ELSEVIER

Image and Vision Computing 20 (2002) 873–888

image AND vision COMPUTING

www.elsevier.com/locate/imavis

On the semantics of visual behaviour, structured events and trajectories of human action

Shaogang Gong*, Jeffrey Ng, Jamie Sherrah

Department of Computer Science, Queen Mary, University of London, London E1 4NS, UK

Abstract

The problem of modelling the semantics of visual events without segmentation or computation of object-centred trajectories is addressed. Two examples are presented. The first illustrates the detection of autonomous visual events without segmentation. The second shows how high-level semantics can be extracted without spatio-temporal tracking or modelling of object trajectories. We wish to infer the semantics of human behavioural patterns for autonomous visual event recognition in dynamic scenes. This is achieved by learning to model the temporal structures of pixel-wise change energy histories using CONDENSATION. The performance of a pixel-energy-history based event model is compared to that of an adaptive Gaussian mixture based scene model.

Given low-level autonomous visual events, grouping and high-level reasoning are required to both infer associations between these events and give meaning to their associations. We present an approach for modelling the semantics of interactive human behaviours for the association of a moving head and two hands under self-occlusion and intersection from a single camera view. For associating and tracking the movements of multiple intersecting body parts, we compare the effectiveness of spatio-temporal dynamics based prediction to that of reasoning about body-part associations based on modelling semantics using Bayesian belief networks.

Keywords: Adaptive Gaussian mixture models; Autonomous visual events; Bayesian belief nets; CONDENSATION; Discontinuous motion trajectories; Dynamic scene models; Pixel-energy-history; Segmentation; Semantics of visual behaviour

1. Problem statement

Understanding visual behaviour is essential for the interpretation of human actions captured in image sequences [7,42,46]. Visual behaviours are often represented as structured patterns of visual events, e.g. ordered sequences or continuous object-centred trajectories of measurable imagery properties including object shape, colour and position. The difficulty in understanding behaviour lies with the ability (or lack of it) to automatically map such measurable visual representations to their semantics, i.e. meaning, which cannot be measured directly from the image. Despite the fact that behaviours are best defined by structured, often discrete and sparse visual events, one usually assumes that behaviour modelling starts with tracking the trajectories of the objects of interest. It is intrinsically flawed to assume that consistent object-centred trajectories can be computed in cluttered scenes with frequent object overlapping and occlusion using visual information alone without invoking semantics. To this end, we consider the following two problems:

Learning the semantics of autonomous visual events: We consider that visual events are localised autonomous changes of meaningful states in the image over time. They are not necessarily reflected by any absolute visual change such as regular pixel colour change in the scene. An *autonomous visual event* is considered to be part of a behaviour with its local semantics. For example, constant rapid scene change observed on a motorway is usually not being perceived as events of significance. However, a sudden absence of change might reveal an accident. Here we consider the problem of learning higher-level semantics through detecting local visual events without explicit object segmentation and motion grouping.

Modelling semantics for interpreting human behaviours: We also consider the problem of modelling semantics for associating erratically overlapping and discontinuous but highly natural behavioural patterns involving multiple intersecting hands, arms and head movement occurring in interactive human actions. The problem is significant because computing trajectories of multiple intersecting and overlapping objects relies heavily on object-centred

* Corresponding author. Tel.: +44-207-975-5249; fax: +44-208-980-6533.

E-mail address: sgg@dcs.qmul.ac.uk (S. Gong).

0262-8856/02/$ - see front matter
PII: S0262-8856(02)00096-3

segmentation and motion grouping, which is severely ill-defined in interactive behaviour without semantics.

2. Related work and motivation

There is an increasing body of work in the literature on modelling visual behaviours of human actions. In visual surveillance, a visual behaviour is considered to be a meaningful interpretation (labelling) of a type of object trajectories (model) extracted from image sequences, often taken by a fixed camera. The task is to learn different common trajectory models and the labelling of these models, which can subsequently be employed to detect abnormal behaviour. The typical approach has been to track individual objects in the scene [20,49,50]. Similarly, reliable markerless tracking of the human head and hands configuration is often a pre-condition for natural gesture recognition. Such tracking usually relies on edges [3,8,16, 37], skin colour or motion segmentation [22,30,34,50], or a combination of these with other cues including depth [1,27, 40,58]. If tracking is to be used for recognition, a 2D model of the body will suffice [22,34]. On the other hand, a 3D model of the body may be required for generative purposes, e.g. to drive an avatar, in which case skeletal constraints can be exploited [8,40,58], or deformable 3D models can be matched to 2D images [16,37].

Provided object-centred trajectories can be reliably extracted, a common approach to model visual behaviours of moving objects employs their motion trajectory templates [2,15,24,57]. A motion trajectory template is a holistic and static trajectory shape model of a cluster (class) of object motion trajectories in space and over time. One significant problem with this approach is uncertainty in temporal scale. It is often ambiguous over what temporal scale behaviours and events, therefore the duration of their trajectories, should be defined without any knowledge about the underlying semantics of the behaviour in question. Another significant problem is object segmentation and motion grouping. Explicitly tracking people in busy scenes such as in a shopping mall requires object segmentation that is both conceptually difficult and computationally ill defined. This is equally true when multiple occluding body parts are encountered in interactive behaviours. Stauffer and Grimson [48,49] proposed an approach for object segmentation based on pixel-wise information alone using an adaptive Gaussian mixture model of the scene background. However, the method does not overcome the difficulties in maintaining a consistent object representation from frame-wise detected regions of foreground pixels. In general, imposing the semantics of behaviours and their context is necessary for robust object segmentation and motion grouping.

The notion of semantics was originally defined in linguistics as the relationship of a representation and its meaning. This concept was further extended by Korzybski to general semantics [55]. Semantics has three basic concerns including the structure (syntax) of a representation, the processes (interpreters) operating on the representation and the meaning (truth) of the representation, also known as its semantic properties. Suppose that the representation of a human behaviour is given by its visual observations. In representational terms, such observations can be defined by either causally or temporally structured sequences of events. Different structure or order can both change the characteristics of behaviour. If behaviours are modelled as a set of discrete events, for example, in the form of a Hidden Markov Model (HMM), the notion of state transition is then regarded as the temporal structure of relating temporally ordered visual events in space and time [17,19]. State transitions are learned from example sequences of visual events often manually clustered and labelled [4,6,13,18,19,29,52]. Methods for automatic temporal clustering of HMM states have also been proposed [5,28,53,54].

Changes in the structure of a representation alters the underlying context therefore its semantics. This can be modelled as belief revision [14]. In particular, Bayesian belief networks have been widely adopted for the task of encoding knowledge as semantics of visual behaviour [18, 23,38,39,46,47]. Alternatively, Ivanov and Bobick [26] proposed to use stochastic grammar to describe high-level behaviour. Their approach tried to learn grammars from data rather than specifying them manually. What they did have to specify were ‘atomic semantic units’. We consider these atomic semantic units to be similar to our notion of visual events. Instead of manual specification, attempts have also been made to learn visual events as hidden Markov states and their transition probabilities using either entropy minimisation [5] or minimum description length [54].

In Section 3, we present a method for modelling the semantics of visual behaviour for interpreting human actions without relying on segmentation or object-centred spatial trajectories. We wish to infer semantics of higher-level behavioural patterns through monitoring visual events captured directly at individual pixels. In Section 3.1, we exploit dynamic scene models using adaptive Gaussian mixtures to bootstrap the detection of visual events. A novel approach is proposed in Section 3.2 for learning the semantics of autonomous visual events in human actions without segmentation. This is achieved by modelling the temporal structures formed by the energy histories of pixel-wise temporal change using CONDENSATION [2].

In Section 4, we present an approach for modelling the semantics of interactive human behaviour for consistent visual association of a moving head and two hands under self-occlusion from a single camera view. Occlusion occurs when a hand passes in front of the face or the other hand. Hand association requires that the hands found in the current frame be matched correctly to the left and right hands. Most existing attempts based on a single camera cope with these problems through temporal prediction which intrinsically assumes temporal order and continuity in measured visual

data. However, such an assumption is often invalid in real environments. We redefine the problem of spatio-temporal prediction as a problem of reasoning about body-part associations based on modelling semantics of interactive visual behaviour of a human body.

Experiments are provided in Section 5 to compare the performance of a pixel-energy-history based event model to that of an adaptive Gaussian mixture-based scene model, similar to that proposed by Stauffer and Grimson [49], for autonomous visual event detection. Experimental comparison of our inference-based interpreter with more commonly adopted dynamical trackers is also presented. We demonstrate that through iterative revision of hypotheses about associations of hands with skin-coloured image regions, such a semantics-based interpreter is able to recover from almost any form of tracking loss. Conclusions are drawn in Section 6.

3. Learning semantics for event detection without segmentation

Multiple object segmentation and motion grouping under occlusion cannot be achieved in general without semantics of the underlying behaviours. This suggests that individual trajectories of occluding multiple objects, which essentially rely upon segmentation and grouping, should not be the starting point for understanding their visual behaviours. We consider that behaviours are not best defined by continuous trajectories of object motion. Instead, behaviours are more closely associated with temporally structured autonomous visual events. The temporal structure of such visual events may not necessarily be governed by strict first temporal order (i.e. all past history is entirely captured by the immediate previous state) dynamics and the kind of spatial proximity typically captured by hidden Markov models. We wish to model such autonomous visual events based on pixel-wise information alone without any explicit attempt for object segmentation or motion grouping.

3.1. Modelling dynamic pixels using adaptive Gaussian mixtures

Let us examine more closely the nature of the pixel-colour spectrums. Dynamic scenes exhibit a wide spectrum of change both in terms of the speed and nature of the change occurring in individual pixels. Short-term change is mostly characterised by its temporal profile. On the other hand, long-term change manifests itself over the predominant components of a pixel's colour distribution. These components can be caused by scintillating static objects in the background or cyclically moving objects, and can be modelled by Gaussian mixtures. More specifically, given a stream of colour values for a given pixel, $\mathbf{x}_t \in \{\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_l\}$, the variation in the (r, g, b) components of $\mathbf{x}_t$ can be described in terms of Gaussian means $\boldsymbol{\mu}$ and co-variances $\boldsymbol{\Sigma}$. However, illumination specularities or swaying objects such as moving tree leaves can cause the colour distributions of pixels to split into multiple modes or clusters [44,49]. A time adaptive Gaussian mixture model $p(\mathbf{x}_t) = \sum_{i=1}^{k} \omega_i \psi(\mathbf{x}_t, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ is an effective way to model the more complex and irregular distributions, where ω_i represents the mixing parameter ($\sum_i \omega_i = 1$) and $\psi(\cdot)$ the Gaussian kernel. In unconstrained environments, the colour distributions of specific pixels rarely remain static. Changes in lighting conditions or patterns of sway in the image cause slow shifts in the parameters of a mixture model. First, we make these parameters adaptive similar to the scheme proposed in [49]. New pixel values are approximated with Gaussian clusters of pre-set co-variance according to the amount of noise and illumination variance present in the particular capturing set-up. Methods such as the EM algorithm are not viable for computing mixtures for thousands of pixels in image sequences over time [44]. Instead, the components are adapted as follows:

(1) New observations $\mathbf{x}_t$ which do not fit into any current Gaussian, i.e. have a small enough Mahalanobis distance, are assigned new Gaussians. A limit of $k_{\max}$ is set for the dynamic set of Gaussians. Once the limit is exceeded, the weakest Gaussian is replaced by the new one.

(2) For each pixel $\mathbf{x}_t$, the closest Gaussian u_c is selected to be responsible for this pixel.

(3) The means and co-variances of Gaussian u_c are updated according to a pre-determined learning rate α :

$$\boldsymbol{\mu}_{u_c,t} = (1 - \alpha)\boldsymbol{\mu}_{t-1} + \alpha \mathbf{x}_t \tag{1}$$

$$\boldsymbol{\Sigma}_{u_c,t} = (1 - \alpha)\boldsymbol{\Sigma}_{t-1} + \alpha(\mathbf{x}_t \mathbf{x}_t^T) \tag{2}$$

where $0 < \alpha < 1$.

(4) This learning rate is also applied to the mixing parameter ω_u, also referred to as the weight, of each Gaussian component u which is updated according to whether the Gaussian is responsible for pixel $\mathbf{x}_t$ at time t:

$$\omega_{u,t} = (1 - \alpha)\omega_{u,t-1} + \alpha M_{u,t} \tag{3}$$

$$M_{u,t} = \begin{cases} 1, & \text{if } u \text{ is the responsible Gaussian} \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

$$\sum_{i}^{k_{\max}} \omega_{i,t} = 1 \tag{5}$$

This is to promote the long-term over the short-term changes in the distribution.

(5) A confidence factor T is used to identify pre-dominant components from short-term components and given as a ratio of predominant Gaussian components in the distribution (between 0–1). The Gaussian components are ordered according to the products of (a) their weights, which reflect the amount of time each has

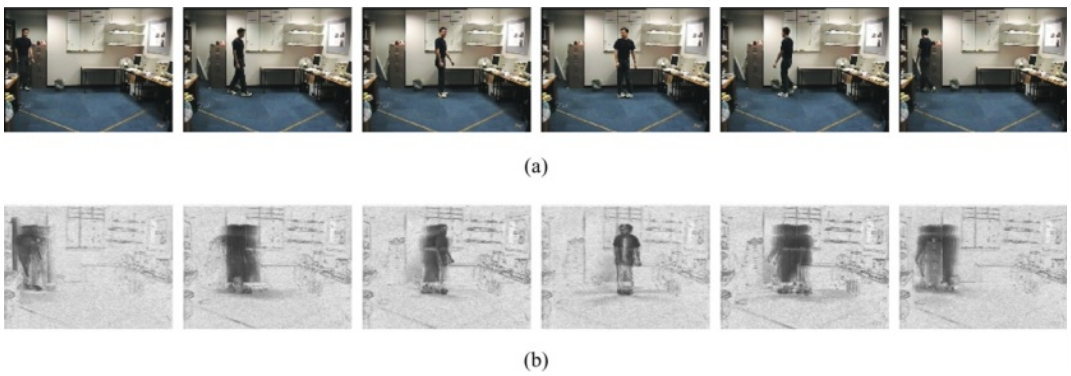

Fig. 1. (a) Selected frames from a sequence containing 20 repetitions of left-right-left movement. (b) Pixel temporal energy extracted from every image frame using a pair of quadrature filters (i.e. temporal size $T = 10$). Energy magnitude is linearly encoded as grey-level where black represents high magnitude. A log-scale has been applied to show small scale structures. Reflective edges can be seen in the image as black lines.

been observed and (b) the inverse of their generalised variances, in order to promote static objects with smaller variances. The first b Gaussians which account for a proportion T of the time are considered as the background components:

$$b = \underset{B}{\overset{k_{\max}}{\operatorname{argmin}}} \left\{ \sum_{i=1}^{B} \omega_i > T \right\} \tag{6}$$

Instead of simply using the Mahalanobis distance in a thresholded binary classification as adopted in [49], Bayes' rule is used here to formulate the probability of pixel values $\mathbf{x}_t$ belonging to a pre-learned set of long-term Gaussian component (Γ_{long}) as opposed to the recent foreground components introduced into the mixture:

$$P(\Gamma_{\text{long}}|\mathbf{x}_t) = \frac{\sum_{i=1}^{b} p(\mathbf{x}_t|i,t)P(i,t)}{p(\mathbf{x}_t)}, \tag{7}$$

where,

$$p(\mathbf{x}_t|i,t) = \sum_{i=1}^{b} \frac{1}{2\pi^{3/2}|\mathbf{\Sigma}_{i,t}|^{1/2}} \exp\left(-\tfrac{1}{2}(\mathbf{x}_t - \boldsymbol{\mu}_{i,t})^T(\mathbf{\Sigma}_{i,t})^{-1}(\mathbf{x}_t - \boldsymbol{\mu}_{i,t})\right) \tag{8}$$

$$P(i,t) = \omega_{i,t} \tag{9}$$

(6) In our case, we use $k_{\max} = 6$ so that the mixture model mostly captures the static components responsible for slow change and a few foreground components.

The predominant set of Gaussians in the mixture stores the accumulated history of the frequency of observation of each component in the mixture over a long-term scale. The state of the set can therefore capture slow changes in the colour distribution of pixels. Provided sufficient training examples are available and depending on the surveillance task, the predominant set can be locked so that new Gaussians are reported as abnormal, e.g. the introduction of a parcel in a busy scene. The long-term models can detect non-fitting fast changes, to be modelled using pixel-energy-histories as follows.

3.2. Learning temporal structures of pixel-energy-history

Rapidly changing visual phenomena exhibited by animated objects typically involves both non-rigid deformations [36] and purposeful trajectories [17,29,48]. Poor object segmentation due to the lack of semantics makes it difficult to associate frame-wise visual change to meaningful scene events. However, the temporal history of change in the appearance of pixel data itself provides useful cues as to the type of event at a higher-level structure occurring at such pixel loci. In particular, pixel-energy information gives a measure of temporal change occurring at a pixel over time. We consider that temporal structures of pixel-energy histories define autonomous visual events at a higher-level. It is important to point out that our notion of computing temporal pixel-energy is not the same as computing visual motion as adopted in [10,21]. Instead of computing motion, our aim is to extract reliable temporal change at individual pixels without attempting to establish correspondence in its local neighbourhood. We then model autonomous events through learning the temporal structures of pixel-energy histories. This is to some extent reminiscent of *topic spotting* in speech recognition.

Pixel-energy can be measured by quadrature filters [41]. We adopt the second order temporal derivatives, Laplacian of Gaussian $g(y)$ and its Hilbert transform $h(y)$ phase-shifted by 90°, as a pair of quadrature filters of temporal size T for

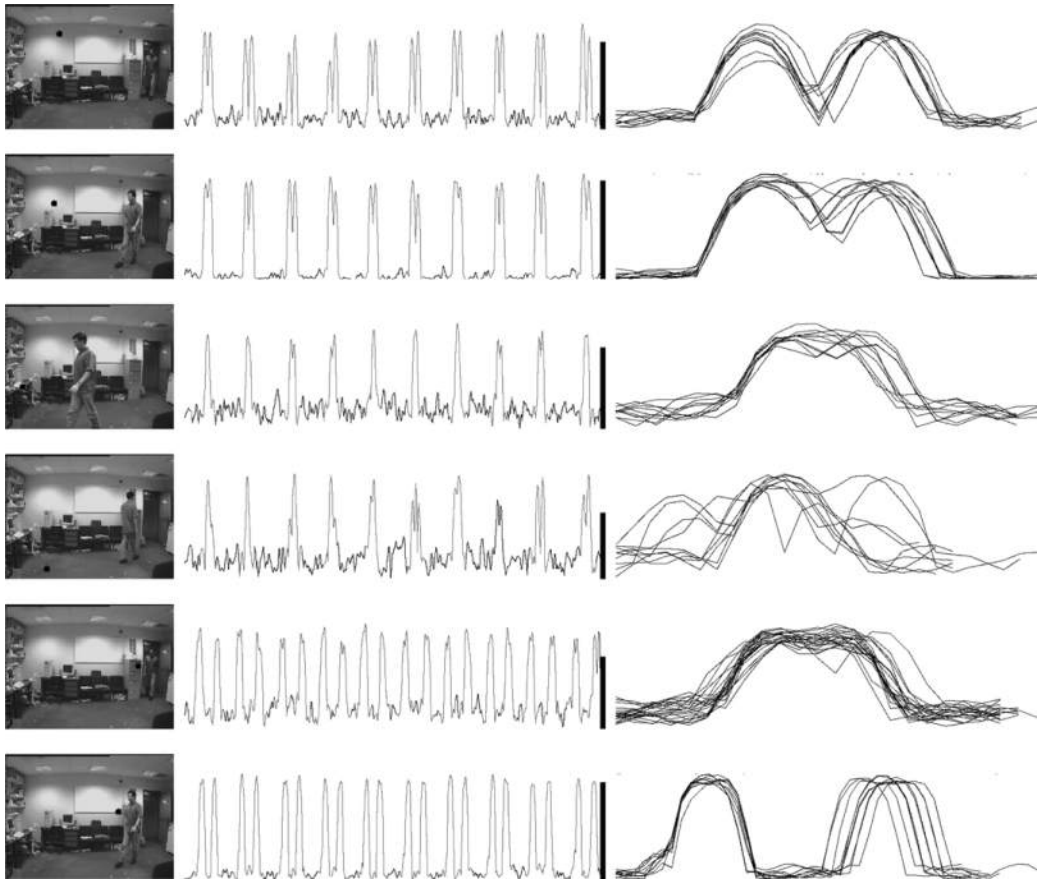

Fig. 2. From left to right. (a) Sample frames from 20 repetitions of a typical walk-about behaviour in an office environment captured by a continuous video footage. A black square dot is drawn in each frame to highlight the pixel whose history is displayed in the right column. (b) Distinctive and repeated temporal energy histories of 6 different highlighted pixels in the image. Notice that the rate of repetition at a different pixel location is different. Grey segments indicate fast change sections while black segments indicate slow change sections in the histories. (c) Duration and scale normalised fast change segments of the energy histories from 6 different pixels.

extracting pixel-energy:

$$Ph(\mathbf{x}_t) = \left(\sum_{\xi=0}^{T} g\left(\frac{3.5(\xi - T)}{T} \right) \mathbf{x}_{t-\xi} \right)^2 + \left(\sum_{\xi=0}^{T} h\left(\frac{3.5(\xi - T)}{T} \right) \mathbf{x}_{t-\xi} \right)^2 \tag{10}$$

The filter masks $g(y)$ and $h(y)$ are, respectively, defined as:

$$g(y) = \eta(2y^2 - 1)\mathrm{e}^{-y^2}, \qquad h(y) = \kappa y + \lambda y^3 \mathrm{e}^{-y^2} \tag{11}$$

where the normalising coefficients are $\eta = 0.9213$, $\kappa = -2.205$ and $\lambda = 0.9780$, as suggested in [12]. These quadrature filters can be extended to multi-scales using wavelets similar to those proposed [31]. Fig. 1 shows the pixel-energy information from an example sequence of a person walking-about in a room.

3.3. A model for detecting autonomous visual events

Fig. 2 shows pixel-energy histories of randomly selected six different pixels from 20 repetitions of a typical walk-about behaviour in an office environment captured by a continuous video footage. The third column of Fig. 2 shows the temporal structures corresponding to fast change segments in the pixel-energy histories. They clearly show to be both distinctive and repeated for each pixel. Adaptive Gaussian mixtures for long-term pixel colour distributions can be used to detect regions of slow change in pixel-energy histories and therefore extract energy histories into discrete pixel-events of 'unexpected' change. We adopt a supervised approach for learning visual events by extracting temporal structures of pixel-energy histories from a set of training sequences of 'normal' behaviour patterns. Probabilistic temporal structure models provide a mechanism for matching new observations to the pre-learned pixel-energy-history models [2,19,24,35]. Multiple hypotheses are generated using the CONDENSATION algorithm [25] to match a backward window from the signal against template windows in the models. The propagation of random samples allows for concurrent hypotheses to be maintained while providing temporal and amplitude scaling for signal-matching cross-correlation flexibility. Here we use such a model for detecting visual events. More

precisely, the matching hypotheses or states $\mathbf{s}_t$ are defined as $(\mu, \phi, \alpha, \rho)$ where μ is the model being matched, ϕ, the time index within the model, α, the amplitude scaling parameter and finally ρ, the temporal scaling parameter. A finite set of k states are propagated over time according to the observation probability, as follows:

$$P(\mathbf{y}_t|\mathbf{s}_t) = \exp\left\{-\sum_{j=0}^{w-1} \frac{\left(\mathbf{y}_{t-j} - \alpha m^{\mu}_{(\phi-\rho j)}\right)^2}{2\sigma_{\mu}(w-1)}\right\} \tag{12}$$

where σ_{μ} is the standard deviation of model μ [2]. States are randomly chosen from a cumulative probability distribution of the normalised observation likelihood of all the states in the set. Then, states with observation likelihood higher than a threshold of probable match (we use a 0.3 likelihood confidence) are propagated to the next time step according to:

$$\mu_t = \mu_{t-1}, \quad \phi_t = \phi_{t-1} + \rho_{t-1} + N, \quad \alpha_t = \alpha_{t-1} + N, \quad \rho_t = \rho_{t-1} + N \tag{13}$$

where N is propagation noise of normal distribution.

In the absence of higher-level object models of shape or appearance, the perceptual process of binding multiple-pixel information together can be facilitated by using the temporal synchrony of change in the pixels [51]. Visual events usually affect multiple pixels in the image simultaneously. Irrespective of the type of visual change occurring at the object level, the temporal energy of the involved stream of pixels should exhibit strongly correlated change, particularly in the time-delay between pixels. Preserving a common time reference for each learned pixel-energy-history model allows for cross-propagation of hypotheses of synchronous change across pixels, i.e. modelling co-occurrence. For modelling autonomous visual events, we define a temporal cross-correlation function through cross-propagation:

> *Co-occurrence dynamic*. Given a pixel $\mathbf{x}_t$ at time t, for all the samples $(\mu_t, \phi_t, \alpha_t, \rho_t)$ satisfying a matching confidence threshold, another pixel $\mathbf{y}_t$, with pixel-energy-history model μ'_t and model time index ϕ'_t, corresponding to ϕ_t is selected. A new sample $(\mu'_t, \phi'_t + \rho_t, \alpha_t, \rho_t)$ is cross-propagated into pixel $\mathbf{y}_t$ at the next time step with similar amplitude and temporal scale as the original sample.

A percentage of the states are reserved for random initialisation and cross-propagation. The probability of the change in a given pixel at time t matching the pre-learned models of normal change is given as the best cross-correlation observation probability over a set of k states:

$$P(\mathbf{y}_t) = \max_{i=1}^{k} (P(\mathbf{y}_t|\mathbf{s}_{i,t})) \tag{14}$$

Unexpected patterns of pixel change are detected when the likelihood of the change at a time instant matching pre-learned 'normal' change falls below a threshold. The likelihood threshold can be adjusted according to the desired sensitivity of detecting 'normal' change versus 'abnormal' change.

For recognising visual events in a new image sequence, this model propagates hypotheses of expected pixel-energy histories (i.e. semantics) to match with pixel-energy computed at each frame of the input sequence. Good hypotheses generate cross-hypotheses in other pixels, which have synchronous short-term fast change (i.e. co-occurrence). Events can therefore sustain adequate recognition by pixels cross-propagating hypotheses to each other. The model provides a solution for learning the binding process of pixels into higher-level visual events without object-level representation.

4. Modelling semantics of interactive human behaviours

Let us now consider the problem of modelling and encoding the semantics of a human body configuration for interpreting and tracking interactive behaviours. We are concerned with the computational task of robust and consistent association of multiple body parts in visually mediated interaction using only a single 2D view without depth information. In typical interactive behaviours, a person's hand, for example, can often move from rest to a distance half the length of their body between one frame and the next! Also, the hands may occupy regions as small as ten pixels or less wide, giving poor and incomplete visual data. Simultaneous and consistent association of erratic but highly natural behaviour patterns involving multiple occluding hands, arms and head is required.

The visual cues we have adopted for locating human head and hands in interactive behaviours are skin colour, image motion and hand orientation. Pixel-wise dynamic background colour models have been shown in Section 3.1 to be a source of robust and inexpensive visual cue for detecting visual events. Adaptive models of specific object foreground colour such as skin provides additional cues for locating and differentiating human head and hands [44,56]. Solving the problem of associating the correct hands (left or right) over time also requires the use of spatial constraints. However, situations arise under occlusion in which choosing the nearest skin-coloured cluster to the previous hand position results in incorrect hand assignment. Therefore the problem cannot be solved purely using colour and motion information. Without depth, we utilise the intensity image of each hand to obtain a very coarse measurement of hand orientation, which is robust even in very low-resolution imagery. The restricted kinematics of the human body are loosely modelled to exploit the fact that only certain hand orientations are likely to occur at a given position in the image relative to the head. The accumulation of a statistical hand orientation model is illustrated in Fig. 3. Assuming

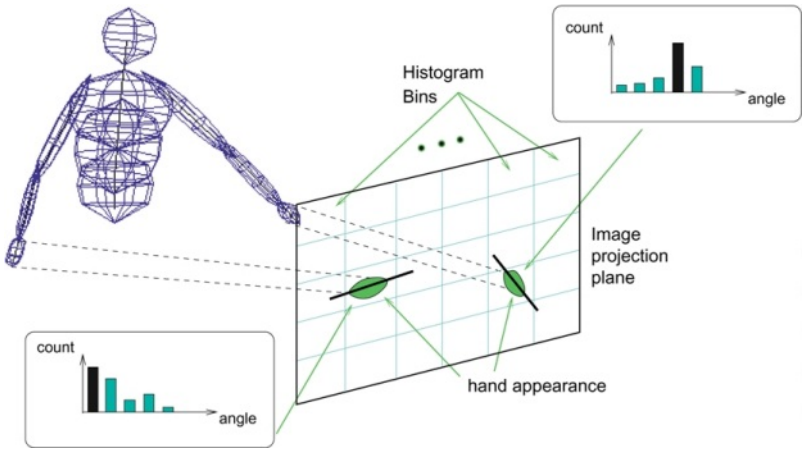

Fig. 3. Schematic diagram of the hand orientation histogram process.

that the subject is facing the camera, the image is divided coarsely into a grid of histogram bins. We then artificially synthesize a histogram of likely hand orientations for each 2D position of the hand in the image projection relative to the head position. To do this, a 3D model of the human body is used to exhaustively sample the range of possible arm joint angles in upright posture [40]. Assuming that the hand extends parallel to the forearm, the 2D projection is made to obtain the appearance of hand orientation and position in the image plane, and the corresponding histogram bin is updated. We would like to point out that the objective here is to establish, through learning, *atemporal* configurational correlations between the 3D orientations of hands and the corresponding 2D positions of both the head and hands in the image space. Such cross-level configurational knowledge aids the process of establishing object-centred trajectories when objects (i.e. the head and both hands) are constantly under occlusion and their motions are highly erratic. Tracking is therefore aided by learned atemporal correlations among different object features in both 2D and 3D. More precisely, during tracking, the quantised hand orientation is obtained according to the maximum response from a bank of oriented Gabor filters, and the tracked hand position relative to the tracked head position is used to index the histogram and obtain the likelihood of the hand orientation given the position. Details of other computations required for extracting the visual cues of human head and hands can be found in [46].

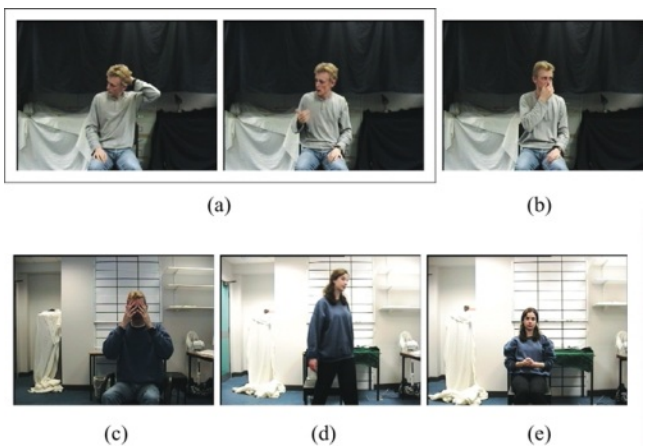

Fig. 4. Examples of the difficulties in interpreting human body parts: (a) motion is discontinuous between frames; (b) one hand occludes the face; (c) both hands occlude the face; (d) a hand is invisible; (e) both hands occlude each other.

4.1. Modelling the semantics of interactive behaviours using Bayes Net

Given the visual cues described above, the problem is now to correctly associate skin colour clusters to the left and right hands. However, only discontinuous information is available (see Fig. 4(a)). Under these conditions, explicit modelling of body dynamics inevitably makes too strong an assumption about image data. Instead, we address this problem of hand association over time under constant occlusion (i.e. discontinuity) through solving the problem of interpreting temporally structured discrete visual events using semantics. This requires full exploitation of both visual cues and the semantics of interactive human behaviour. For instance, we know that at any given time a hand is either (1) associated with a skin colour cluster, or (2) it occludes the face (and is therefore 'invisible' using only skin colour) as in Figs 4(a) and (b), or (3) it has disappeared from the image as in Fig. 4(d). When considering both hands, the possibility arises that both hands are associated with the same skin colour cluster, as when one clasps the hands together for example, shown in Fig. 4(e).

A mechanism is required for modelling semantics and reasoning about the visual cues. The obvious method of incorporating semantics into the hand association problem would be through a fixed set of rules. However there are two unpleasantries associated with this approach: brittleness and global lack of consistency. Hard rules are notoriously sensitive to noise due to their dependencies on fixed thresholds. A rule-based approach also suffers from global consistency problems because commitment to a single

decision precludes feedback of higher-level knowledge to refine lower-level uncertain observations. As a result, subsequent decision-making is isolated from the contending unchosen possibilities.

An alternative approach to reasoning is based on soft, probabilistic decisions. Under such a framework all semantics are considered to some degree but with an associated probability. Bayesian Belief Networks (BBNs) provide a rigorous framework for combining semantic and sensor-level reasoning under conditions of uncertainty [9,11,43]. Given a set of variables[1] $\mathbf{W}$ representing the scenario, the assumption is that all our knowledge of the current state of affairs is encoded in the joint distribution of the variables conditioned on the existing evidence, $P(\mathbf{w}|\mathbf{e})$. Explicit modelling of this distribution is unintuitive and often infeasible. Instead, conditional independencies between variables can be exploited to sparsely specify the joint distribution in terms of more tangible conditional distributions between variables.

A BBN is a directed acyclic graph that explicitly defines the statistical (or 'causal') dependencies between all variables.[2] These dependencies are known *a priori* and used to create the network architecture. Nodes in the network represent random variables, while directed links point from conditioning to dependent variables. For a link between two variables, $X \rightarrow Y$, the distribution $P(y|x)$ in the absence of evidence must be specified beforehand from contextual knowledge. As evidence is presented to the network over time through variable instantiation, a set of beliefs are established to reflect both prior and observed information:

$$\mathrm{BEL}(x) = P(x|\mathbf{e}) \tag{15}$$

where $\mathrm{BEL}(x)$ is the belief in the value of variable X given the evidence $\mathbf{e}$. A BBN can subsequently be used for prediction and queries regarding values of single variables given current evidence. However, if the most probable joint configuration of several variables given the evidence is required, then a process of belief revision[3] (as opposed to belief updating) must be applied to obtain the most probable explanation of the evidence at hand, $\mathbf{w}^*$, defined by the following criterion:

$$P(\mathbf{w}^*|\mathbf{e}) = \max_{\mathbf{w}} P(\mathbf{w}|\mathbf{e}) \tag{16}$$

where $\mathbf{w}$ is any instantiation of the variables $\mathbf{W}$ consistent with the evidence $\mathbf{e}$, termed an *explanation* or *extension* of $\mathbf{e}$, and $\mathbf{w}^*$ is the most probable explanation/extension. This corresponds to the locally computed function expressing the local belief in the extension:

$$\mathrm{BEL}^*(x) = \max_{\mathbf{w}'_X} P(x, \mathbf{w}'_X|\mathbf{e}) \tag{17}$$

where $\mathbf{W}'_X = \mathbf{W} - X$.

4.2. Interpreting visual events by inference

The BBN for modelling the semantics for interpreting human head and hands in interactive behaviour is shown in Fig. 5. Abbreviations are: LH = left hand, RH = right hand, LS = left shoulder, RS = right shoulder, C1 = skin cluster 1, C2 = skin cluster 2. There are 29 variables, $\mathbf{W} = \{X_0, X_1, \ldots, X_{28}\}$. The first point to note is that some of the variables are conceptual, namely X_0, X_1, X_5 and X_8, while the remaining variables correspond to image-measurable quantities, constituting $\mathbf{e}$. All quantities in the network are or have been transformed to discrete variables. The conditional probability distributions attributed to each variable in the network are specified beforehand using either domain knowledge or statistical sampling. The total state space size of the set of variables $\mathbf{W}$ is 9.521245×10^{12}, which is the number of probabilities required to explicitly represent $P(\mathbf{W})$. However, to populate the conditional and prior probability tables of the network required specification of only 456 probabilities, yet any query on the full joint distribution can still be made.

At each time step, all of the measurement variables are instantiated from observations. C1 and C2 refer to the two largest skin clusters in the image (apart from the head). Absence of clusters is handled by setting the variables X_5 and X_8 to have zero probability of being a hand. The localised belief revision method is then employed until the network stabilises and the most probable joint explanation of the observations is obtained:

$$P(\mathbf{w}^*|\mathbf{e}) = \max_{\mathbf{w}} P(\mathbf{w}|\mathbf{e}) \tag{18}$$

This yields the most likely joint values of X_0 and X_1, which can be used to set the left and hand box position.

Note that the network structure is not singly connected, due to undirected cycles formed through X_0 and X_1. Consequently the simple belief revision algorithm of Pearl [43] cannot be used due to non-convergence. Instead, we apply the more general inference algorithm of Lauritzen and Spiegelhalter [11,32]. This inference method transforms the network to a join tree, each node of which contains a sub-set of variables called a clique. The transformation to the join tree needs to be performed only once off-line. Inference then proceeds on the join tree via a message-passing mechanism similar to the method proposed by Pearl. The complexity of the propagation algorithm is proportional to the span of the join tree and the largest

[1] Upper-case is used to denote a random variable, lower-case to denote its instantiation, and boldface is used to represent sets of variables. In the rest of this section, $\mathbf{W}$ represents a set of random variables, whilst $\mathbf{w}$ represents a particular instantiation of that set of variables. X represents a single random variable, and x is a particular instantiation of that variable.

[2] Therefore the statistical independencies are implicitly defined as well.

[3] The difference between belief updating and belief revision comes about because in general, the values for variables X and Y that maximise their joint distribution are not the values that maximise their individual marginal distributions.

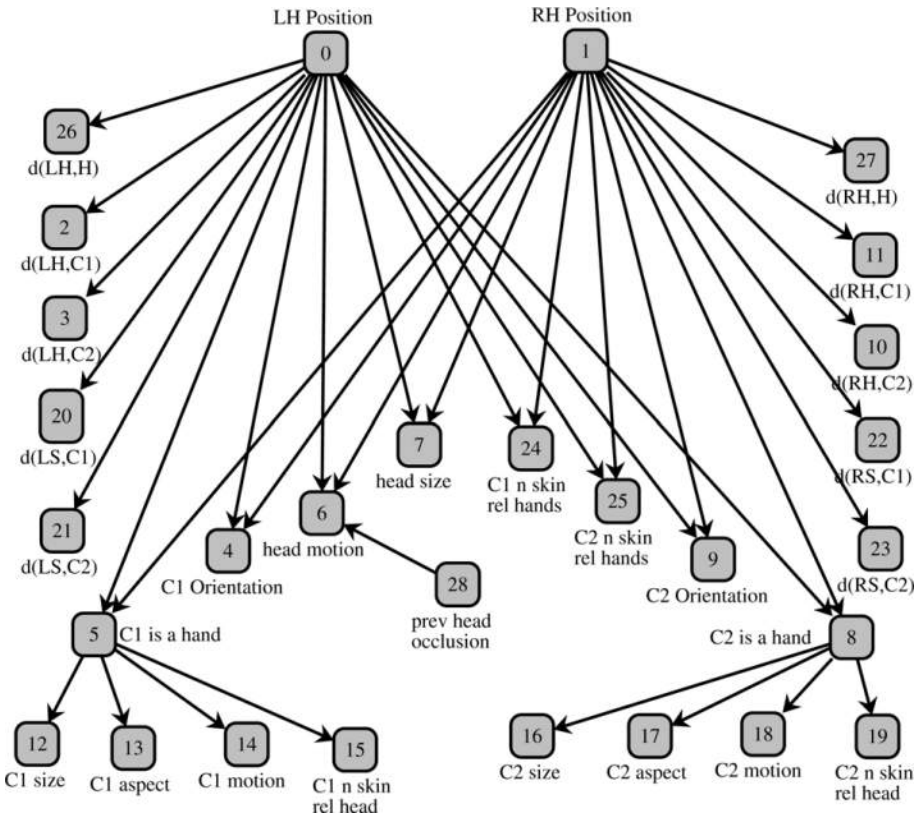

Fig. 5. A Bayesian Belief Network representing dependencies amongst variables in human body-parts configuration.

state space size amongst the cliques. The variables and their dependencies are now explained as follows:

X_0 and X_1: the primary hypotheses regarding the left and right hand positions respectively. These variables are discrete with values {CLUSTER1, CLUSTER2, HEAD} which represent skin cluster 1, skin cluster 2 and occlusion of the head, respectively. Disappearance of the hands is not modelled here for simplicity.

X_2, X_3, X_{10}, X_{11} : the distance in pixels of the previous left- or right-hand box position from the currently hypothesized cluster. The dependency imposes a weak spatio-temporal constraint that hands are more likely to have moved a small distance than a large distance from one frame to the next.

X_{20}, X_{21}, X_{22}, X_{23} : the distance in pixels of the hypothesized cluster from the left or right shoulder. The shoulder position is estimated from the tracked head box. This dependency specifies that the hypothesized cluster should lie within a certain distance of the shoulder, given by the length of the arm.

X_{26} and X_{27} : the distance in pixels of the previous left or right hand box position from the current head box position. A hand that was previously close to the current head position may now be occluding the head.

X_5, X_{12}, X_{13}, X_{14}, X_{15}; X_8, X_{16}, X_{17}, X_{18}, X_{19} : these variables determine whether each cluster is a hand. X_5 and X_8 are Boolean variables specifying whether or not their respective clusters are hands or noise. The variables have an obvious dependency on X_0 and X_1 : if either hand is a cluster, then that cluster must be a hand. The descendants of X_5 and X_8 provide evidence that the clusters are hands. X_{15} and X_{19} are the number of skin pixels in each cluster relative to the head, which has some distribution depending on whether or not the cluster is a hand. X_{14} and X_{18} are the number of motion pixels in each cluster, expected to be high if the cluster is a hand. Note that these values can still be non-zero for non-hands due to shadows, highlights and noise on skin-coloured background objects. X_{13} and X_{17} are the aspect ratios of the clusters which will have a certain distribution if the cluster is a hand, but no constraints if the cluster is not a hand. X_{12} and X_{16} are the spatial areas of the enclosing rectangles of the clusters. For hands, these values have a distribution in terms relative to the size of the head box, but for non-hands there are no expectations.

X_6 and X_7 : the number of moving pixels and number of skin-coloured pixels in the head box respectively. If either of the hands is hypothesized to occlude the head, we expect more skin pixels and some motion.

X_{28} : the previously inferred judgment of whether the head was occluded by a hand. This influences the judgment of whether motion in the head region implies a hand is now occluding or uncovering the head region.

X_{24} and X_{25} : the number of skin pixels in clusters 1 and 2 compared with a running average of observed skin pixel counts in the hand clusters.

X_4 and X_9 : orientation of the respective hand, which

Table 1
Event detection results from applying the pixel-energy-history (PEH) based autonomous event model and an adaptive Gaussian mixture (AGM) based dynamic scene model to 5 different test sequences totalling over 1700 image frames.

Visual events	No. of event occurrence	No. of image frames	PEH-based detection	AGM-based detection
Slow down	6	615	6	6
Fast move	6	255	6	6
Pause	6	362	5	0
Jump	6	356	4	0
Falling box	1	108	1	0

depends to some extent on its spatial position in the screen relative to the head box. This orientation is calculated for each hypothesized hand position. The histogram described early (shown in Fig. 3) is used to assign a conditional probability table.

In this network, all of the visual cues can be considered simultaneously and consistently to arrive at a most probable explanation of both hands and the head. BBNs lend the benefit of being able to 'explain away' evidence. For example, if the belief that the right hand occludes the face increases, this decreases the belief that the left hand also occludes the face because it explains any motion of growth in the number of skin pixels in the head region. This comes about through the indirect coupling of the hypotheses X_0 and X_1 and the fixed amount of probability attributable to any single piece of evidence. Hence probabilities are consistent and evidence is not 'double counted' [43].

5. Experiments

5.1. Learning semantics of visual events without object segmentation

To illustrate our pixel pixel-energy-history based temporal structure model for learning semantics of visual events defined by behaviour patterns in a constantly changing scene background, we designed a set of visual event detection tasks purely based on pixel information alone without any object segmentation and motion grouping.

Our model described in Section 3.2 was trained using repeated image sequences of two different people carrying out their normal routine of actions (behaviours) in an office environment. These behaviours include (a) walking from the door to the desk in the right corner of the room, (b) walking-about in the room and (c) leaving the room by the door (see examples in Fig. 6). The autonomous event model was trained to learn pixel change caused by these expected behaviour patterns in the scene based on 10 repetitions each performed by 2 different people. The 20 training sequences contained 1000 frames in total, approximately 50 frames per sequence.

The model was tested on five different sequences of behaviour patterns performed by three different people, one of whom was not present during training. The testing sequences contained similar actions to the training sequences but with differences in the characteristics of the performed movement as to add novel events to the behaviour patterns. These visual events include:

(1) *Slow down*. The testing subject suddenly walked at a slower speed for a short period of time while keeping to the same trajectory of motion.
(2) *Fast move*. The testing subject suddenly walked at a faster speed for a short period of time along the same trajectory.
(3) *Pause*. The testing subject walked as usual except for a brief pause in the middle of his walk-about activity.
(4) *Jump*. A quick jump was introduced in the middle of the walk.
(5) *Falling box*. The scene contained a number of static objects including a box on the floor. Whilst the movement and changes to image background caused by normal walk-about do not constitute an event, the

Fig. 6. Examples of training sequences illustrating a typical behaviour pattern in an indoor office environment consisting of three stages: (a) walking from the door to the desk in the left corner of the room, (b) walking-about in the room and (c) walking back to the door and leaving.

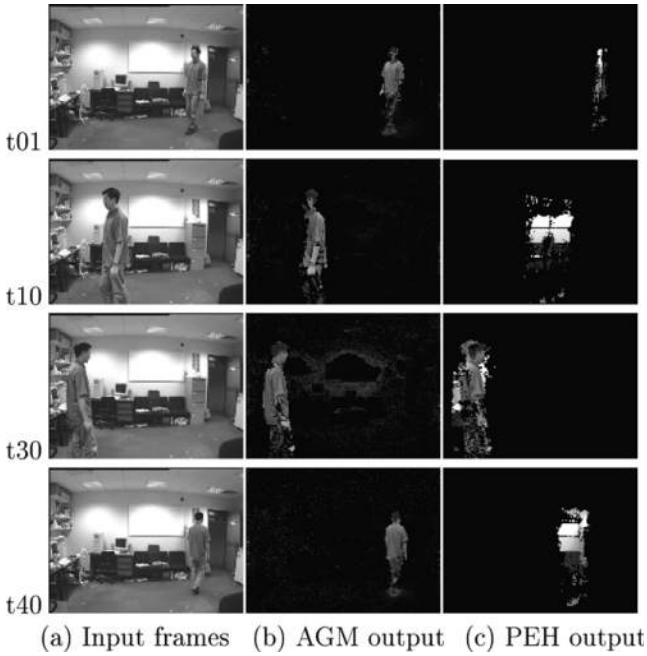

Fig. 7. Detecting the *slow down* event. (a) Image frames, (b) adaptive Gaussian mixture (AGM) model output, (c) pixel-energy-history (PEH) model output.

movement and background change caused by the box falling over do.

Table 1 shows some results from both Pixel-Energy-History (PEH) based and Adaptive Gaussian Mixture (AGM) based event recognition. Both models were able to successfully detect both *slow down* and *fast move* events in most image frames (see examples in Fig. 7). However, the PEH-based autonomous event model performs differently

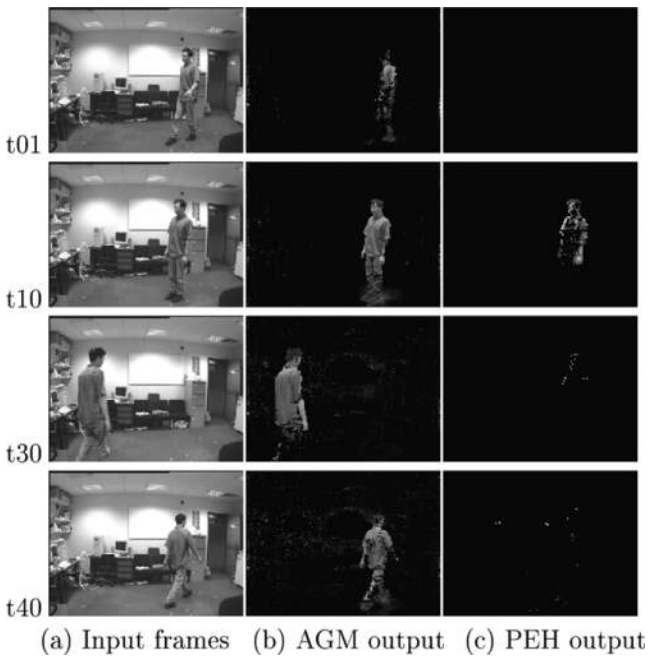

Fig. 8. Detecting the *pause* event.

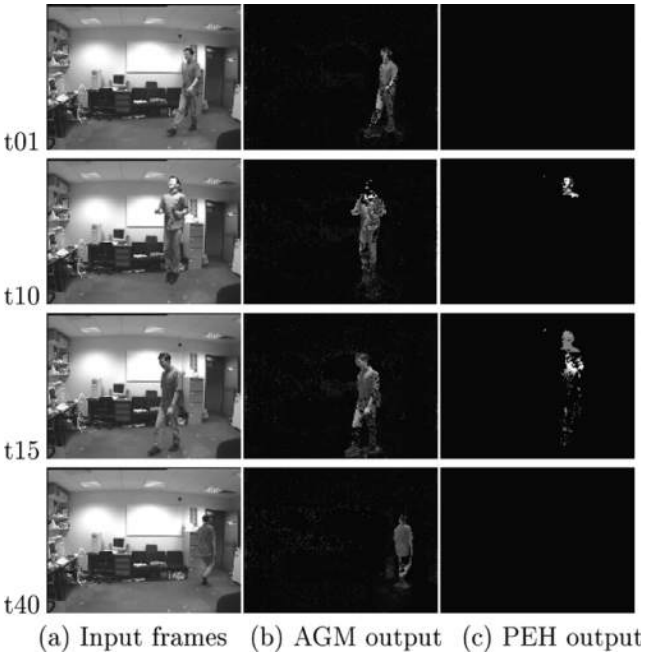

Fig. 9. Detecting the *jump* event.

on sequences containing pause and jump events as shown in Figs. 8 and 9. In the two example sequences shown most of the frames containing movement of normal walk-about behaviour. Two *pause* and *jump* events occurred only briefly in the sequences and were detected by the PEH-based autonomous event model. The process implicitly invoking higher-level scene semantics learned from individual pixel-energy histories. The AGM-based dynamic scene model did not model any knowledge of context, it detected all the movement as expected and could not differentiate between movement and visual events. In Fig. 10, the movement caused by the walking person did not constitute any event, whilst the image change caused by the *falling box* did. Both the PEH and the AGM models detected the occurrence of the *falling box* but the latter could not differentiate such an event from all other movement caused by the walking person.

These experiments suggest that higher-level scene semantics can be learned indirectly by learning the normal temporal structures of individual pixel-energy histories without the need for object segmentation and motion grouping. Such a temporal structure model has shown to be able to differentiate some primitive visual events from merely visual movement and background scene change caused by expected behaviour patterns in a given scene.

5.2. On modelling semantics for human hands association

To evaluate the performance of modelling semantics using a BBN for human hands association in natural interactive behaviours, we compare the performance of our BBN based model with two dynamical tracking methods. Note that to make our point about the difficulty of

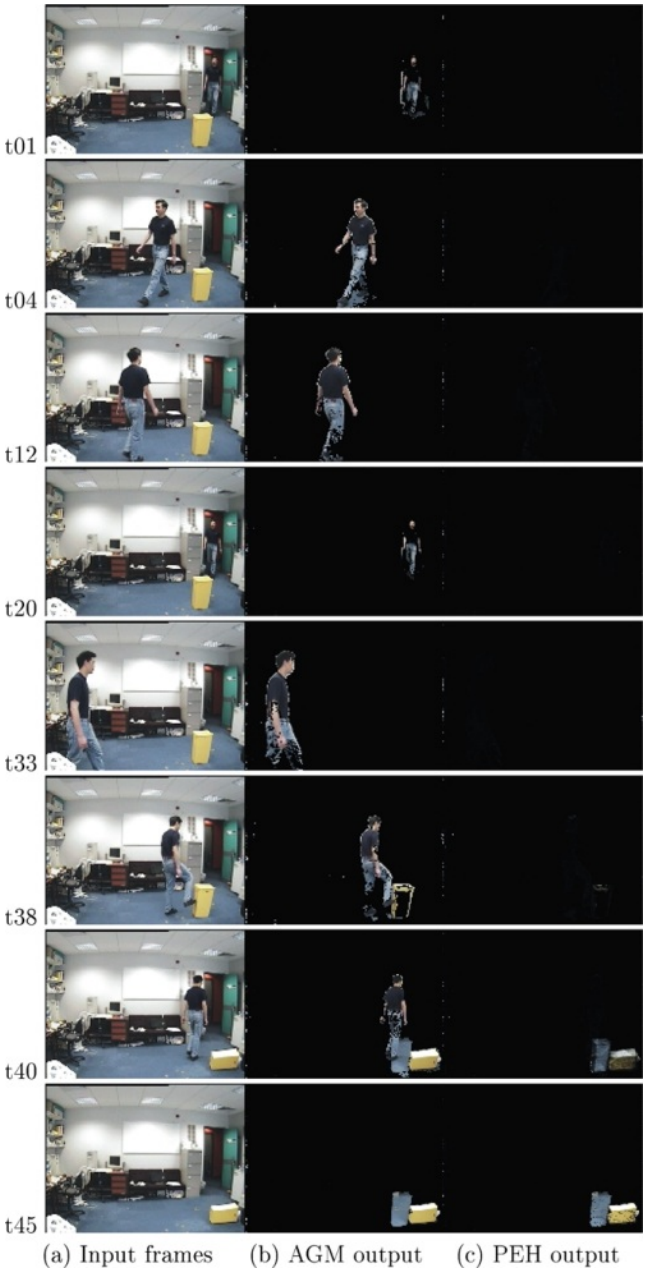

Fig. 10. Detecting the *falling box* event.

Table 2
Comparative results of the four tracking methods. Computation times are shown for a PII-330 MHz and are approximate.

Method	Incorrect frames		Time per frame (ms)
	Number	%	
Semantics-based	439	13	200
CONDENSATION-based	602	18	24,100
Dynamical	728	22	40
Non-contextual	995	30	50

discontinuous motion more compelling, we captured all video data at a relatively high frame rate of 18 fps and used off-line processing.

First, we show some results from applying a BBN based semantics model to perform hands movement association. Example frames from four different video sequences consisting of 141– 367 frames per sequence are shown in Fig. 11. Each sub-figure shows frames from one sequence temporally ordered from left to right, top to bottom. It is important to note that the frames are not consecutive. In each image a box frames the head and each of the two hands. The hand boxes are labelled left and right, showing the correct assignments. In the first example, Fig. 11(a), the hands were accurately interpreted before, during and after mutual occlusion. In Fig. 11(b), typical coughing and nose-scratching movements bring about occlusion of the head by a single hand. In this sequence the two frames marked with 'A' are adjacent frames, exhibiting the significant motion discontinuity that was encountered. Nevertheless the BBN semantics model was able to correctly interpret the hands. In Fig. 11(c) the subject undergoes significant whole body motion to ensure that the model works while the head is constantly moving. With the hands alternately occluding each other and the face in a tumbling action, the model is still able to correctly interpret the body parts. In the third-to-last frame both hands simultaneously occlude the face. The example of Fig. 11(d) has the subject partially leaving the screen twice to fetch and then offer a book. Note that in the frames marked 'M' one hand is not visible in the image. Since, this case is not explicitly modelled by the BBN semantics model, occlusion with the head or the other hand is deduced. After these periods of disappearance, the hand is once again correctly interpreted.

Second, we compared the atemporal semantics-based interpreter experimentally with three other dynamics-based tracking methods:

Dynamical: assuming temporal continuity exists between frames over time and linear dynamics, this method uses Kalman filters for each body part to match boxes at the pixel level between frames.
Non-contextual: similar to the semantics-based method, this method assumes temporal continuity but does not attempt to model the dynamics of the body parts. The method matches skin clusters based only on spatial association without the use of high-level knowledge.
CONDENSATION-based: based on the approach taken in [33], this method uses CONDENSATION with simple dynamics to track the hands simultaneously as a joint state, and employs an exclusion principle for the case of occlusion. More details can be found in [45].

It is difficult to compare the tracking methods fairly in this context. Comparison of the average deviation from the true hand and head positions would be misleading because

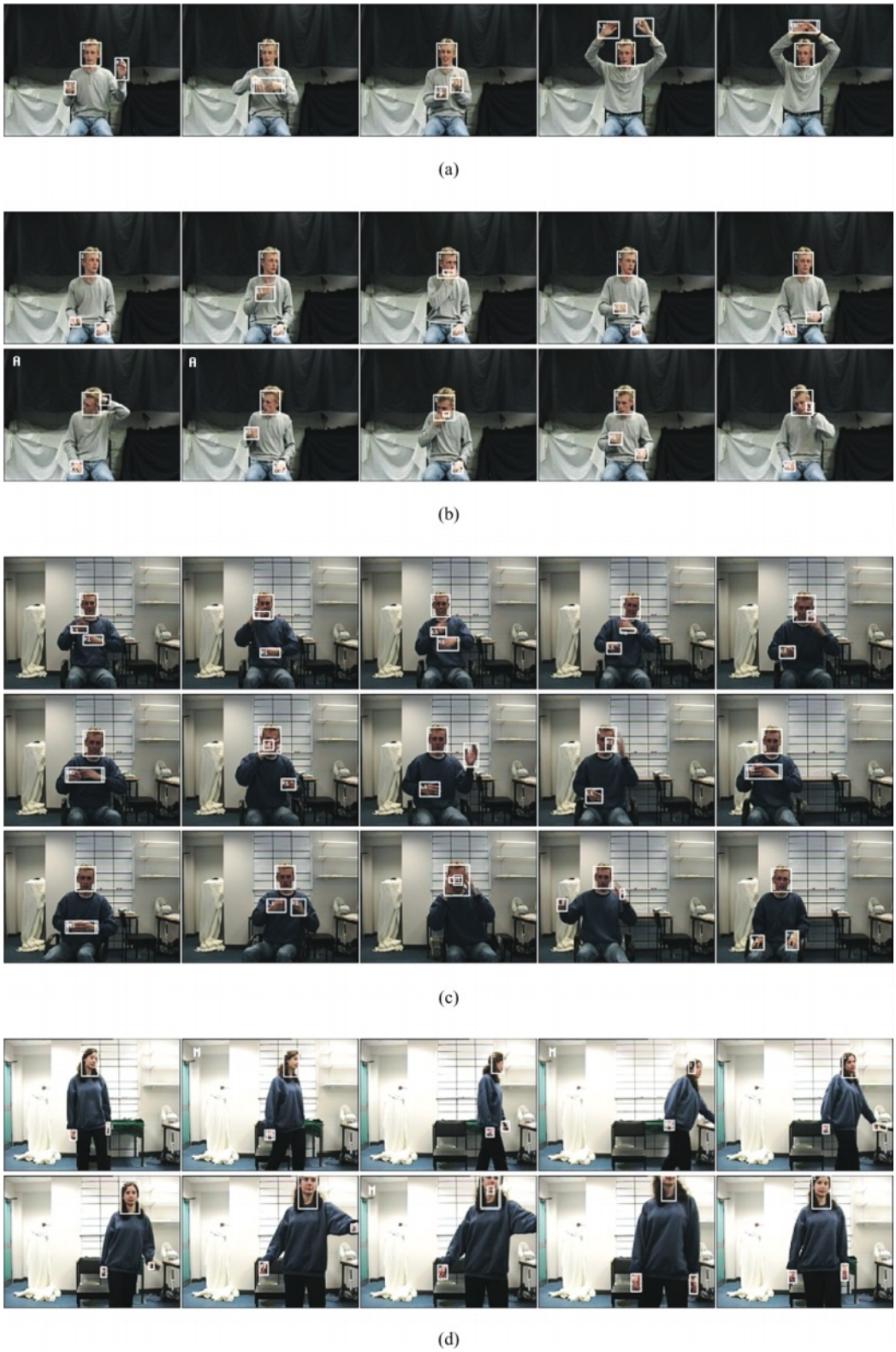

Fig. 11. Examples of discontinuous motion tracking.

of the all-or-nothing nature of matching to discrete clusters. Another possible criterion is the number of frames until loss-of-track, but this is somewhat unfair since a tracker may lose lock at the start of the sequence and then regain it and perform well for the rest of the sequence. The criterion we chose for comparison is the total number of frames on which at least one body part was incorrectly tracked, or the hands were mismatched. The comparison was performed on

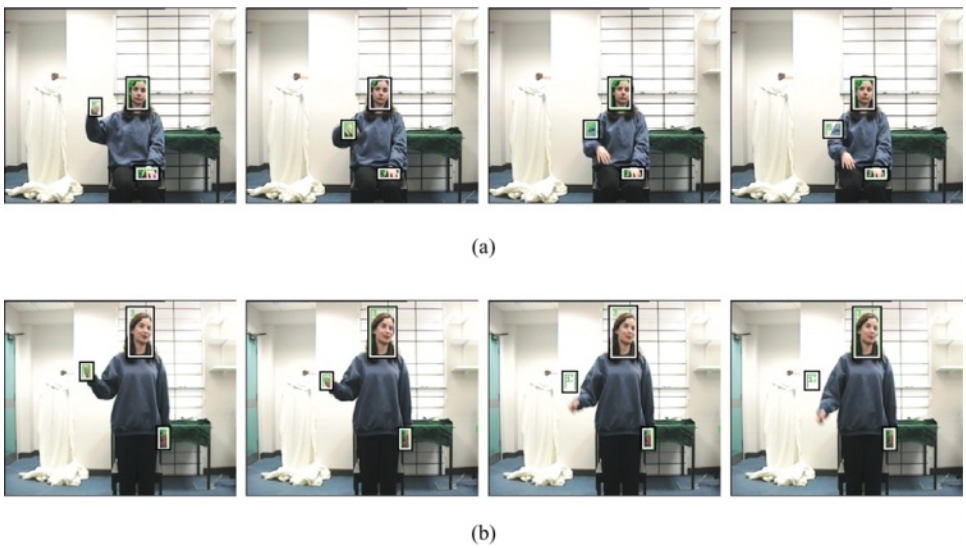

Fig. 12. Two examples of the failure of the dynamic Kalman filter tracker.

14 sequences containing two different people totalling 3300 frames.

Table 2 shows the number of frames incorrectly tracked by each method, in absolute terms and as a percentage of the total number of frames, and the approximate computation time per frame in milliseconds on a Pentium II 330 MHz computer. The semantics-based interpreter performs significantly better than the other methods, even though the data was captured at a high frame rate. Therefore, the benefits of modelling semantics rather than temporal continuity for performing association of visual observation of hands under discontinuous motion are significant. One would expect even better improvements if only low frame-rate data were available. The most common failure modes for the non-contextual-based, CONDENSATION-based and dynamical-based trackers were incorrect assignment of the left and right hands to clusters, and locking on to background noise when one hand was occluded. The dynamics-based tracker often failed due to inaccurate temporal prediction of the hand position. Two examples of this failure are shown in consecutive frames in Fig. 12. Although one could use more sophisticated dynamical models, it is still very unlikely they will ever be able to feasibly capture the full gamut of human behaviour, let alone accurately predict under heavily discontinuous motion. For example, the body-parts tracker in [58] switches in appropriate high-level models of behaviour for improved tracking, but the computational cost increases with the number of possible behaviours modelled.

Regarding computational cost, the table shows that the CONDENSATION-based method is two orders of magnitude more expensive than the semantics-based method. The enormous increase in computational expense was mainly due to the re-use of observations in statistical sampling, in particular the local hand orientations, which require an expensive filtering operation. This highlights the important computational advantage of the Bayesian network approach: an enormous state space can be fully modelled using efficient computation, while the resources required for particle filtering methods such as CONDENSATION grow exponentially with the state space size and are largely out of the designer's control.

6. Conclusion

Modelling behaviours and recognising events often require object-level representations to interpret visual data. However, object segmentation and trajectory extraction rely upon spatial proximity (region-growing) and temporally constrained 'blob' correlations (linear or second-order dynamics) respectively. Using such assumptions to interpret complex visual phenomena in busy scenes might not be sufficient. Indeed, it is questionable how much an automated system can learn and perceive objects from single image frames without pre-learned object models.

We described an approach to learning semantics of scene context in order to interpret novel visual events without object segmentation and motion grouping. Pixel-energy histories provide a condensed variable-length representation of temporal fast change in single pixels. Our experiments show that they can be used to semantically discriminate motion caused by different types of scene background change and detect events of significance without segmentation and grouping. We have also used adaptive Gaussian Mixture Models to separately model and recognise slow change such as illumination cycles under a less computationally taxing framework. The ambiguity inherent in viewing a complex world through a single pixel has been addressed by incorporating the modelling of synchronous change in multiple pixels during events (i.e. co-occurrence)

to perform pixel-stream hypotheses and matching of fast change.

Visual observations of human body motion from interactive behaviour can often be jerky and discontinuous. Semantics of the underlying behaviour and context can be used to overcome ambiguities and uncertainties in measurement. We presented a method to model the semantics of human body configuration using Bayesian belief networks. The model is used to perform robust human body parts association under discontinuous motion from a single 2D view. Rather than modelling spatio-temporal dynamics, the problem of visual tracking is addressed by reasoning about the observations using a semantics-based inference model. This semantics-based interpreter was tested and compared with more traditional dynamical and non-contextual trackers. The results indicate that modelling semantics significantly improves the robustness and consistency of tracking and associating visual observations under uncertainty and discontinuity.

References

[1] Y. Azoz, L. Devi, R. Sharma, Tracking hand dynamics in unconstrained environments, IEEE International Conference on Automatic Face and Gesture Recognition, Japan (1998) 247–279.

[2] M.J. Black, A.D. Jepson, A probabilistic framework for matching temporal trajectories: condensation-based recognition of gestures and expressions, European Conference on Computer Vision, Freiburg, Germany (1998) 909–924.

[3] A. Blake, M. Isard, Active Contours, Springer, Berlin, 1998.

[4] A. Bobick, A.D. Wilson, A state-based approach to the representation and recognition of gesture, IEEE Transactions on Pattern Analysis and Machine Intelligence 19 (12) (1997) 1325–1337.

[5] M. Brand, V. Kettnaker, Discovery and segmentation of activities in video, IEEE Transactions on Pattern Analysis and Machine Intelligence 22 (8) (2000) 844–851.

[6] M. Brand, N. Oliver, A. Pentland, Coupled hidden Markov models for complex action recognition, Computer Vision and Pattern Recognition, Puerto Rico, June (1997) 994–999.

[7] H. Buxton, S. Gong, Visual surveillance in a dynamic and uncertain world, Artificial Intelligence 78 (1995) 431–459.

[8] T. Cham, J. Rehg, Dynamic feature ordering for efficient registration, IEEE International Conference on Computer Vision, Corfu, Greece, September 2 (1999) 1084–1091.

[9] E. Charniak, Bayesian networks without tears, AI Magazine 12 (4) (1991) 50–63.

[10] O. Chomat, J. Crowley, A probabilistic sensor for the perception and the recognition of activities, European Conference on Computer Vision, Dublin, Ireland, September (2000).

[11] R.G. Cowell, A.P. Dawid, S.L. Lauritzen, D.J. Spiegelhalter, Probabilistic Networks and Expert Systems, Springer, New York, 1999.

[12] W.T. Freeman, E.H. Adelson, The design and use of steerable filters, IEEE Transactions on Pattern Analysis and Machine Intelligence 13 (9) (1991) 891–906.

[13] A. Galata, N. Johnson, D. Hogg, Learning structured behaviour models using variable length Markov models, IEEE International Workshop on Modelling People, Corfu, Greece, September (1999) 95–102.

[14] P.G. Gärdenfors, The dynamics of belief systems. Foundations vs. coherence theories, Revue Internationale de Philosophic 172 (1990) 24–46.

[15] D.M. Gavrila, The visual analysis of human movement: a survey, Computer Vision and Image Understanding 73 (1) (1999).

[16] D.M. Gavrila, L.S. Davis, 3-D model-based tracking of human motion in action, Computer Vision and Pattern Recognition (1996) 73–80.

[17] S. Gong, H. Buxton, On the expectations of moving objects, European Conference on Artificial Intelligence, Vienna, Austria, August (1992) 781–785.

[18] S. Gong, H. Buxton, Bayesian nets for mapping contextual knowledge to computational constraints in motion segmentation and tracking, British Machine Vision Conference, Guildford, UK, September (1993) 229–239.

[19] S. Gong, M. Walter, A. Psarrou, Recognition of temporal structures: learning prior and propagating observation augmented densities via hidden Markov states, IEEE International Conference on Computer Vision, Corfu, Greece, September 1 (1999) 157–162.

[20] I. Haritaoglu, D. Harwood, L.S. Davis, W^4: Real-time surveillance of people and their activities, IEEE Transactions on Pattern Analysis and Machine Intelligence 22 (8) (2000) 809–830. August.

[21] D.J. Heeger, Optical flow from spatiotemporal filters, IEEE International Conference on Computer Vision, London, UK, June (1987) 181–190.

[22] K. Imagawa, S. Lu, S. Igi, Color-based hands tracking system for sign language recognition, IEEE International Conference on Automatic Face and Gesture Recognition, Nara, Japan (1998) 462–467.

[23] S. Intille, A. Bobick, Representatio, ECCV Workshop on Perception of Human Action, Freiburg, Germany, June (1998).

[24] M. Isard, A. Blake, Contour tracking by stochastic propagation of conditional density, European Conference on Computer Vision, Cambridge, UK (1996) 343–356.

[25] M. Isard, A. Blake, CONDENSATION—conditional density propagation for visual tracking, International Journal on Computer Vision 29 (1) (1998) 5–28.

[26] Y.A. Ivanov, A. Bobick, Recognition of visual activities and interactions by stochastic parsing, IEEE Transactions on Pattern Analysis and Machine Intelligence 22 (8) (2000) 852–872.

[27] C. Jennings, Robust finger tracking with multiple cameras, IEEE International Workshop Recognition, Analysis and Tracking of Faces and Gestures in Real-Time Systems, Corfu, Greece, September (1999) 152–160.

[28] N. Johnson, Learning Object Behaviour Models, PhD thesis, University of Leeds, Leeds, UK, 1998.

[29] N. Johnson, D. Hogg, Learning the distribution of object trajectories for event recognition, Image and Vision Computing 14 (8) (1996) 609–615.

[30] N. Jojic, M. Turk, T. Huang, Tracking self-occluding articulated objects in dense disparity maps, IEEE International Conference on Computer Vision, Corfu, Greece, September 1 (1999) 123–130.

[31] P. Kovesi, Image correlation from local frequency information, The Australian Pattern Recognition Society Conference, Brisbane, December (1995) 336–341.

[32] S. Lauritzen, D. Spiegelhalter, Local computations with probabilities on graphical structures and their application to expert systems, in: G. Shafer, J. Pearl (Eds.), Readings in Uncertain Reasoning, Morgan Kaufmann, Las Altos, CA, 1990, pp. 415–448.

[33] J. MacCormick, A. Blake, A probabilistic exclusion principle for tracking multiple objects, IEEE International Conference on Computer Vision, Corfu, Greece, September 1 (1999) 572–578.

[34] J. Martin, V. Devin, J. Crowley, Active hand tracking, IEEE International Conference on Automatic Face and Gesture Recognition, Nara, Japan (1998) 573–578.

[35] S.J. McKenna, S. Gong, Gesture recognition for visually mediated interaction using probabilistic event trajectories, British Machine Vision Conference, Southampton, England, September (1998) 498–507.

[36] S.J. McKenna, S. Jabri, Z. Duric, H. Wechsler, Tracking interacting

people, IEEE International Conference on Automatic Face and Gesture Recognition, Grenoble, France, March (2000) 348–353.
[37] D. Metaxas, Deformable model and HMM-based tracking, analysis and recognition of gestures and faces, IEEE International Workshop Recognition, Analysis and Tracking of Faces and Gestures in Real-Time Systems, Corfu, Greece, September (1999) 136–140.
[38] D. Moore, I. Essa, M. Hayes III, Exploiting human actions and object context for recognition tasks, IEEE International Conference on Computer Vision, Corfu, Greece, September 1 (1999) 80–86.
[39] N.M. Oliver, B. Rosario, A. Pentland, A Bayesian computer vision system for modelling human interactions, IEEE Transactions on Pattern Analysis and Machine Intelligence 22 (8) (2000) 831–843.
[40] E.-J. Ong, S. Gong, A dynamic human model using hybrid 2D–3D representations in hierarchical PCA space, British Machine Vision Conference, Nottingham, UK, September 1 (1999) 33–42. BMVA.
[41] A.V. Oppenheim, R.W. Schafer, Digital Signal Processing, Prentice-Hall, Englewood Cliffs, NJ, 1975.
[42] PAMI, Special issue on video surveillance, IEEE Transactions on Pattern Analysis and Machine Intelligence, August 22 (2000).
[43] J. Pearl, Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference, Morgan Kaufmann, San Mateo, CA, 1988.
[44] Y. Raja, S.J. McKenna, S. Gong, Tracking and segmenting people in varying lighting conditions using colour, IEEE International Conference on Automatic Face and Gesture Recognition, Nara, Japan, April (1998) 228–233.
[45] J. Sherrah, S. Gong, Resolving visual uncertainty and occlusion through probabilistic reasoning, British Machine Vision Conference, Bristol, UK, September 1 (2000) 252–261. BMVA.
[46] J. Sherrah, S. Gong, Tracking discontinuous motion using Bayesian inference, European Conference on Computer Vision, Dublin, Ireland, June 2 (2000) 150–166.
[47] J. Sherrah, S. Gong, Continuous global evidence-based Bayesian modality fusion for simultaneous tracking of multiple objects, IEEE International Conference on Computer Vision, Vancouver, Canada, July (2001).
[48] C. Stauffer, W.E.L. Grimson, Using adaptive tracking to classify and monitor activities in a site, Computer Vision and Pattern Recognition, Los Alamitos, USA (1998) 22–29.
[49] C. Stauffer, W.E.L. Grimson, Adaptive background mixture models for real-time tracking, Computer Vision and Pattern Recognition, Colorado, USA, June (1999) 246–252.
[50] G.D. Sullivan, Visual interpretation of known objects in constrained scenes, Philosophical Transactions of Royal Society of London B 337 (1992) 361–370.
[51] M. Usher, N. Donnelly, Visual synchrony affects binding and segmentation in perception, Letter to Nature 394 (1998) 179–182.
[52] T. Wada, T. Matsuyama, Multiobject behaviour recognition by event driven selective attention method, IEEE Transactions on Pattern Analysis and Machine Intelligence 22 (8) (2000) 873–887.
[53] M. Walter, A. Psarrou, S. Gong, An incremental approach towards automatic model acquisition for human gesture recognition, IEEE Workshop on Human Motion, December (2000).
[54] M. Walter, A. Psarrou, S. Gong, Auto-clustering for unsupervised learning of atomic gesture components using Minimum Description Length, IEEE International Workshop on Recognition, Analysis and Tracking of Faces and Gestures in Real-time Systems, Vancouver, Canada, July (2001).
[55] H.L. Weinberg, Levels of knowing and existence: Studies in general semantics, Harper and Brothers, New York, 1959.
[56] C. Wern, A. Azarbayejani, T. Darrell, A. Pentland, Pfinder: Real-time tracking of the human body, IEEE Transactions on Pattern Analysis and Machine Intelligence 19 (1) (1997) 780–785.
[57] C. Wern, A. Pentland, Dynamic models of human motion, IEEE International Conference on Automatic Face and Gesture Recognition, Nara, Japan (1998) 22–27.
[58] C. Wern, A. Pentland, Understanding purposeful human motion, IEEE International Workshop on Modelling People, Corfu, Greece (1999) 19–25.