ELSEVIER

Image and Vision Computing 21 (2003) 125–136

image AND vision COMPUTING

www.elsevier.com/locate/imavis

Learning and understanding dynamic scene activity: a review

Hilary Buxton*

School of Cognitive and Computing Sciences, University of Sussex, Falmer, Brighton BN1 9QH, UK

Abstract

We are entering an era of more intelligent cognitive vision systems. Such systems can analyse activity in dynamic scenes to compute conceptual descriptions from motion trajectories of moving people and the objects they interact with. Here we review progress in the development of flexible, generative models that can explain visual input as a combination of hidden variables and can adapt to new types of input. Such models are particularly appropriate for the tasks posed by cognitive vision as they incorporate learning as well as having sufficient structure to represent a general class of problems. In addition, generative models explain all aspects of the input rather than attempting to ignore irrelevant sources of variation as in exemplar-based learning. Applications of these models in visual interaction for education, smart rooms and cars, as well as surveillance systems is also briefly reviewed.

Keywords: Cognitive computer vision; Generative model; Visual reasoning; Visual control; Visual learning

1. Introduction

How can we develop *visual intelligence*? What are the basic research issues when we go from live video input to meaningful behaviour in our vision systems? We know that the gap cannot be closed by structural computer vision techniques alone. Such techniques are concerned with direct space-time aspects of the physical world and reconstruction of the observed scene. However, we are concerned with system purpose to integrate our vision modules into complete intelligent systems, which can understand observed activities as well as track the objects and people in the scene. Ballard's landmark paper [2] proposes that vision is best understood in the context of the visual behaviours engaging the system without requiring detailed internal representations of the scene. His approach also recognises that it is important to have a system framework that integrates visual processing within the task context. In general, it is not necessary to have perfect object reconstruction before we achieve the visual understanding required for real-world applications.

What do cognitive vision systems entail? The basic approaches combine techniques from symbolic or subsymbolic AI with computer vision techniques in some way. Naturally, we then encounter many of the major issues in AI such as knowledge representation and reasoning, control and the handling of uncertainty, as well as machine learning. Much of the work assumes that knowledge drives reasoning in visual interpretation (*seeing as*), thus visual context is seen as essential for understanding what is depicted in images or image sequences. If we are to build efficient systems that can tackle many different tasks, high-level attention and control (*seeing for*) is also seen as essential. In addition, if we are to incorporate scene and task knowledge, we have to address the question of how such knowledge can be acquired. Knowledge structures were traditionally designed by hand but here we argue for the importance of automating their construction (*learning*) since learnt models are well suited to on-line analysis.

Learning in a vision system can be at the level of object models, their movements and actions, and how to control views and processing in the system. Some kinds of knowledge can be hand-crafted, for example, *explicit models* designed to provide a solution to one particular application. Such models are necessarily fixed after construction, with no flexibility to adapt to new kinds of data. A typical explicit model is given in early work by Hogg [33] on an articulated cylinder model of a walking man with spatio-temporal constraints on the movement patterns. This was advanced research for model-based vision in 1983, however, such models are slow to match to the image data and cannot adapt in the on-line system. Even with more recent work where aspects of the model are learned from training data [19], the fact that the model is

* Tel.: +44-1273-678569; fax: +44-1273-671320.
E-mail address: hilaryb@cogs.susx.ac.uk (H. Buxton).

0262-8856/03/$ - see front matter © 2002 Elsevier Science B.V. All rights reserved.
PII: S0262-8856(02)00127-0

explicit means that human intervention is required for labeling the examples.

An alternative approach has been developed more recently with *exemplar-based models* that have an associated distance metric. For this kind of model, it is necessary to remove irrelevant variation by preprocessing to derive the set of exemplars. Typical exemplar-based techniques include nearest-neighbour methods and Roweis and Saul's local linear embedding [76], which yields non-linear dimensionality reduction in the learnt model. While such models are fast to match using the distance metric and can adapt, they are still crucially dependent on choice of preprocessing for the application. This involves human intervention in deciding what is irrelevant, thus leaving only relevant variables or parameters to be learned.

Exciting new work on fully automated learning of *flexible models*, such as the partially learnt deformable and graphical models reviewed here, could be extended. Although learning of the underlying hidden variables as well as the data relationships has so far only been applied to more basic vision tasks such as segmentation, digit recognition, tracking and 'sprite' modeling [25], the approach is general and could be taken further. Flexible models have evolved in the machine learning community and cover a wide class of parametric models for signal processing, which includes PCA and Gaussian mixtures (GM) as well as hidden Markov models (HMM) and Bayesian belief networks (BBN), etc. Here we are interested in those that are 'generative', that is, with probability distributions, estimated from input data, over a set of hidden variables. A clear advantage of such models is that some knowledge of the complexity of the problem, such as number of dimensions of the hidden space, can be built in. These generative models can then adapt to the different types of input required by the application as well as predict and explain the data.

2. Models: reasoning and learning

Essentially, what is proposed here is that a powerful way of tackling these issues of learning and inference as well as the use of context and active task control is to use generative models. In cognitive vision, there are two highly related approaches to generative models which have been developed, deformable and graphical models. These are to a large extent complementary as deformable models have been primarily used to represent static and dynamic shape, texture and other physically observable parameters, while more general graphical models have been used to capture more abstract relationships. Both use well-established statistical learning theory in either off-line or on-line learning. A unifying review of theory and techniques for such models from the machine learning perspective is given in Roweis and Ghahramani [75]. They explain the relationships between the models and even offer a generative model for generative models. Table 1 gives an overview of how the flexible, generative models in this review can be regarded as extensions of each other, giving increasing capability at the cost of greater complexity.

The Gaussian is the initial underlying measurement model in automated learning of more structured generative models, where variability in the data can be explained (or predicted) from inferred (hidden) model parameters. HMM models can be regarded as extending GM models by having learned dynamic dependencies between states. These have a chain of simple dependencies on the immediately previous state. They can be extended to either 1) coupled dependencies with states in another HMM to form a Coupled Hidden Markov Model (CHMM) or 2) possible longer term temporal dependencies with previous states in the same HMM to form a variable length Markov model (VLMM). PCA models are commonly used to characterise data by a reduced set of dimensions or model parameters. BBN and DBN models assume statistical independence plus a hierarchy of dependencies between the hidden model variables. We can regard a DBN as an extension of an HMM with hierarchy or an extension of a BBN with learnt dynamic dependencies between states of some kind, e.g. with a simple Markov relationship over time. Adding utility theory allows decision support for rational agents using utility nodes and decision nodes for actions to form a dynamic decision network (DDN) of some kind, although this extension is usually hand-coded.

PCA is the commonest way of modeling data by using a linear (Karhunen–Loeve) transformation to find a reduced number of 'effective' features that retain most of the intrinsic information. That is, we have m-dimensional vector $\mathbf{x}$ and want to model this using l features or parameters, so we want to find transformation $\mathbf{Tx}$ such that truncation causes as little increase in mean-square error as possible. This can be achieved by simple Hebbian learning or a variety of other techniques. The transformation yields a set of orthogonal eigenvectors, which are the parameters of interest for the representation. The different combinations of eigenvalues then represent particular instances of the parameterised, generative model. In Section 3, the use of these techniques in the deformable models approach is described, followed in Section 4 by extensions

Table 1
Generative model relationships

Initial	Extension	Final
Gaussian	Mixture	GM
Gaussian	Reduce dimension	PCA
GM	Dynamic	HMM
HMM	Coupling	CHMM
HMM	Variable length	VLMM
HMM	Hierarchy	DBN
DBN	Utility	DDN

to more general reasoning, control and learning for more complex graphical models.

3. Deformable models

Original work by Terzopoulos [80] pioneered the use of deformable models in realistic simulations of human heads using biomechanical facial muscles and skin movements. This work led to early methods for shape and non-rigid motion estimation going from synthesis to an interpretive task [55]. An alternative approach was developed by Edwards and colleagues, based on work by Cootes and Taylor [19]. This simply models the appearance of the face in 2D with parameterised texture variations and 2D deformations to simulate different expressions and viewpoints [22] rather than model the underlying 3D structure. Such models can provide good graphical representations for synthesis of appearance and motion, but reversing this for interpretation is less well understood. However, further work has shown we can both learn adaptive shape models for humans and use them in finding likely spatio-temporal patterns of activity for tracking, as in the early work by Baumberg and Hogg [4,6]

So what is a deformable model and how is it used in this visual interpretation work? Physically based vibration modes are fundamental for the description of such models, and building in known constraints into the perception process can allow difficult problems in computer vision to be solved. The constraints are mainly learned from observation for visual interpretation, e.g. Baumberg and Hogg [4]. However, in general, these constraints need not even be based on real physical properties but just act to condition the interpretation. For tracking a non-rigid deformable object, early work by Pentland and Horowitz [65] showed how to recover motion and structure using finite element techniques. This method relies on assumptions about elasticity and density distribution, with vibration modes derived via mass and stiffness matrices in the main equation governing the physical behaviour. Further work by Nastar [58] used such 'modal analysis' in a wider range of applications. However, it is the combination of acquisition of these modes of variation by observation and their use in generative perceptual control that has proved crucial for more general application in vision.

The use of training in deformable model analysis has much in common with training in neural networks. Cootes and Taylor [20], in their early work, introduced the point distribution model (PDM) which is derived from a set of characteristic training data for the problem at hand and parameterised by a set of orthogonal 'modes of variation'. The training shapes are then represented by a subset of vectors which account for the majority of the observed variations. The PDM has proved extremely useful for image sequence analysis for tracking contours [4] and locating structures in medical images [32]. However, there was originally no intrinsic time dimension and, in early work, characteristic points for the training sets were selected by hand. More recent work by Baumberg and Hogg [6] successfully tackled both these drawbacks and has led to the development of not only tracking of walking people but also generalisation of the techniques using probability density functions (PDFs) on extracted trajectories. These are used in event analysis (e.g. Johnson and Hogg [45]) and the full visual interaction modeling we describe below, which uses extended statistical learning techniques.

3.1. Point distribution model

PDMs are a well-established tool for statistical analysis of visual data. The PDM is based on a set of example shapes of a given object (or behaviour pattern if temporal extension). Traditionally each shape is defined by landmark points that are selected as important, corresponding to features of the object (or trajectory). This approach then allows a class of objects (or behavior patterns) to be characterised by a small set of parameters. The original linear PDM [20] was trained on a set of shapes with n landmark points aligned to the mean shape $\bar{\mathbf{x}}$. The distance from the mean is calculated using PCA to get the vector $\mathbf{d} = \mathbf{x} - \bar{\mathbf{x}}$ for the n 2D landmark vectors $\mathbf{x} = (x_1, y_1, \ldots, x_n, y_n)$. Next the $2n \times 2n$ covariance matrix $C = E(\mathbf{d} \cdot \mathbf{d}^{\mathrm{T}})$ is found, where E is the expectation operator over the training examples. Eigenvectors of C then correspond to the variation modes for this data with the largest eigenvalue describing the most significant mode, etc. The PDM model is simply the mean shape $\mathbf{x}$, together with the t eigenvectors $\mathbf{p}_1, \ldots, \mathbf{p}_t$ characterising a number of modes of variation sufficient to describe the training data for the task at hand.

The PDM can be extended to parameterised curves such as the cubic B-spline [45]. They show that the point measurements can be mapped to the modes of variation using $\mathbf{b} = P^{\mathrm{T}}(\mathbf{x} - \bar{\mathbf{x}})$, where $\mathbf{b}$ is a vector of t shape parameters, $\mathbf{x}$ is a vector of $2n$ positions and P is a matrix with columns corresponding to eigenvectors of the covariance matrix. Using a set of parameterised shapes for training the model, represented by n control points, allows the automated placing of a set of measurements at regular intervals along a continuous curve. However, in contrast to landmark points set by hand, any particular measurement will affect several control points, which violates the independence assumption implicit in the measurement model (Gaussian sampling noise with variance r). The formal solution to this problem is to use the covariance matrix $R = r\mathscr{H}^{-1}$, obtained using the interpolation function H_i for the parameter [9]

$$\mathscr{H}_{ij} = \int H_i(u) H_j(u) \mathrm{d}u.$$

In practice a covariance matrix S is estimated from training data to obtain $2n$ eigenshapes using the relationship $S\mathscr{H}\mathbf{e}_i = \lambda_i \mathbf{e}_i$. The eigenvectors $\mathbf{e}_i$ are made orthonormal to

the inner product $\langle \mathbf{u}, \mathbf{u} \rangle = \mathbf{u}^{\mathrm{T}} \mathscr{H} \mathbf{u}$. The shape vectors are defined by $\mathbf{u} = \sum_{i=1}^{2n} \mu_i + \bar{\mathbf{u}}$, with coefficient $\mu_i = \langle \mathbf{u} - \bar{u}, \mathbf{e}_i \rangle$. Over the training set these coefficients can then be shown to be linearly independent and allow analysis similar to that in the traditional PDM. That is, the coefficients for each eigenshape mode are updated independently and a subset t can be used sufficient for the task.

3.2. Tracking objects

Deformable shape models, such as the PDM, are derived from a set of training data containing representative objects. The boundary of each segmented object is traced automatically and fitted with a cubic B-spline contour with consistent control points or fitted with an active 'snake' [48]. As we have seen, using the PDM approach above requires PCA on the spline data so that a small set of the largest components characterise the data. This model can be re-parameterised to make it even more compact for on-line tracking. Kalman filters can then be used to track objects using position, scale and orientation of the model in addition to the compact shape parameters. The iterated Kalman filter is used by Baumberg [3] to adaptively set these parameters for rapid convergence in the tracking. Using the ground-plane constraint allows the 2D height and position parameters to be mapped into approximate world coordinates. This approach was developed for deformable car models by Sullivan and colleagues [79] and brought together for both humans and cars in a joint project [71]. In the combined system, cars can stop moving and be incorporated into the background by using a median filter over time, thus allowing the detection of moving objects in front of these stationary vehicles.

3.3. Tracking humans

A promising scheme using PDMs has been mentioned above and tested for tracking human figures in image sequences, starting with the early work of Baumberg and Hogg [5,6]. The flexible 2D prior model of the human outline shape can be used to recognise and track pedestrians in dynamic scenes. This uses a learnt model in contrast to the explicit, hand-built 3D articulated model used in human tracking by Hogg in 1983 [33]. The deformable shape model incorporates the observed modes of variation as constraints, so that tracking becomes less sensitive to partial occlusions and image noise. This model captures the apparent changes in shape found in images due to underlying human pose and deformation variation, as well as relative viewpoint variation of the walking human with respect to the camera in the 3D scene. Wren and Pentland have also built deformable models with a region-based approach in which moving 'blobs' are tracked in real-time, used in their 'Pfinder' system [86]. Further work by Wren allows more detailed analysis of human body movements which incorporates some real physical constraints due to the skeletal structure [87]. This is closer to the work using biomimetic models of Terzopoulos and colleagues cited above, which incorporates models of the underlying muscle structure and elasticity.

3.4. Behaviour analysis

The main systems here to illustrate the deformable models approach are taken from the work of Johnson and Hogg [45], Morris and Hogg [56] and Galata et al., [27]. Their work emphasises acquisition and encoding of spatial, temporal and procedural knowledge by passive observation of video sequences of typical interactions. The basis for the work is development of spatio-temporal models automatically learned from video training data. The models have no explicit or hand-crafted element and are usually low dimensional linear transformations of the image space. These visual models of interaction have been applied to both surveillance applications and human computer interaction.

A basic component of these systems is representation of a given target activity by training a model with the video sequences depicting repeated examples of this target activity. The sequences must contain examples that span the range of ways in which the activity may be carried out and there will typically need to be hundreds of these examples. If the activity is on the ground-plane, such as pedestrians crossing a pathway or carpark, the constructed models must be sufficiently detailed to evaluate whether the pedestrian is deviating from the model and to allow prediction of the immediately following trajectory, up to some choice-point or junction for example. This kind of model, then, would be of use in surveillance systems to raise alarms and provide short-term occlusion handling in tracking.

3.4.1. Visual tracking

The *visual tracker* of Baumberg and Hogg [5,6] is used in analysis of this kind and closely related to the active shape models of Cootes and Taylor [19] and active contours of Blake and Isard [10]. A closed B-spline contour is used to represent the image profile and consists of a set of control points. There is initial segmentation using image differencing to extract the moving objects for training, then the B-spline contour is wrapped around each extracted profile so that the first control point is located in a similar position in each example for training. The number and spacing of the control points is also held constant so that variations in control point location from example to example are due to observable, intrinsic shape variations. An eigenshape model then consists of the largest principal components that account for these variations. For example, about 10–20 components were sufficient for the walking pedestrian, with the most salient feature being the gap appearing and disappearing between the legs. Tracking using this kind of model can provide a set of example trajectories by using a Kalman filter to constrain the search from frame to frame.

For example, in Fig. 1 trajectory from such tracking is shown ready for further processing. For more details see Johnson and Hogg [45], who demonstrate its use in assessing the typicality of a given trajectory from the PDF.

3.4.2. Visual interaction

The *visual interaction* of Morris and Hogg [56] builds on this work to allow the capability of recognising, for example, 'suspicious' behaviour. Trajectory landmarks are found using closest approach to the nearest object and the distribution of speed and distance with respect to these landmarks, called an 'event', is used to characterise the underlying typical behaviour during the training phase. In the recognition phase, the probability of a given event is assessed using simple statistical tests based on observing the measured speed and distance, where low speed and distance are associated with suspicious behaviour. That is, moving quickly down a row of cars is not unusual nor is standing still a long way off but being both close and near stationary is noted as atypical (low probability). The trajectories are first observed as an ordered sequence of events but are re-sorted in order of increasing probability for assessment. The last stage requires characterisation of cummulative probability for the overall trajectory under the expectation that people get out of their cars and walk to the exit. This value will usually start with low probability events like getting out of a car but will change if all other interactions are at a reasonable speed and distance. However, if atypical behaviour is taking place, there will be a large number of low probability events and the cummulative value will change less. Supervised learning is used to provide the classification boundary to decide if the overall trajectory value is 'atypical'.

3.4.3. Virtual reality

The *virtual reality* systems of Johnson et al. [46] also build on this approach to model interaction between people for applications in human computer interaction. Here a probabilistic model of the joint behaviours is learned by observing sets of typical interactions off-line in the training phase, much as above, but using the 'eigenshape' B-spline profile as the base [5]. The contours of the two interacting people are then represented by the concatenated set of control points from the B-spline (third order) contours. This time the state vector uses separation and height of the left and right individuals together with the first derivative of these measures. From the data, a set of prototype states are again derived by vector quantisation and linked into a Markov chain to form the behaviour model. As in the surveillance work, the model can be used in a variety of ways in the on-line recognition and interpretation phase. For example, extrapolation forward in time from a tracked behaviour is obtained probabilistically from the model in order to generate a set of more or less likely continuations. The model can also fill in missing parts of the behaviour interaction using a Bayesian framework in a similar manner to representing and updating the *a posteriori* density used in the tracking algorithm [9]. In general, the set of plausible state hypotheses can be found from this density function, where the maximum represents the most likely hypothesis for the state of the interaction.

The main application of their work is in the synthesis of a virtual partner and this has been demonstrated for handshaking and turn-taking in speaker interaction. Further work in this direction by Galata et al. [27] has developed this approach with the deformable model combined with a graphical model, a VLMM. This can be applied to highly structured behaviour such as dance, aerobics and sign language. As we see in Section 4 on graphical models, there are many variants of using an HMM and this approach then is close to work coupling dynamic models with the Markov chain representing long-term constraints [66,74]. However, the extension here uses VLMMs to overcome problems in the iterative optimisation found in learning standard models, where there can be local maxima. The flexible models first encode sequences of about 20 ms and then the higher level sequences of atomic behaviours of about 1 s.

Fig. 1. Results from tracking using deformable model to show trajectory extracted for further analysis [45].

4. Graphical models

Probabilistic frameworks have much to offer in dealing with the pervasive problem of uncertainty in visual evidence, and allow full information integration as proposed by Pearl [63]. The representation of constraints in a Bayesian Belief Network, BBN or 'Bayes net', can be achieved by mapping into a graph structure so that the nodes represent concepts or parameters of interest and dependencies are given by causal links. Bayes nets can be learned and model dependencies for either static (BBN) or dynamic (DBN) domains, as well as incorporate decision theory in DDN. A simpler chained structure with single causal dependencies over time, the HMM, is often used for speech analysis [70] and has been extensively adapted for analysis of dynamic scenes and perceptual control as described below. In probabilistic reasoning [17], the likelihood of classes of objects or events is inferred by propagation of belief values in the light of changing evidence. Early work incorporating Bayes nets was developed by Levitt and Binford [50,51] to make model-based vision reliable, while remaining computationally tractable.

Bayes nets have been widely adopted in vision systems as they are applicable to all levels of processing, due to fast numerical updating in singly connected trees. There are techniques to decompose complex models and handle networks with multiple causes as well as learn the parameters for such networks [77]. Rimey and Brown introduced such techniques for active vision, with both control of camera movement and the selective processing required in task-based perceptual control [72,73]. As discussed by Gong and Buxton [30], Bayes nets provide a clear way to map contextual constraints from the scene onto the computation of the visual interpretation by combining known causal dependencies with estimated statistical knowledge. They are essentially providing closed-loop control using both top-down and bottom-up messages in the propagation of belief values. They also provide the possibility of learning and refining visual representations by observation [14,85]. Bayes nets have been used in many demanding applications such as BATmobile [24] and TEA system [73].

HMMs are also widely used in visual processing, as seen in the review of recent work on behaviour analysis below. The advantage here is that the 'hidden' purposes of regular behaviour patterns can be learned from examples, i.e. the structure of the model as well as the parameters are easily learned using the Baum-Welch (or EM) algorithm [70]. For more general Bayes nets, the dependency structure may be unknown, which complicates the learning process [31] but it can be solved by using 'structural' EM with local search in the M step [26]. Conditional probability learning for the state transitions in HMMs is straightforward, for example, in early work by Gong [28,29] the movement patterns of vehicles on an airport ground-plane were learned and provided a generative model used in prediction. More recent work uses *coupled* CHMMs to learn models for interactions [13], *parameterised* HMMs for gesture interpretation [12], and *variable length* VLMMs for virtual reality systems [27].

4.1. Belief propagation

BBN are directed acyclic graphs (DAG) in which each node represents an uncertain quantity using variables with multiple possible values. The arcs connecting the nodes signify the direct causal influences between the linked variables, with the strengths of such influences quantified by associated conditional probabilities. If we assume a variable in the network is X_i, and a selection of variables Π_{X_i} are the direct causes of X_i, the strengths of these direct influences are quantified by assigning the variable X_i a link matrix consisting of the values $P(x_i|\Pi_{X_i})$, given any combination of instantiations of the parent set Π_{X_i}. The conjunction of all the local link matrices of variables X_i in the network (for $1 \leq i \leq n$ where n is the total number of the variables) specifies a complete and consistent global model which provides answers to all the probabilistic queries. Such a conjunction is given by the overall joint distribution function over the variables $X_1, \ldots, X_n$

$$P(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} P(x_i|\Pi_{X_i}),$$

where lower case symbols stand for a particular instantiation of the corresponding variables.

In a belief network, if we quantify the degree of coherence between the expectations ($\mathbf{X}$) and the evidence ($\mathbf{e}$) by a measure of local belief $\mathrm{BEL}(x) = P(x|\mathbf{e})$, and define belief commitments as the tentative acceptance of a subset of hypotheses that together constitute a most satisfactory explanation of the evidence at hand. Then, Bayesian belief revision amounts to the updating of belief commitments by distributed local message passing operations. Instead of associating a belief measure with each individual hypothesis locally, belief revision identifies a composite set of hypotheses that best explains the evidence. We call such a set the most-probable-explanation (MPE). In computational terms, this means finding the most probable instantiations of all hypothetical variables given the observation.

Let $\mathbf{W}$ stand for the set of all the variables concerned, inclusive of those in $\mathbf{e}$. Any particular instantiation of variables in $\mathbf{W}$ that is also consistent with $\mathbf{e}$ will be regarded as an *extension* or *explanation* of $\mathbf{e}$. The problem then is to find an extension $\mathbf{w}^*$ that maximises the conditional probability $P(\mathbf{w}|\mathbf{e})$. In other words, $\mathbf{W} = \mathbf{w}^*$ is the MPE of the evidence if $P(\mathbf{w}^*|\mathbf{e}) = \max_{\mathbf{w}} P(\mathbf{w}|\mathbf{e})$. Here, $\mathbf{w}^*$ is obtained by (1) locally computing the belief function for each variable X mentioned above, i.e.

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

where $\mathbf{w}'_X = \mathbf{w}/x$, i.e. the set $\mathbf{w}'_X$ is equal to the set $\mathbf{w}$ minus the element x; (2) propagating local messages, where these

are defined as: if X has n parents $U_1, U_2, \ldots, U_n$ and m children $Y_1, Y_2, \ldots, Y_m$, then node X receives messages $\pi^*_X(u_i)$, $i = 1, \ldots, n$ from its parents and $\lambda^*_{Y_j}(x)$, $j = 1, \ldots, m$ from its children, where: $\pi^*_X(u_i)$ is the probability of the most probable tail-extension of the hypothetical value $U_i = u_i$ relative to the link $U_i \rightarrow X$ and is known as an *explanation*, $\lambda^*_{Y_j}(x)$ is the conditional probability of the most probable head-extension of the hypothetical value $X = x$ relative to the link $X \rightarrow Y_j$, known as a *forecast*.

4.2. Tracking objects

Blake and Isard learn to track visual contours for general applications [10], without using a graphical model but with the possibility of building constraints into the prior term to influence interpretation. However, it can be argued that for application-specific tracking, there should be modeling of dependencies and explicit association of entities over time [59]. More recently, Black and Fleet [8] have developed a full generative Bayesian framework for tracking motion boundaries. Buxton and Gong [14] have developed a systematic methodology for the design and integration of advanced vision systems using Bayes nets. These networks allow dynamic updating of values in evidence and interpretation nodes, but not specification of the temporal constraints themselves. Howarth and Buxton, as discussed below, used dynamically reconfigured nets to model evolving spatial relationships of vehicles as they move through the scene before using a standard 'tasknet' BBN in the behaviour evaluation. The usual approach is to use map-like knowledge of the environment to develop expectations of likely object motion, as in Buxton and Gong [30], for the segmentation and tracking of vehicles. For example, in Fig. 2 top node of the graph is image grid position which causally affects two daughter nodes coding orientation and size of the vehicle. These, in turn causally affect the observed flow vector features to be grouped under these contextual constraints for tracking. Others [24] have adopted dynamic probabilistic networks [21], which make use of the simple Markov property that the future is independent of the past given the present state.

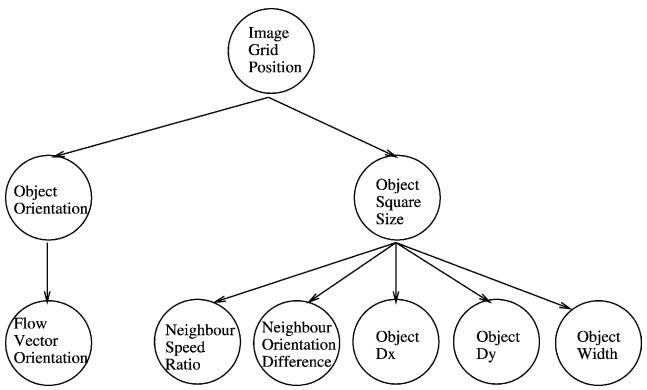

Fig. 2. Belief network that captures dependent relationships for motion segmentation and tracking [30].

4.3. Tracking humans

Isard and Blake have continued to develop general approaches to tracking that are well-suited to tracking fast human movement where it is particularly important to keep multiple likely explanations of motion patterns [41,42]. Conditional expectation maximisation (CEM) and conditional density estimation, 'condensation', trackers have their roots in recursive filtering and control theory and fit in very well with a probabilistic approach to full behaviour analysis. They provide a framework for representing density as a set of samples drawn from the underlying distribution in a Bayesian manner. The main difficulty is that these trackers must maintain a large number of samples, e.g. 15,000 to track a hand while drawing, in order to fully represent the underlying density. Recent developments by Hogg and colleagues (in Section 6) can improve on this using more knowledge in the form of contour models. In general, we should note that where the shape and dynamics are known, they can be modeled to simplify tracking, for example Wren [87].

4.4. Behaviour analysis

Some of the recent work using graphical models in behaviour analysis has been mentioned above. For example, support of spatial and temporal reasoning to generate dynamic scene descriptions [16], which uses hand-coded Bayes nets. Bayesian systems can also be used for intelligent vehicle control as in the BATmobile [24] or recent work by Oliver and Pentland [60]. We have also briefly introduced the coupled HMMs of Brand and colleagues [13] for understanding behaviour interactions, although this approach requires a great deal of training data. This is also true of parameterised HMMs [12], which can suffer from lack of stability in the interpretation compared to deformable model tracking and analysis seen in the last section. More recent work by Oliver, Rosario and Pentland [61,62] has developed reliable Bayesian vision systems. Two exciting recent developments are: (1) work by Galata, Johnson and Hogg using deformable models with HMM behaviour models for virtual actors [27] and (2) the action reaction learning of Jebara and Pentland [43,44], which models interactions and exploits new ideas from support vector machines (SVMs) in conjunction with generative Bayesian theory.

4.4.1. Visual representation

The main system we look at in some detail here, as an exemplar of graphical models, is the HIVIS-watcher system [38,40]. When reasoning about the behaviour of dynamic objects, it is useful if the representation of the properties related to each object are described in a local relative coordinate system. This involves recognising each moving object so that an 'intrinsic-front', such as the leading edge of a car, can be identified together with its spatial extent for

attentive processing. In a surveillance system we can obtain the poses of the scene objects via model-matching, making local reasoning attractive. The *local-form* is representation and reasoning that uses the intrinsic frame-of-reference of a perceived object. The *global-form* is representation and reasoning that uses the perceiver's frame-of-reference, which operates over the whole field-of-view. In active surveillance where we are inherently concerned with the 'here-and-now', it is important to form a consistent, task relevant interpretation of the observed behaviour. The system achieves this by task-level control policies that use typical-object behaviour models to specify both the preattentive (global) and attentive (local) processing to be performed in watching for evolving behaviour interactions. In particular, 'deictic' relationships, relative to their intrinsic frames of reference, are used to describe evolving contexts for observer and scene objects [15]. In this situated approach, it is not necessary to name and describe every object, but register only those relevant to the task so that the information registered is proportional to properties of interest rather than everything derivable in the dynamic scene.

4.4.2. Visual control

In HIVIS-watcher, there are three separate elements: the 'virtual-world' which holds data about the world, the 'peripheral-system' with operators that access the world, and the 'central-system' which controls system behaviour. The peripheral-system is based on Ullman's [84] visual routine processor, with reactive planning following the approach of Agre and Chapman [1] using spatial indexes [68]. Event detection operators are only run when selected by the task-level control system in this kind of situated vision approach [69]. The operators in the peripheral-system are separated into preattentive ones that are global, simple, and of low-cost and attentive ones which are applied to a single object and are more complex. The preattentive operators are used to guide application of attentive ones [37,38].

Bayes nets are the main mechanism used to update the context for interpretation of the dynamic scene at the behaviour level. This framework allows knowledge representation (symbolic and probabilistic) as well as playing a role in the on-line control of visual processing. For example, *gross-change-in-motion* is the preattentive cue to watch for in analysis of dynamic 'giveway' behaviour and to start gathering detailed evidence for the Bayesian tasknet. Similarly, *mutual-proximity* is the cue for 'overtaking' or 'following' so these are treated as starting the attentive processing of evidence gathering when these are the active policy. Evidence in this system is in terms of deictic relationships such as 'ref-obj3 is moving faster than ref-obj2', which help confirm or reject decisions about the relevant behaviour. Prior probabilities are not very relevant in such decision-making but statistical knowledge is learned as conditional probabilities for the dynamic relationships. However, how likely the different classes are *a priori* is relevant at the movement level [30] to initialise interpretation. All Bayes nets are updated using both parent to child (expectation) and child to parent (evidence) to give maximum likelihood explanation *a posteriori*. The complete dynamic updating cycle plays a role in controlling evidence collection as nodes can be parameters that affect the perceptual processing. For example, in Fig. 3 the same data is processed in different ways depending on the control policy coded in a DDN.

4.4.3. Visual learning

The learning above is off-line but is exploited on-line in the HIVIS system. So far, the processing is structured by analysis of what was involved in verifying that some behaviour is taking place, both what should be attended (via the preattentive cues and attentional markers) and how the reasoning should be structured using *typical object behaviour models*. However, this hand-coding is very difficult and involves a lot of empirical evaluation in order to get an effective processing scheme. Ongoing work learns the on-line visual cues for behaviours using a mixture of connectionist and Bayesian learning techniques to develop open classes of behavioural models. For complex models it helps to learn the detailed weights using probabilistic neural networks (PNNs) from an initial DAG structure. New work by Frey and Jojic [25] on flexible models and learning with variational methods in the graphical models community [47, 57] is rapidly developing techniques which can be applied to learn structure as well as parameters for cognitive vision tasks. For example, Yacoob and Black [88] have developed modeling and recognition with PCA 'activity-bases'.

5. Applications

Generative models can be used to advance cognitive vision for applications in human–computer interaction, education and tutoring, smart rooms and cars, and surveillance systems.

Many researchers are developing useful techniques for the rapidly growing area of multimodal and multimedia interaction. Turk's work at Microsoft with visual control of virtual actors is one example [83], as is Blake's use of sophisticated multi-object trackers [52] in hand tracking [53] and new work by Frey and Jojic on learning flexible sprites [25]. Another well-known centre for such work is the MIT Media lab led by Pentland. Some of this research has been mentioned earlier, especially under the graphical models section. MIT Media lab is also very active in the development of education and tutoring systems. For example, there has been ongoing research on sign language interpretation by Starner and Pentland [78] using HMMs. Bobick and others developed the exciting KidsRoom project while at MIT [11]. This was a perceptually based

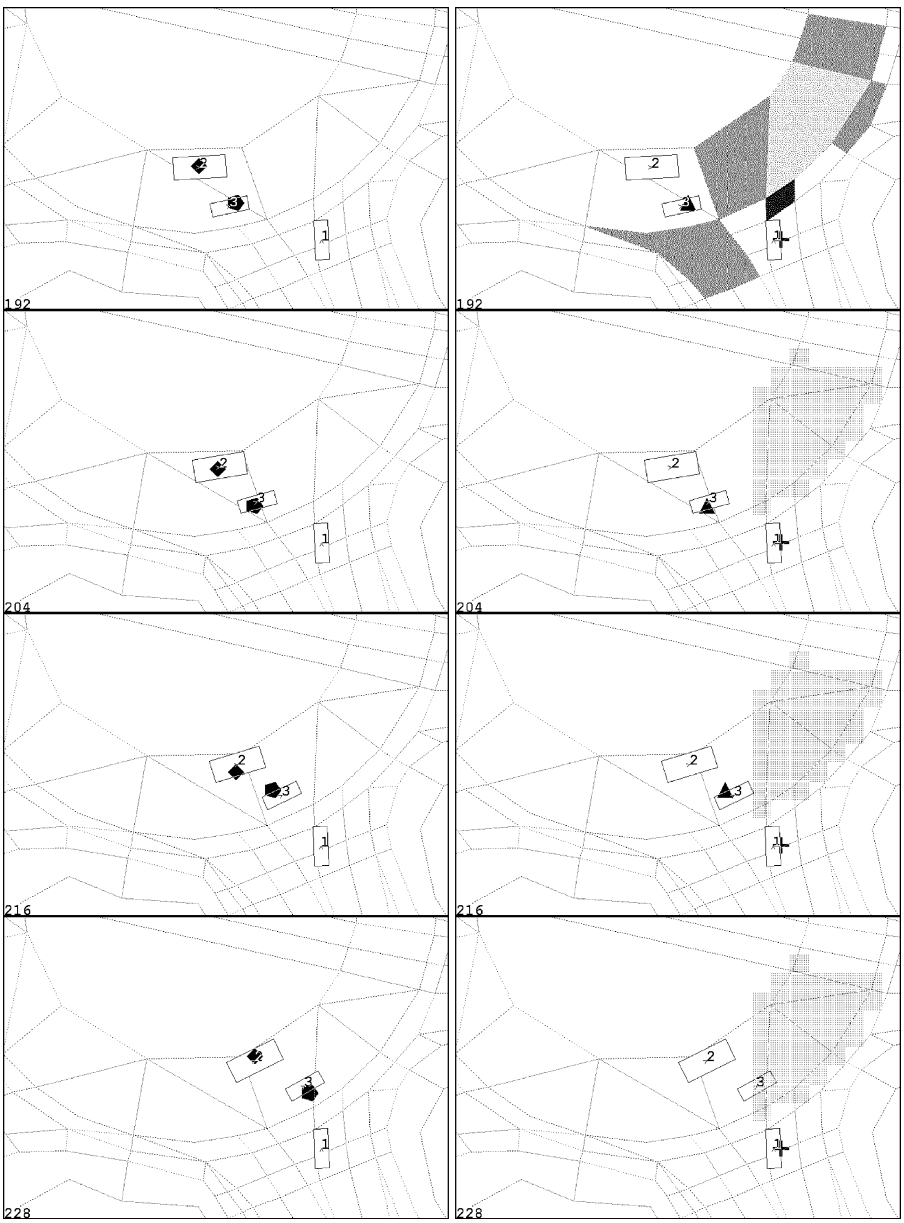

Fig. 3. Results from HIVIS-watcher [15] showing overtaking policy (left) and giveway policy (right).

environment in which children could interact while playing in a story-telling scenario.

Smart rooms and intelligent environments are now a rapidly growing area and a visionary overview of general issues raised by early research is given by Pentland [64]. A good example is given by the intelligent studios of Pinhanez and Bobick [67] who use their approximate world models and incorporate linguistic information to select vision routines and control of camera views. Smart cars using new sensor technology for vehicle control are being developed in conjunction with traffic monitoring in intelligent highway systems by Malik and colleagues [7, 54]. Also a SmartCar testbed with learning techniques to model and recognise driving behaviour has been developed by Oliver and Pentland [60]. The techniques here are again exploiting graph-based models, both HMM and CHMMs for: *passing*, *changing lanes*, *turning*, *starting and stopping* with prediction a full second before the manoeuver starts. This anticipation is essential for control and driver assistance systems.

A great deal of cognitive vision research has been done in the field of surveillance, for example early work VIEWS discussed in Section 4 [18]. Here again there is the need to integrate vision and language to deliver conceptual descriptions of the scene, which means that language must be grounded in vision and there must be a clear ontology for events, activities and scenarios. For example, recent work by Nevatia's group has used a Bayesian approach to develop multi-agent event recognition [34,35]. In addition, Bayes nets and dynamic decision nets have been developed for

task-based control [36] and many of the reasoning tasks required in advanced surveillance systems by Buxton et al. [14,39,40]. These allow ongoing interpretation of the visual evidence in a control loop to actively direct processing for tracking or behaviour evaluation and deliver a conceptual description of the dynamic scene which is relevant to the surveillance task.

6. Conclusion

To summarise the arguments in the earlier sections, we are suggesting that generative models are essential in the design and integration of cognitive vision systems. For the interpretation of a dynamic scene, these systems not only track and analyse movements of objects and people in the scene, but also evaluate their behaviour in the context of the vision system task. Thus, models are used in both interpretation and control of the visual processing. In addition, these models should ideally be learned for the interpretation and control of cognitive vision tasks to automate construction and support adaptive processing.

We reviewed work using the deformable models approach and using statistical learning techniques to extract the models of typical behaviour at the level of movement trajectories, events, and more abstract sequences of movement. However, note that all this analysis was in the image plane and does not involve active visual processing or task control. There are approaches using fuller biomechanical models by Terzopoulos and colleagues [81] aimed more at an artificial life level of modeling for behaviour. These biomimetic agents can be situated in realistic environments to develop a software approach to the design of active vision systems [82]. Nevertheless, integration of cognitive vision systems involves more than the deformable models, as they must be combined with some kind of situated vision framework to allow action, or within a full cognitive vision framework to allow long-term reasoning and evaluation of behaviour.

A useful approach to cognitive vision system design with full information integration is offered by graphical models, where the representations reflect dynamic dependencies for interpretation of visual evidence. We reviewed work where Bayes nets and HMMs are partially structured using contextual knowledge [30] and detailed parameters learned for prior and conditional probabilities between the important concepts in dynamic scene interpretation. Complex models suffer from the combinatorics of top-down and bottom-up message passing in on-line evaluation of beliefs required to compute the most likely explanation of the visual evidence. However, if singly connected trees of limited depth are used in a decomposition or simplification of the problem, they are effective for real-time systems, incorporating active selection for behaviour interpretation [38]. The contextual constraints and control of processing in updating beliefs from evidence in these graphical models is easily understood. Modeling essential aspects of situated vision by using deictic markers in conjunction with the Bayes nets and task-based control [40] is also possible. This type of approach has all the required characteristics for cognitive vision: contextual processing and control with learning capabilities.

Finally, for applications, it seems that there is a need for interdisciplinary work in cognitive science, HCI and AI approaches to vision. We have seen how cognitive science research on deictic descriptions and attentional markers in situated vision [1,68] has been used in the work of Howarth and Buxton [15]. This is particularly important for the generation of conceptual descriptions of dynamic scenes [40]. We also mentioned the role of active viewpoint control in cognitive vision systems, which is an important area of research but not reviewed here [23]. Recent research on overt attention in natural tasks, such as Land and McLeod's study of a cricket batsman trying to hit the ball [49], gives great insight into perceptual performance for such active camera systems. Future progress in this area depends on multi-disciplinary projects that combine such cognitive studies with learnt generative models to develop effective cognitive vision systems.

Acknowledgements

The author gratefully acknowledges the invaluable discussion and help of co-workers Shaogang Gong and Richard Howarth who completed much of the original work in the section on graphical models reviewed here. Also thanks to David Hogg for permission to reproduce figures and discussion of approaches for the section on deformable models. Finally, thanks are due for funding under EU IST cognitive vision initiative ActIPret project.

References

[1] P.E. Agre, D. Chapman, Pengi: an implementation of a theory of activity, American Association for Artificial Intelligence (1987) 268–272.
[2] D. Ballard, Animate vision, Artificial Intelligence 48 (1991) 57–86.
[3] A. Baumberg, Hierarchical shape fitting using an iterated linear filter. British Machine Vision Conference, 1996, Edinburgh, Scotland, p. 313–323.
[4] A. Baumberg, D.C. Hogg, Learning flexible models from image sequences, European Conference on Computer Vision, 1994, Stockholm, Sweden, p. 299–308.
[5] A. Baumberg, D.C. Hogg, An adaptive eigenshape model, British Machine Vision Conference, Birmingham, UK, 1995, p. 87–96.
[6] A. Baumberg, D.C. Hogg, Generating spatiotemporal models from training examples, Image and Vision Computing 14 (1996) 525–532.
[7] D. Beymer, P. McLauchlan, B. Coifman, J. Malik, A real-time computer vision system for measuring traffic parameters, IEEE Conference on Computer Vision and Pattern Recognition, San Juan, Puerto Rico, 1997.

[8] M.J. Black, D.J. Fleet, Probabilistic detection and tracking of motion boundaries, International Journal of Computer Vision 38 (2000) 231–245.
[9] A. Blake, M. Isard, Active Contours, Springer, Berlin, 1998.
[10] A. Blake, M. Isard, D. Reynard, Learning to track the visual motion of contours, Artificial Intelligence 78 (1995) 179–212.
[11] A. Bobick, S. Intille, J. Davies, F. Baird, C. Pinhanez, L. Campbell, Y. Ivanov, A. Schutte, A. Wilson, The Kidsroom: a perceptually-based interactive and immersive story environment, Presence: Teleoperators and Virtual Environments 8 (1999) 367–1391.
[12] A. Bobick, A. Wilson, A state-based technique for the summarization and recognition of gesture. In IEEE International Conference on Computer Vision, Cambridge, MA, 1995, p. 382–388.
[13] M. Brand, N. Oliver, A. Pentland, Coupled hidden Markov models for complex action recognition, IEEE Conference on Computer Vision and Pattern Recognition, San Juan, Puerto Rico, 1997.
[14] H. Buxton, S. Gong, Visual surveillance in a dynamic and uncertain world, Artificial Intelligence 78 (1995) 431–459.
[15] H. Buxton, R. Howarth, Watching behaviour: the role of context and learning. In International Conference on Image Processing, Lausanne, Switzerland, 1996, p. II: 797–800.
[16] H. Buxton, R. Howarth, Spatial and temporal reasoning in the generation of dynamic scene descriptions, in: P. Olivier, K.-P. Gapp (Eds.), Representation and Processing of Spatial Expressions, Lawrence Erlbaum Associates, Hillsade, NJ, 1997.
[17] E. Charniak, Bayesian networks without tears, AI Magazine 12 (4) (1991) 50–63.
[18] VIEWS consortium, The VIEWS project and wide-area surveillance, ESPRIT Workshop at ECCV, Genoa, Italy, 1992.
[19] T.F. Cootes, D. Cooper, C.J. Taylor, J. Graham, Active shape models—their training and application, Computer Vision and Image Understanding 61 (1995) 38–59.
[20] T.F. Cootes, C.J. Taylor, D. Cooper, J. Graham, Training models of shape from sets of examples, British Machine Vision Conference, Springer, Berlin, 1992, p. 9–18.
[21] T. Dean, K. Kanazawa, Probabilistic temporal reasoning, American Association for Artificial Intelligence, St Paul, Minnesota, 1988, p. 524–528.
[22] G.J. Edwards, C.J. Taylor, T.F. Cootes, Learning to identify and track faces in image sequences, British Machine Vision Conference, Colchester, UK, 1997, p. 130–139.
[23] J.O. Eklundh, P. Nordlund, T. Uhlin, Issues in active vision: attention and cue integration/selection, British Machine Vision Conference, Cambridge, UK, 1996, p. 1–12.
[24] J. Forbes, T. Huang, K. Kanazawa, S. Russell, The BATmobile: towards a Bayesian automated taxi, International Joint Conference on Artificial Intelligence, Montreal, Canada, 1995, p. 1878–1885.
[25] B.J. Frey and N. Jojic, Flexible models: a powerful alternative to exemplars and explicit models, CVPR Workshop on Models versus Exemplars, Kauai, Hawaii, 2001.
[26] N. Friedman, Learning belief networks in the presence of missing values and hidden variables, International Conference on Machine Learning, 1997, p. 125–133.
[27] A. Galata, N. Johnson, D.C. Hogg, Learning variable length Markov models of behaviour, Computer Vision and Image Understanding 81 (2001) 398–413.
[28] S. Gong, Visual behaviour: modelling hidden purposes in motion, International Conference on Neural and Stochastic Methods in Image and Signal Processing, San Diego, CA, 1992, p. 45–57.
[29] S. Gong, Visual observation as reactive learning, International Conference on Adaptive and Learning Systems, Orlando, FL, 1992, p. 175–187.
[30] S. Gong, H. Buxton, Bayesian nets for mapping contextual knowledge to computational constraints in motion segmentation and tracking, British Machine Vision Conference, Guildford, UK, 1993, p. 229–238.
[31] D. Heckerman, A tutorial on learning with Bayesian networks, Learning in Graphical Models, NATO Science Series, 1998, p. 301–354.
[32] A. Hill, A. Thornham, C.J. Taylor, Model-based interpretation of 3D medical images, British Machine Vision Conference, Guildford, UK, 1993, p. 339–349.
[33] D.C. Hogg, Model-based vision: a program to see a walking person, Image and Vision Computing 1 (1983) 5–20.
[34] S. Hongeng, F. Bremond, R. Nevatia, Bayesian framework for video surveillance applications, International Conference on Pattern Recognition, Barcelona, Spain, 2000.
[35] S. Hongeng, R. Nevatia, Multi-agent event recognition, IEEE International Conference on Computer Vision, Vancouver, Canada, 2001.
[36] R. Howarth, Interpreting a dynamic and uncertain world: task-based control, Artificial Intelligence 100 (1998) 5–85.
[37] R. Howarth, H. Buxton, Selective attention in dynamic vision, International Joint Conference on Artificial Intelligence, Chambery, France, 1993, p. 1579–1585.
[38] R. Howarth, H. Buxton, Visual surveillance monitoring and watching, European Conference on Computer Vision, Cambridge, UK, 1996, p. 321–334.
[39] R. Howarth, H. Buxton, Attentional control for visual surveillance, in: S. Maybank, T. Tan (Eds.), ICCV Workshop on Visual Surveillance, IEEE Press, London, 1997.
[40] R. Howarth, H. Buxton, Conceptual descriptions from monitoring and watching image sequences, Image and Vision Computing 18 (2000) 105–135.
[41] M. Isard, A. Blake, Contour tracking by stochastic propagation of conditional density, European Conference on Computer Vision, Cambridge, UK, 1996, p. 343–356.
[42] M. Isard, A. Blake, A mixed-state condensation tracker with automatic model-switching, IEEE International Conference on Computer Vision, Bombay, India, 1998, p. 107–112.
[43] T. Jebara, A. Pentland, Action reaction learning: automatic visual analysis and synthesis of interactive behaviour, International conference on Vision Systems, Gran Canaria, Spain, 1999, p. 273–292.
[44] T. Jebara, A. Pentland, On Reversing Jensen's Inequality, Advances in Neural Information Processing Systems, vol. 13, Denver, Colorado, 2000.
[45] N. Johnson, D.C. Hogg, Learning the distribution of object trajectories for event recognition, British Machine Vision Conference, Birmingham, UK, 1995, p. 583–592.
[46] N. Johnson, D.C. Hogg, The acquisition and use of interaction behaviour models, IEEE Conference on Computer Vision and Pattern Recognition, Santa Barbara, 1998, p. 866–871.
[47] M.I. Jordan, Learning in Graphical Models, NATO Science Series, 1998.
[48] M. Kass, A. Witkin, D. Terzopoulos, Snakes: active contour models. In IEEE International Conference on Computer Vision, London, UK, 1987, p. 259–268.
[49] M.R. Land, P. McLeod, From eye movements to actions: How batsmen hit the ball, Nature Neuroscience 3 (2000) 1340–1345.
[50] T.S. Levitt, J.M. Agosta, T.O. Binford, Model-based influence diagrams for machine vision, Uncertainty in Artificial Intelligence. Machine Intelligence and Pattern Recognition Series, vol. 10, North-Holland, Amsterdam, 1990.
[51] T.S. Levitt, T.O. Binford, G.J. Ettinger, Utility-based control for computer vision, Uncertainty in Artificial Intelligence. Machine Intelligence and Pattern Recognition Series, vol. 9, North-Holland, Amsterdam, 1990.
[52] J. MacCormick, A. Blake, A probabilistic exclusion principle for tracking multiple objects, IEEE International Conference on Computer Vision, Vancouver, Canada, 1999, p. 572–578.

[53] J. MacCormick, M. Isard, Partitioned sampling, articulated objects, and interface-quality hand tracking, European Conference on Computer Vision, Dublin, Ireland, 2000, p. II: 3–19.
[54] J. Malik, J. Weber, T. Luong, D. Koller, Smart cars and smart roads, British Machine Vision Conference, Birmingham, UK, 1995, p. 367–382.
[55] D. Metaxas, D. Terzopoulos, Shape and non-rigid motion estimation though physics-based synthesis, IEEE Transactions on Pattern Analysis and Machine Intelligence 15 (1993) 580–591.
[56] R.J. Morris, D.C. Hogg, Statistical models of object interaction, International Journal of Computer Vision 37 (2000) 209–215.
[57] K. Murphy, Learning Bayes net structure from sparse data sets, Technical report, University of Berkeley, CA, Computer Science Technical Report, 2001.
[58] C. Nastar, Vibration modes for nonrigid analysis in 3D images, European Conference on Computer Vision, Stockholm, Sweden, 1994, p. 231–238.
[59] A.E. Nicholson, J.M. Brady, The data association problem when monitoring robot vehicles using dynamic belief networks, European Conference on Artificial Intelligence, Vienna, Austria, 1992, p. 689–693.
[60] N. Oliver, A. Pentland, Graphical models for driver behaviour recognition in a smartcar, Intelligent Vehicles, Detroit, US, 2000.
[61] N. Oliver, B Rosario, A. Pentland, Graphical models for recognising human interactions, Advances in Neural Information Processing Systems, Denver, Colorado, 1998.
[62] N. Oliver, B. Rosario, A. Pentland, A Bayesian computer vision system for modeling human interactions, International Conference on Vision Systems, Gran Canaria, Spain, 1999.
[63] J. Pearl, Probabilistic Reasoning in Intelligent Systems, Networks of Plausible Inference, Morgan Kaufmann, Los Altos, CA, 1988.
[64] A. Pentland, Smart rooms, Scientific American 274 (1996) 54–62.
[65] A. Pentland, B. Horowitz, Recovery of non-rigid motion and structure, IEEE Transactions on Pattern Analysis and Machine Intelligence 13 (1991) 730–742.
[66] A. Pentland, A. Liu, Modeling and prediction of human behaviour, Neural Computation 11 (1999) 229–242.
[67] C. Pinhanez, A. Bobick, Intelligent studios: modeling space and action to control TV cameras, Applications of Artificial Intelligence 11 (1997) 285–306.
[68] Z. Pylyshyn, The role of location indexes in spatial perception: a sketch of the FINST spatial-index model, Cognition 32 (1989) 65–97.
[69] Z. Pylyshyn, Visual indexes, preconceptual objects, and situated vision, Cognition 80 (2001) 127–158.
[70] L.R. Rabiner, A tutorial on hidden Markov models, Proceedings of the IEEE 77 (1989) 257–286.
[71] P. Remagnino, A. Baumberg, T. Grove, D. Hogg, A. Worrall, K. Baker, An integrated traffic and pedestrian model-based vision system, British Machine Vision Conference, Colchester, UK, 1997, p. 380–389.
[72] R.D. Rimey, C.M. Brown, Where to look next using a Bayes net: incorporating geometric relations, European Conference on Computer Vision, Genoa, Italy, 1992, p. 542–550.
[73] R.D. Rimey, C.M. Brown, Control of selective perception using Bayes nets and decision theory, International Journal of Computer Vision 12 (1994) 173–209.
[74] J. Rittschler, A. Blake, Classification of human body movement, IEEE International Conference on Computer Vision, 1999, p. 634–639.
[75] S. Roweis, Z. Ghahramani, A unifying review of linear Gaussian models, Neural Computation 11 (1999) 305–345.
[76] S. Roweis, L.K. Saul, Nonlinear dimensionality reduction by locally linear embedding, Science 290 (2001) 2323–2326.
[77] D.J. Spiegelhalter, R.G. Cowell, Learning in probabilistic expert systems, Bayesian Statistics 4, Oxford University Press, Oxford, 1992.
[78] T. Starner, A. Pentland, Real-time American sign language recognition using Hidden Markov Models, International Symposium on Computer Vision, Coral Gables, FL, 1995, p. 265–270.
[79] G.D. Sullivan, A. Worrall, J.M. Ferryman, Visual object recognition using deformable models of vehicles, ICCV Workshop on Context-based Vision, Cambridge, MA, 1995, p. 75–86.
[80] D. Terzopoulos, Physically-based facial modeling, analysis, and animation, Journal of Visualisation and Computer Animation 1 (1990) 73–80.
[81] D. Terzopoulos, Visual modeling for computer animation: Graphics with a vision, Computer Graphics 33 (1999) 42–45.
[82] D. Terzopoulos, T. Rabie, Animat vision: active vision in artificial animals, Videre: Journal of Computer Vision Research 1 (1997) 2–19.
[83] M. Turk, Visual interaction with lifelike characters, IEEE International Conference on Automatic Face and Gesture Recognition, Killington, VE, 1996.
[84] S. Ullman, Visual routines, Cognition 18 (1984) 97–159.
[85] S.D. Whitehead, D.H. Ballard, Learning to perceive and act by trial and error, Machine Learning 7 (1991) 45–83.
[86] C. Wren, A. Azarbayejani, T. Darrell, A. Pentland, Pfinder: real-time tracking of the human body, IEEE Transactions on Pattern Analysis and Machine Intelligence 19 (1997) 780–785.
[87] C. Wren, A. Pentland, Dynamic models of human motion, IEEE International Conference on Automatic Face and Gesture Recognition, Nara, Japan, 1998.
[88] Y. Yacoob, M.J. Black, Parameterized modelling and recognition of activities, Computer Vision and Image Understanding 73 (1999) 232–247.