mathbf{o}_t|\mathbf{q}_t) = p(\mathbf{o}_t|\mathbf{q}_t, \mathcal{O}_{t-1}) \tag{2}$$

We can then propagate the state probability $p(\mathbf{q}_t|\mathcal{O}_t)$ based on Bayes' rule as follows:

$$p(\mathbf{q}_t|\mathcal{O}_t) = k_t p(\mathbf{o}_t|\mathbf{q}_t) p(\mathbf{q}_t|\mathcal{O}_{t-1}) \tag{3}$$

where $p(\mathbf{q}_t|\mathcal{O}_{t-1})$ is the prior from the accumulated observation history up to time $t-1$, $p(\mathbf{o}_t|\mathbf{q}_t)$ is the conditional observation density and k_t is a normalisation factor. The prior density $p(\mathbf{q}_t|\mathcal{O}_{t-1})$ for accumulated observation history can be regarded as a prediction taken from the posterior at the previous time $p(\mathbf{q}_{t-1}|\mathcal{O}_{t-1})$ and the state transition probability $p(\mathbf{q}_t|\mathbf{q}_{t-1})$:

$$p(\mathbf{q}_t|\mathcal{O}_{t-1}) = \int_{\mathbf{q}_{t-1}} p(\mathbf{q}_t|\mathbf{q}_{t-1}) p(\mathbf{q}_{t-1}|\mathcal{O}_{t-1}) \tag{4}$$

In condensation, Eq. (4) is implemented using factored sampling and the posterior $p(\mathbf{q}_{t-1}|\mathcal{O}_{t-1})$ is approximated by a fixed number of state density samples [11]. The prediction is more accurate as the number of samples increases but there is a corresponding increase in computational cost.

In a hidden Markov model, sequences are modelled by assuming that the observations depend upon a discrete, hidden state, $\mathbf{q}_t$. The HMM hidden states are indexed by a single multinomial label that can take one of N discrete values, $\mathbf{q}_t \in \{1,\ldots,N\}$. Each of the hidden states has its own conditional probability density function $p(\mathbf{o}_t|\mathbf{q}_t)$.

4. Propagating conditional densities

Condensation does not make any strong parametric assumptions about the form of the state density, $p(\mathbf{q}_t)$, and can therefore track multiple, ambiguous targets simultaneously over time [11]. However, based on the accumulated history of the current observations $\mathcal{O}_t$ alone without any prior knowledge, the state propagation density $p(\mathbf{q}_t|\mathbf{q}_{t-1})$ is usually given as the previous estimations plus arbitrary Gaussian noise. Consequently, meaningful estimation of the history accumulated prior $p(\mathbf{q}_t|\mathcal{O}_{t-1})$ can only be obtained by propagating a very large set of conditional densities over time [2]. 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, priors on temporal structures learned from training examples should be used.

4.1. Learning prior using HMMs and EM

One solution to the problem of over-sampling in condensation is to learn and impose a priori knowledge of both observation covariance and the underlying state transition structure over time in order to constrain ambiguities in the sampling and propagation of observation conditional state densities. This is the notion of propagating observation conditional densities *with priors* (based on *landmarks*).

An HMM serves this purpose well. In other words, an HMM can be used to learn the prior knowledge of both observation covariance and state transition probabilities between a set of sparse and discrete landmark locations in the state space in order to constrain the ambiguity in continuous propagation of conditional state densities over time.

An HMM model $\lambda = (\mathbf{A}, \mathbf{b}, \pi)$ can be fully described by a set of probabilistic parameters as follows:

1. $\mathbf{A}$ is a matrix of state transition probabilities where

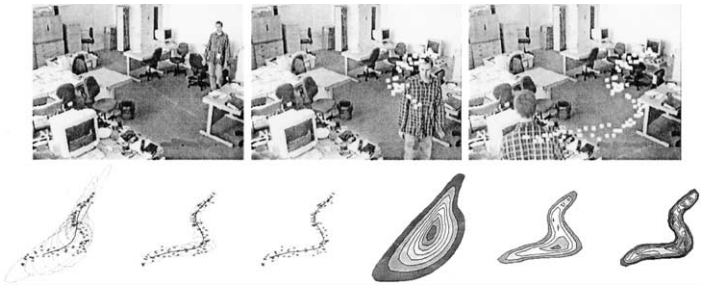

Fig. 2. Learning the spatio-temporal structure of a walking behaviour going from station C to station B using HMM and EM clustering (top). The process of iteration is shown in the bottom row from left to right. The images on the left show the clustering on positions. The images on the right show the corresponding density distributions over the entire structure based on the clustered hidden states and their distributions in space and time.

element a_{ij} describes the probability $p(\mathbf{q}_{t+1} = j|\mathbf{q}_t = i)$ and $\sum_{j=1}^{N} a_{ij} = 1$.
2. $\mathbf{b}$ is a vector of observation density functions $b_j(\mathbf{o}_t)$ for each state j where $b_j(\mathbf{o}_t) = p(\mathbf{o}_t|\mathbf{q}_t = j)$. The observation density $b_j(\mathbf{o}_t)$ can be discrete or continuous, e.g. a Gaussian mixture $b(\mathbf{o}_t) = \sum_{k=1}^{K} c_k \mathcal{G}(\mathbf{o}_t, \mu_k, \sum_k)$ with mixture coefficient c_k, mean μ_k and covariance Σ_k for the kth mixture in a given state.
3. $\boldsymbol{\pi}$ is a vector of initial probabilities of being in state j at time $t = 1$, where $\sum_{i=1}^{N} \pi_i = 1$.

Let us define the condensation state vector at time t as $\mathbf{q}_t = (\mathbf{q}_t, \lambda)$, given by the current hidden Markov state $\mathbf{q}_t$ for a model λ. By training HMMs on a set of observed trajectories of activities, a priori knowledge on both state propagation and conditional observation density can be learned by assigning the hidden Markov state transition probabilities $p(\mathbf{q}_t = j|\mathbf{q}_{t-1} = i)$ of a trained model λ to the condensation state propagation densities of

$$p(\mathbf{q}_t|\mathbf{q}_{t-1}) = p(\mathbf{q}_t = j|\mathbf{q}_{t-1} = i, \lambda) = a_{ij} \tag{5}$$

Similarly, the prior on the observation conditional density $p(\mathbf{o}_t|\mathbf{q}_t)$ is given by the Markov observation densities at each hidden state as

$$p(\mathbf{o}_t|\mathbf{q}_t) = p(\mathbf{o}_t|\mathbf{q}_t = j, \lambda) = b_j(\mathbf{o}_t) \tag{6}$$

The Markov observation density at each Markov state $b_j(\mathbf{o}_t)$ is used to provide the prior knowledge about the observation covariance. As a result, the process of both sampling and propagating condensation states is made not only more focused (guided) but also robust against observation noise.

Learning the prior involves (i) automatic hidden state segmentation through temporal clustering, estimation of (ii) hidden state transition distribution and (iii) conditional observation density distribution at each hidden state. This can be achieved using the Baum–Welch method, an iterative model that maximises the likelihood $P(\mathbf{O}|\lambda)$ for a given model λ. 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 $\boldsymbol{\pi} = \{1, 0, \ldots, 0\}$ and the state transition matrix $\mathbf{A}$ 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 $a_{ii} = 1 - (1/\hat{t})$ and $a_{ii+1} = 1 - a_{ii}$.
For a first-order HMM, the average time $\hat{t}$ in a state is given by

$$\hat{t} = \sum_{n=1}^{\infty} n a_{ii}^{n-1}(1 - a_{ii}) = \frac{1}{1 - a_{ii}} \tag{7}$$

and estimated as the ratio between the mean trajectory duration $\hat{T}$ of a behaviour in the training set and the number of states N, $\hat{t} = \hat{T}/N$.
2. Use the EM algorithm over a set of M training examples $O = \{O^1, \ldots, O^M\}$ to iteratively perform temporal clustering on the states and estimate model probability distributions $\mathbf{A}$, $\mathbf{b}$ and $\boldsymbol{\pi}$.

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

4.2. 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 $p(\mathbf{q}_t|\mathbf{q}_{t-1})$ in Eq. (4) to be augmented by the current observation, $p(\mathbf{q}_t|\mathbf{q}_{t-1}, \mathbf{o}_t) = p(\mathbf{q}_t|\mathbf{q}_{t-1}, \mathcal{O}_t)$. Assuming observations are independent over time and future observations have no

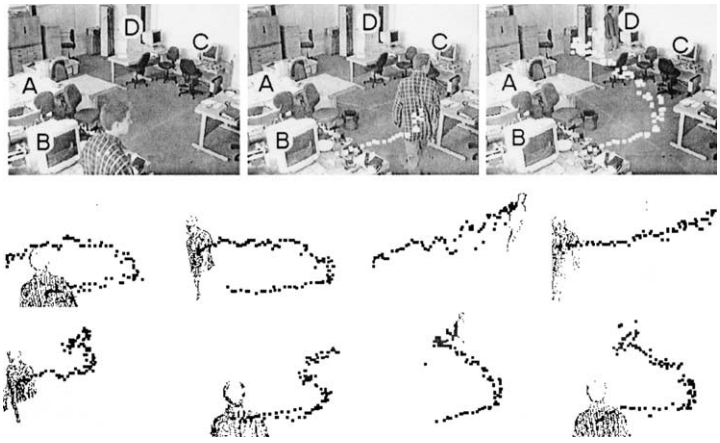

Fig. 3. An example trajectory of a person walking from station B to station D (top). Examples of the eight typical behaviours that the subjects perform in an office. From right to left and top to bottom: (A to B), (B to A), (A to C), (C to A), (D to A), (C to B), (B to D), (D to B).

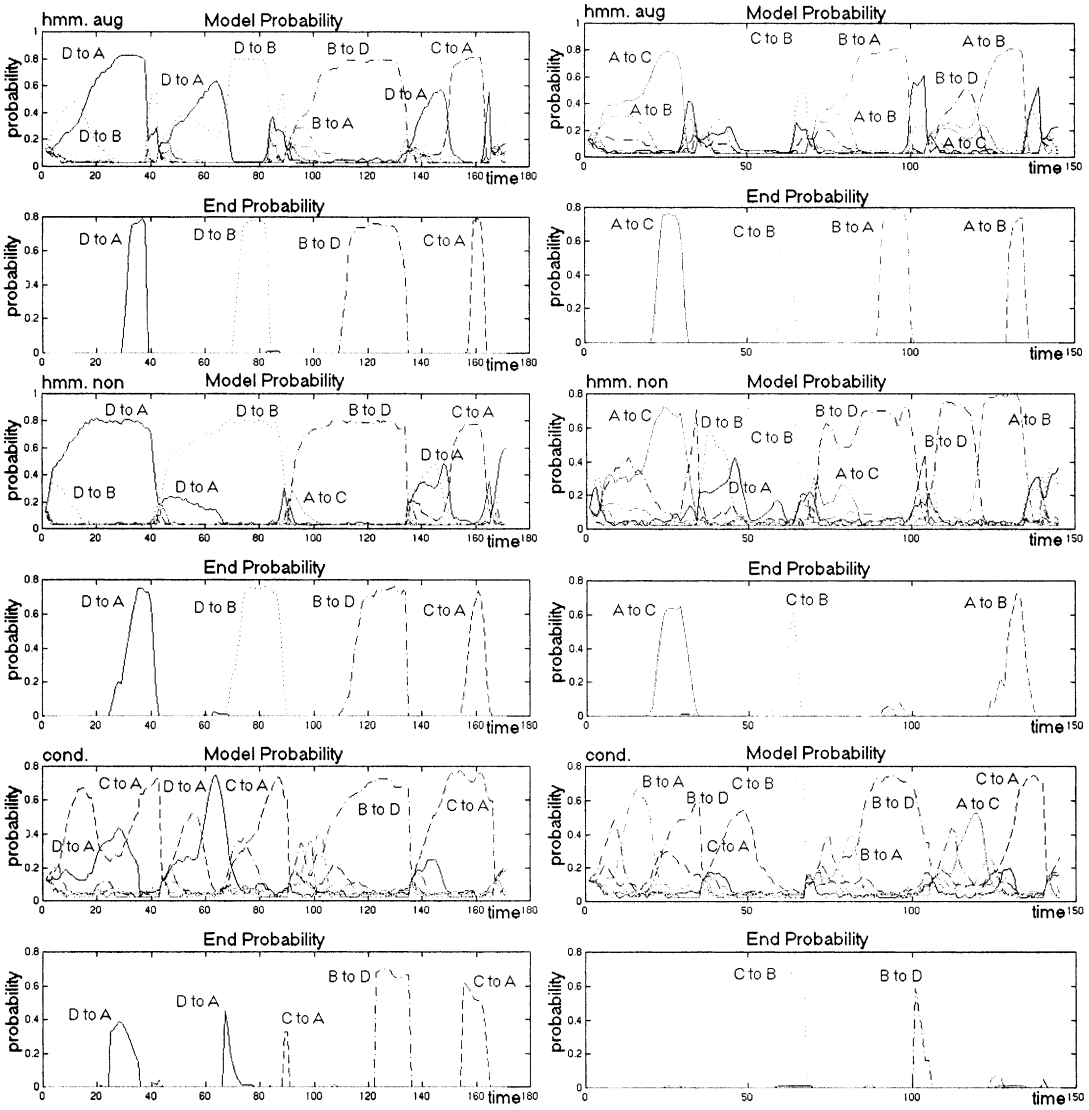

Fig. 4. Behaviour likelihoods estimated over time and final probability estimation for the walking sequences (D to A), (D to B), (B to D), (C to A), (A to C), (C to B), (B to A), (A to B), using observation augmented density propagation (top two rows), non-augmented density propagation using prior only (middle two rows) and the condensation algorithm (bottom two rows). The number of samples used for these experiments was 80.

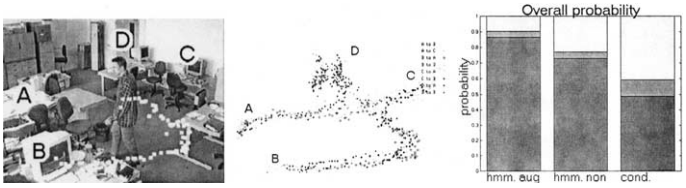

Fig. 5. The office environment (left), the overlaid sample trajectories of the walking behaviours performed within such an environment (middle) and the overall recognition and misclassification rate for these trajectories (right).

effect on past states $p(\mathbf{q}_{t-1}|\mathcal{O}_t) = p(\mathbf{q}_{t-1}|\mathcal{O}_{t-1})$, the prediction process of Eq. (4) can then be replaced by

$$p(\mathbf{q}_t|\mathcal{O}_t) = \sum_{\mathbf{q}_{t-1}} p(\mathbf{q}_t|\mathbf{q}_{t-1}, \mathbf{o}_t)p(\mathbf{q}_{t-1}|\mathcal{O}_{t-1})$$

$$= \sum_{\mathbf{q}_{t-1}} k_t p(\mathbf{o}_t|\mathbf{q}_t)p(\mathbf{q}_t|\mathbf{q}_{t-1})p(\mathbf{q}_{t-1}|\mathbf{o}_{t-1}) \quad (8)$$

where $k_t = 1/p(\mathbf{o}_t|\mathbf{q}_{t-1})$ and

$$p(\mathbf{q}_t|\mathbf{q}_{t-1}, \mathbf{o}_t) = \frac{p(\mathbf{o}_t, \mathbf{q}_t|\mathbf{q}_{t-1})}{p(\mathbf{o}_t|\mathbf{q}_{t-1})} = \frac{p(\mathbf{o}_t|\mathbf{q}_t, \mathbf{q}_{t-1})p(\mathbf{q}_t|\mathbf{q}_{t-1})}{p(\mathbf{o}_t|\mathbf{q}_{t-1})}$$

$$= \frac{p(\mathbf{o}_t|\mathbf{q}_t)p(\mathbf{q}_t|\mathbf{q}_{t-1})}{p(\mathbf{o}_t|\mathbf{q}_{t-1})} \quad (9)$$

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

$$p(\mathbf{q}_t|\mathbf{q}_{t-1}, \mathbf{o}_t) = p(\mathbf{q}_t = j|\mathbf{q}_{t-1} = i, \mathbf{o}_t) = \frac{a_{ij}^{\lambda} b_j^{\lambda}(\mathbf{o}_t)}{\sum_{n=1}^{N} a_{in}^{\lambda} b_n^{\lambda}(\mathbf{o}_t)} \quad (10)$$

The observation augmented prediction unifies the processes of innovation and prediction in condensation given by Eqs. (3) and (4). 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 both improves the recognition rate and reduces the number of samples used for propagation.

5. Experiments

We have applied our model to a set of extensive experiments on the recognition of walking behaviours and gestures. During our experiments any gesture or behaviour is recognised once the number of samples in the last two states is above a preset threshold. A gesture or behaviour is misclassified if it is incorrectly recognised, whereas it is not classified if the number of samples in the last two states is below the preset threshold for all gesture and behaviour models.

Walking behaviours: These behaviours were defined within an office environment where four stations were identified as shown in Fig. 3. 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 filtering and stored in a vector $\mathbf{o} = \{x, y, \mathrm{d}x, \mathrm{d}y\}$ containing the centre of mass, relative to the initial starting position and the displacement of the moving person between two consecutive frames. Examples for some of the behaviours and how they overlap can be seen in Figs. 3 and 5.

Prior knowledge on the state propagation and conditional observation density was learned using four example trajectories from each behaviour. 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.

Fig. 4 shows the probability likelihoods for the behaviours shown in Fig. 3. They are 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 (middle two rows) and (iii) the condensation algorithm (bottom two rows). For each algorithm we show the model probability estimation for each behaviour during the recognition process and the estimated final probability for the recognised behaviour.

It can be seen that the observation augmented propagation algorithm was able to recognise all behaviours whereas non-augmented propagation algorithm was not able to 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 was not able to recognise behaviours (A to C) and (A to B) and misclassified behaviour (D to B) as (D to A) and (C to A) and behaviour (B to A) as (B to D). In general, the probability estimated for a recognised behaviour by the condensation algorithm was much lower to that estimated by both observation-augmented propagation and non-augmented propagation using prior.

The block diagrams in Fig. 6 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 88 novel trajectories. The observation augmented density propagation algorithm (hmm.aug) recognised

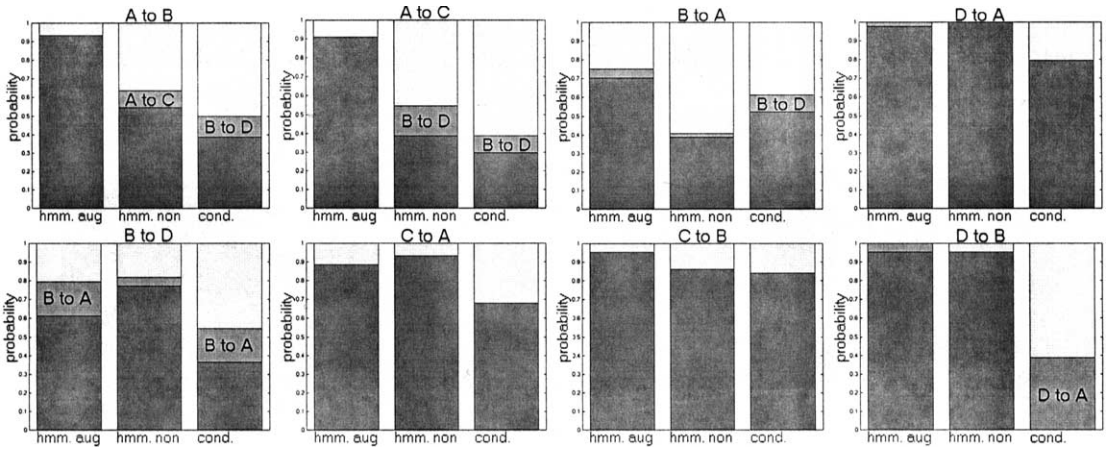

Fig. 6. Recognition (black area) and misclassification (dark grey area) rate for the walking behaviour and all novel trajectories of the data set using (i) observation augmented density propagation, (ii) non-augmented propagation using only prior, (iii) the condensation algorithm.

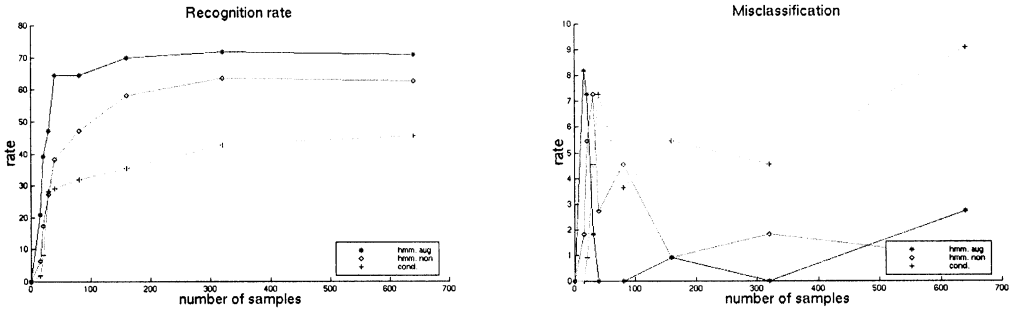

Fig. 7. Recognition rate (left) and misclassification rate (right) for the walking behaviour with respect to the number of samples used for the observation augmented, non-augmented and condensation algorithm.

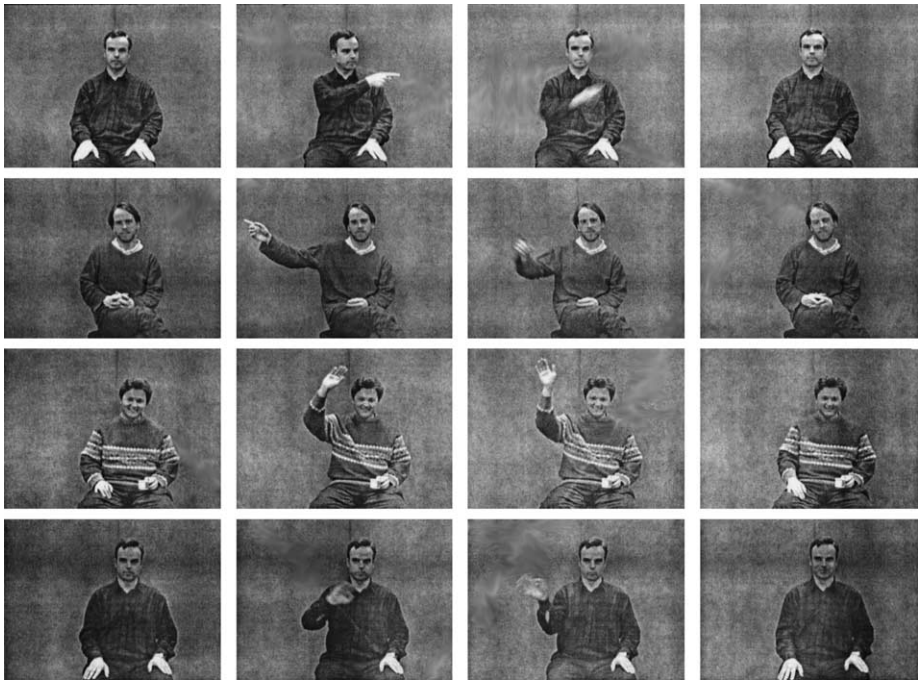

Fig. 8. Examples for the four communicative gestures in visually mediated interaction. From top to bottom: pointing left, pointing right, waving high and waving low.

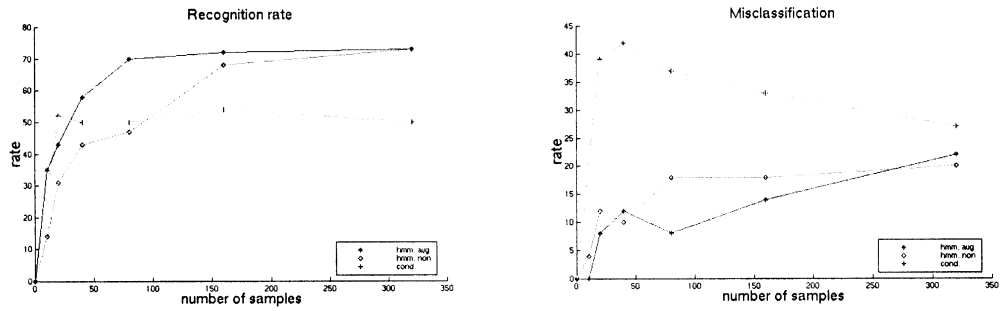

Fig. 9. Recognition rate (left) and misclassification rate (right) for the communicative gestures with respect to the number of samples used for the observation augmented, non-augmented and the condensation algorithm.

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.

Fig. 7 shows the recognition and misclassification rate for all walking behaviours with respect to the number of samples propagated. Using only 160 samples the results in Fig. 7 (left) illustrate that estimating prior knowledge and incorporating it to our behaviour models increases the overall probability estimation to 60%. Further using observation augmented propagation of density functions increases the overall probability estimation to 70%. This translates to an improvement of 60 and 100%, respectively, compared to the recognition rate achieved by the condensation algorithm. Compared to the non-augmented propagation algorithm the observation augmented recognition rate is increased by 25%. It is also significant that the observation augmented propagation algorithm achieves a 64% recognition rate using only 40 samples compared to the 38% recognition rate achieved by the non-augmented algorithm and 29% rate achieved by the condensation 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, observation augmented propagation gives a recognition rate over 70%. Fig. 7 (right) shows the misclassification rate with respect to the number of samples for the three algorithms. The graphs illustrate that the misclassification rate is much higher for the condensation algorithm compared to both observation augmented and non-augmented algorithms. It should be noted that the increase in the misclassification rate as the number of samples increase is against the non-classification rate, which is decreased accordingly.

Gestures: In addition to the walking behaviours two sets

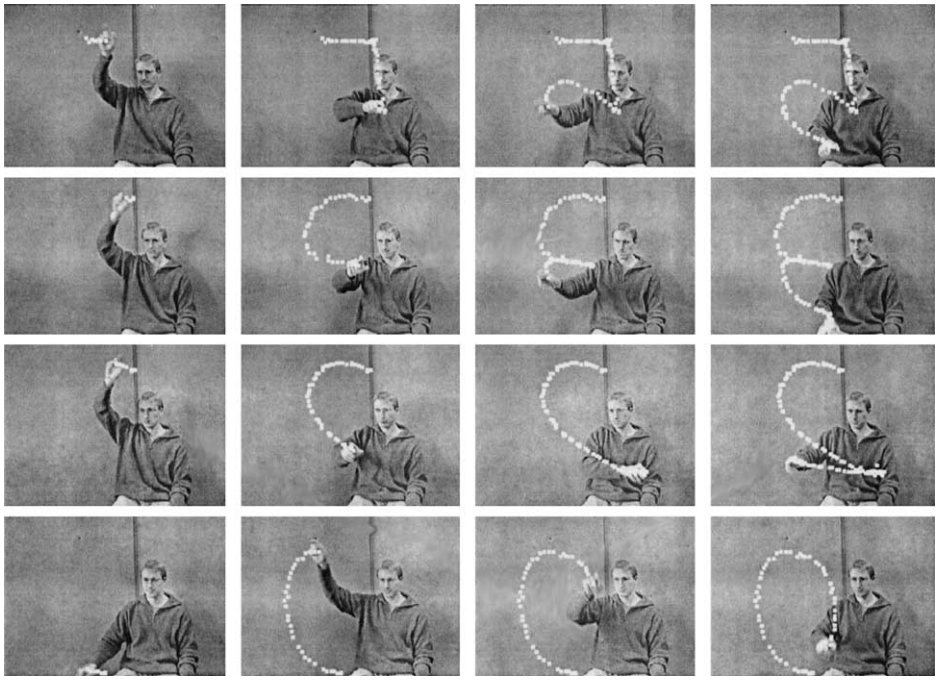

Fig. 10. Examples for the four symbol gestures defined in three sets. From top to bottom: numerals '5', '3', '2' and letter '0'.

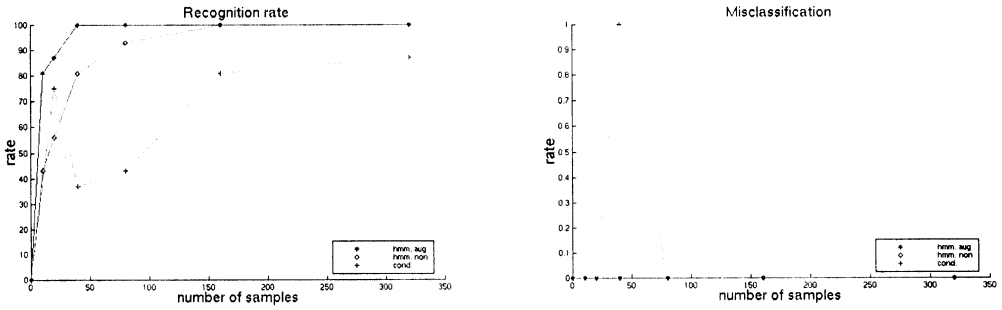

Fig. 11. Recognition rate (left) and misclassification rate (right) for the symbol gestures with respect to the number of samples used for the observation augmented, non augmented and the condensation algorithm.

of gestures have been used: (i) a set of communicative gestures defined within the context of visually mediated interaction and they are: pointing left, pointing right, waving high up and waving low down [14] and (ii) a set of symbol gestures similar to that defined in [2] and they are the numerals '2', '3', '5' and letter 'l'. In the communicative gestures the object-centred position and displacement $\{x, y, \mathrm{d}x, \mathrm{d}y\}$ of a gesture in time t is determined using moment features estimated from image motion as described in [14]. In the symbol gestures, in addition to image motion the skin colour of the hand was also used for extracting the observation vector $\{x, y, \mathrm{d}x, \mathrm{d}y\}$. As a result, the symbol gestures are less noisy than the communicative gestures. A database of image sequences was collected and for the purpose of these experiments we build HMMs using four examples of each symbol gesture and six examples of each subject performing communicative gestures. Each sequence has on average 40 frames captured at 12 Hz. Examples of the communicative gestures and symbol gestures can be seen in Fig. 8 and 10, respectively.

Fig. 9 and 11 show the recognition and misclassification rate for the communicative and symbol gestures, respectively. The results for the communicative gestures shown in Fig. 9 (left) illustrate that using 160 samples the observation augmented and non-augmented propagation density algorithms increase the overall probability estimation by 40% compared to the probability estimated by the condensation algorithm. It is important to note that using only 80 samples the observation augmented algorithm achieves a 70% recognition rate. Fig. 11 (left) illustrates that using 160 samples to recognise the symbol gestures, the observation augmented and non-augmented algorithms increase the overall probability estimation by 25% compared to the probability estimated by the condensation algorithm. The improved performance of the condensation algorithm is due to the less noisy nature of the symbol gestures.

Figs. 9 and 11 (right) show the misclassification rate with respect to the number of samples for the three algorithms. The graphs illustrate that for both communicative and symbol gestures the misclassification rate is much higher for the condensation algorithm compared to both observation augmented and non-augmented algorithms.

6. Conclusions

We described a statistical dynamic framework to model and recognise human activities in a state space based on learning prior and the continuous propagation of density models of behaviour patterns. Prior is learned from training sequences using hidden Markov models recognition is made more robust using density models augmented by current visual observation. The ability of the framework to address the problems of uncertainty, non-linear scaling and temporal segmentation intrinsic to temporal structures, is illustrated through the recognition of a set of walking behaviours and gestures and compared to that of condensation. From the experiments we have shown that both the observation augmented and non-augmented algorithms achieve a much higher recognition rate compared to condensation. The recognition rate of the walking behaviour is increased by 70% and the recognition rate of the communicative and symbol gestures is increased by 40% and 25%, respectively. In addition we have shown that we can achieve a high recognition rate for the walking behaviour and gestures with using only a small number of samples (40). 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, A.D. Jepson, A probabilistic framework for matching temporal trajectories: condensation based recognition of gestures and expressions, in: H. Burkhardt, B. Neumann (Eds.), Lecture Notes in Computer Science, vol. 1406, Springer, Berlin, 1998, pp. 909–924.

[2] M.J. Black, A.D. Jepson. Recognising temporal trajectories using the condensation algorithm, IEEE Conference on Face and Gesture Recognition, Japan, 1998, pp. 16–21.

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

[4] H. Buxton, S. Gong, Visual surveillance in a dynamic and uncertain world, Artif. Intell. 78 (1–2) (1995) 431–459.

[5] T. Darell, A. Pentland, Space-time gestures, CVPR, 1993, pp. 335–340.

[6] J. Davis, A. Bobick, The representation and recognition of action

using temporal templates, CVPR, Puerto Rico, June, 1997, pp. 928–934.
[7] J.L. Elman, Finding structure in time, Cogn. Sci. 14 (1990) 179–211.
[8] S. Gong, H. Buxton, On the expectations of moving objects, ECAI, Vienna, Austria, August, 1992, pp. 781–786.
[9] S. Gong, A. Psarrou, I. Katsoulis, P. Palavouzis, Head tracking and dynamic face recognition, European Workshop on Combined Real and Synthetic Image Processing for Broadcast and Video Production, Hamburg, Germany, 1994, pp. 96–111.
[10] S. Haykin, Neural Networks: A Comprehensive Foundation, MacMillan, New York, 1994.
[11] M. Isard, A. Blake, Contour tracking by stochastic propagation of conditional density, ECCV, Cambridge, UK, April, 1996, pp. 343–357.
[12] M. Isard, A. Blake, Icondensation: unifying low-level and high-level tracking in a stochastic framework, ECCV, Freiburg, Germany, June, 1998, pp. 893–909.
[13] Y.A. Ivanov, A.F. Bobick, Recognition of visual activities and interactions by stochastic parsing, IEEE PAMI 22 (8) (2000) 852–872.
[14] S. McKenna, S. Gong, Gesture recognition for visually mediated interaction using probabilistic event trajectories, BMVC, vol. 2, Southampton, UK, 1998, pp. 498–508.
[15] K.H. Munk, E. Granum, On the use of context and a priori knowledge in motion analysis for visual gesture recognition, Lecture Notes in Artificial Intelligence, vol. 1731, Springer, Berlin, 1998 pp. 123–134.
[16] N. Oliver, B. Rosario, A. Pentland, Statistical modelling of human interactions, IEEE Workshop on the Interpretation of Visual Motion, 1998.
[17] R. Polana, R. Nelson, Detecting activities, CVPR, New York, USA, 1993, pp. 2–7.
[18] A. Psarrou, S. Gong, H. Buxton, Modelling spatio-temporal trajectories and face signatures on partially recurrent neural networks, IEEE ICNN, Perth, Australia, November, 1995, pp. 2226–3321.
[19] R.D. Rimey, C.M. Brown, Controlling eye movements with hidden markov models, Int. J. Comp. Vis. 7 (1) (1991) 47–66.
[20] T. Starner, A. Pentland, Visual recognition of American sign language using hidden Markov models, IEEE Conference on Face and Gesture Recognition, Zurich, Switzerland, 1995, pp. 189–194.
[21] J. Yamato, J. Ohya, K. Ishii, Recognising human action in time-sequential images using hidden Markov models, CVPR, Champaign, IL, USA, June, 1992, pp. 379–385.