Tracking Discontinuous Motion using Bayesian Inference

Jamie Sherrah and Shaogang Gong

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

Abstract. Robustly tracking people in visual scenes is an important task for surveillance, human-computer interfaces and visually mediated interaction. Existing attempts at tracking a person's head and hands deal with ambiguity, uncertainty and noise by intrinsically assuming a consistently continuous visual stream and/or exploiting depth information. We present a method for tracking the head and hands of a human subject from a single view with no constraints on the continuity of motion. Hence the tracker is appropriate for real-time applications in which the availability of visual data is constrained, and motion is discontinuous. Rather than relying on spatio-temporal continuity and complex 3D models of the human body, a Bayesian Belief Network deduces the body part positions by fusing colour, motion and coarse intensity measurements with contextual semantics.

1 Introduction

Tracking human body parts and motion is a challenging but essential task for modelling, recognition and interpretation of human behaviour. In particular, tracking of at least the head and hands is required for gesture recognition in human-computer interface applications such as sign-language recognition and visually mediated interaction. Existing methods for markerless tracking can be categorised according to the measurements and models used [9]. In terms of measurements, tracking usually relies on intensity information such as edges [10, 2, 17, 5], skin colour and/or motion segmentation [24, 14, 11, 16], or a combination of these with other cues including depth [13, 25, 19, 1]. The choice of model depends on the application of the tracker. If the tracker output is to be used for some recognition process then a 2D model of the body will suffice [16, 11]. On the other hand, a 3D model of the body may be required for generative purposes, to drive an avatar for example, in which case skeletal constraints can be exploited [25, 19, 5], or deformable 3D models can be matched to 2D images [10, 17].

Colour-based tracking of body parts is a relatively robust and inexpensive approach. Nevertheless the loss of information involved induces problems of noise, uncertainty, and ambiguity due to occlusion and distracting "skin-coloured" background objects. The two most difficult problems to deal with when tracking the head and hands are occlusion and correct hand association. Occlusion occurs when a hand passes in front of the face or intersects with 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 at tracking cope with these problems using temporal prediction and/or depth information. Temporal prediction intrinsically assumes temporal order and continuity in measured data, therefore a consistent, sufficiently high frame rate is required. The use of depth information requires more than one camera and solution of the correspondence problem which is computationally non-trivial.

We argue that robust, real-time human tracking systems must be designed to work with a source of *discontinuous visual information.* Any vision system operates under constraints that attenuate the bandwidth of visual input. In some cases the data may simply be unavailable, in other cases computation time is limited due to finite resources. A further and more significant computational constraint is associated with complexity and stability of behavioural models. Exhaustive modelling of the world would be prohibitively complex; rather it is more realistic to establish economical models or *beliefs* about the environment which are iteratively updated by visual observations. Since the models are not exhaustive, not all of the visual information needs to be processed. In fact, it may be undesirable to absorb all available visual information into belief structures because instability, or "catastrophic unlearning", may result. Therefore a robust vision system should be based on selective attention mechanisms to filter out irrelevant information and use only salient visual stimuli to update its beliefs [23]. While selective attention is traditionally considered in the spatial domain, in this work we cast the notion into the temporal domain in order to relax the underlying constraint of temporal order and continuity required in tracking visual events over time.

We achieve the goal of tracking discontinuous human body motion by replacing the problem of spatio-temporal prediction with reasoning about body-part associations based on contextual knowledge. Our approach

uses *Bayesian Belief Networks* (BBNs) to fuse high-level contextual knowledge with sensor-level observations. Belief networks are an effective vehicle for combining user-supplied semantics with conflicting and noisy observations to deduce an overall consistent interpretation of the scene. BBNs have been used previously as a framework for tracking multiple vehicles under occlusion using contextual information [4]. In [18], a naive BBN was used to characterise and classify objects in a visual scene. For tracking body parts under discontinuous motion the BBN framework is ideal because unlike other tracking methods such as Kalman filtering or CONDENSATION [12] that explicitly model the dynamics through change, Belief Networks model absolute relationships between variables and can make deductive leaps given limited but significant evidence. Nevertheless, the accumulated beliefs still implicitly reflect all currently observed evidence over time. We demonstrate that through iterative revision of hypotheses about associations of hands with skin-coloured image regions, such an *atemporal belief-based tracker* is able to recover from almost any form of track loss. In Section 2 we describe the context, assumptions and measurements used by the body tracker. In Section 3 we present the framework for combining these observations with contextual knowledge using BBNs. An experimental comparison of our tracker with a dynamic tracker and a non-contextual tracker is presented in Section 4, and the conclusion is given in Section 5.

2 Tracking Discontinuous Motion from 2D Observations

The merits of any given behavioural modelling method are established according to the purpose for which it is used, therefore it is appropriate at this point to introduce the context for our tracking approach and the assumptions made. We are interested in modelling individual and group behaviours for visually mediated interaction using only a single 2D view, therefore depth information is unavailable. Behaviour models are used to interpret activities in the scene and change the view to focus on regions of interest. Therefore we have the luxury of not requiring full 3D tracking of the human body parts, which would rely on expensive matching to unreliable intensity observations. On the other hand, the system is required to simultaneously track several people which generally results in a variable and relatively low frame rate. From our experience with these conditions, 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, in images of manageable resolution containing several people (all images used in this work are 320×240 pixels), the hands may occupy regions as small as ten pixels or less wide, making appearance-based methods unreliable.

To illustrate the nature of the discontinuous body motions under these conditions, Figure 1 shows the head and hands positions and accelerations (as vectors) for two video sequences, along with sample frames. The video frames were samples at 18 frames per second (fps). Even so, there are many significant temporal changes in both the magnitude and orientation of the acceleration of the hands. It may be unrealistic to attempt to model the dynamics of the body under these circumstances.

We propose that under the following assumptions, the ambiguities and uncertainties associated with tracking a person's discontinuous head and hand movement can be overcome using only information from a single 2D view without modelling the full dynamics of the human body:

1. the subject is oriented roughly towards the camera for most of the time.
2. the subject is wearing long sleeves.
3. reasonably good colour segmentation of the head and hands is possible, and
4. the head and hands are the largest moving skin colour clusters in the image.

The robust visual cues used for tracking are now described, followed by a description of the head-tracking and bootstrapping methods.

2.1 Computing Visual Cues

Real-time vision systems have two chief practical requirements: computational efficiency and robustness. Computational constraints exclude the use of expensive optimisation methods, while robustness requires tolerance of assumption violation. To meet these requirements we adopt a philosophy of perceptual fusion: independent, relatively inexpensive visual cues are combined to benefit from their mutual strengths and achieve some invariance to their assumptions [7]. The cues that are used to drive our body tracker are skin colour, image motion and coarse intensity information, namely hand orientation. Pixel-wise skin colour probability has been previously shown to be a robust and inexpensive visual cue for identification and tracking of people under varying lighting conditions [22]. Skin colour probabilities can be computed for an image and thresholded to obtain a binary skin image, an example is shown in Figure 2(b). Here image motion is naively computed as the thresholded difference between pixel intensities in successive frames; an example is shown in Figure 2(c).

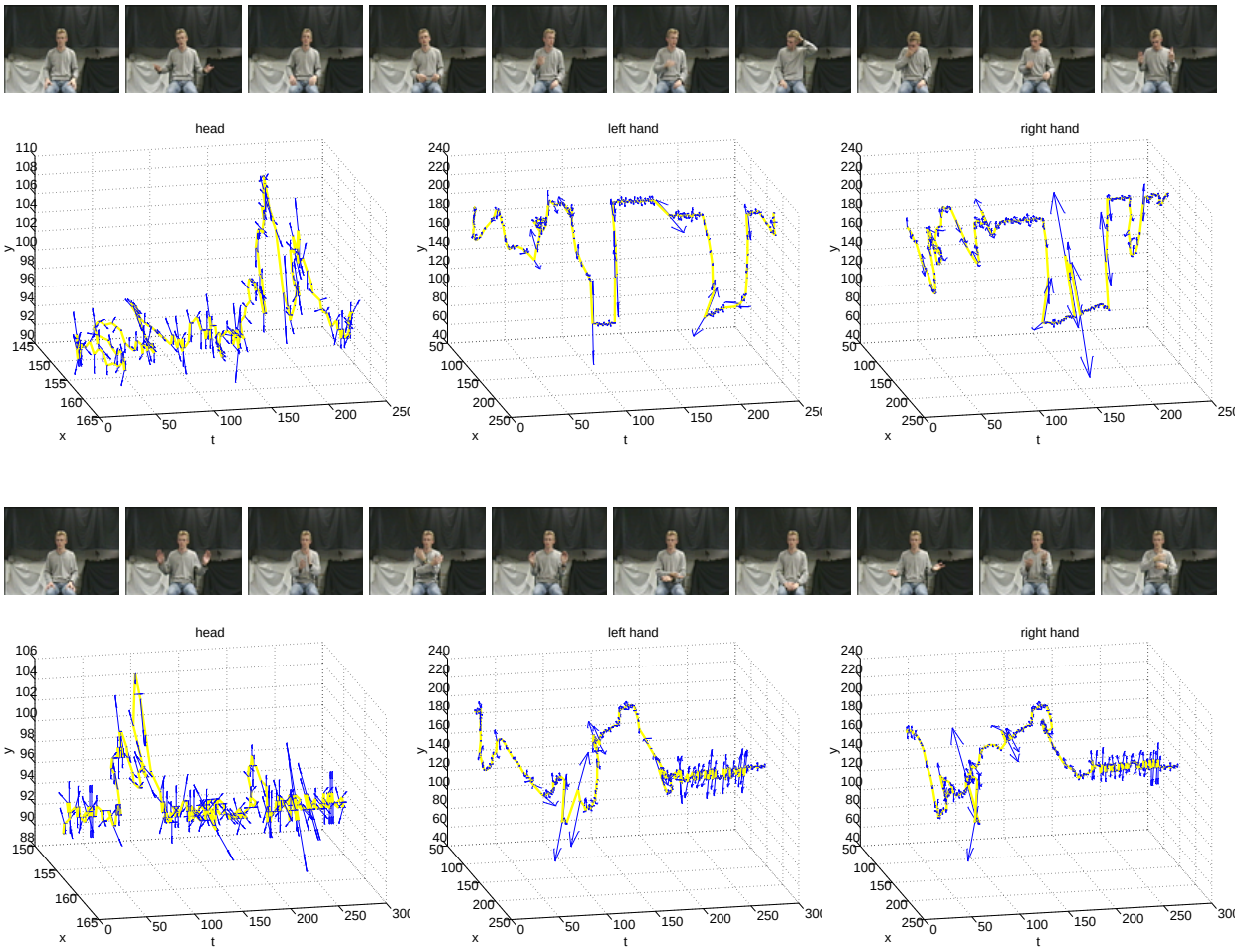
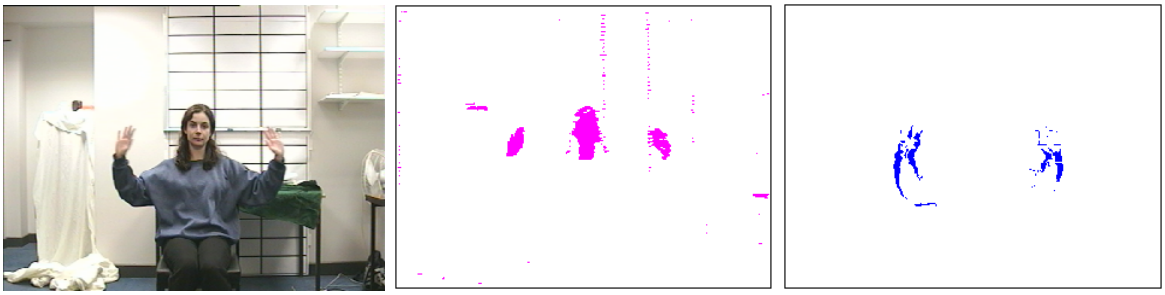


Fig. 1. Two examples of behaviour sequences and their tracked head and hand positions and accelerations. At each time frame, the 2D acceleration is shown as an arrow with arrowhead size proportional to the acceleration magnitude. From left to right, the plots correspond to the head, left hand and right hand.

Skin colour and motion are natural cues for focusing attention and processing resources on salient regions in the image. Note that although distracting noise and background clusters appear in the skin image, these can be eliminated at a low level by "AND"ing directly with motion information. However, fusion of these cues at this low level of processing is premature due to loss of information. For example, the motion information generally occurs only at the edges of of the moving object, making the fused information too sparse.

The problem of associating the correct hands over time can usually be solved using 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. In the absence of depth information or 3D skeletal constraints, we use intensity information to assist in resolving incorrect assignment. The intensity image of each hand is used to obtain a very coarse measurement of hand orientation which is robust even in low resolution imagery. The restricted kinematics of the human body are loosely modelled to exploit the fact that only certain hand orientations are likely at any position in the image relative to the head.

The accumulation of a statistical hand orientation model is illustrated in Figure 3. Assuming that the subject is facing the camera, the image is divided coarsely into a grid of histogram bins. We then artificially synthesise 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. 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. 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



(a) (b) (c)

Fig. 2. Example of visual cues measured from video stream. (a) original image; (b) binary skin colour image; and (c) binary motion image.

relative to the tracked head position is used to index the histogram and obtain the likelihood of the hand orientation given the position.

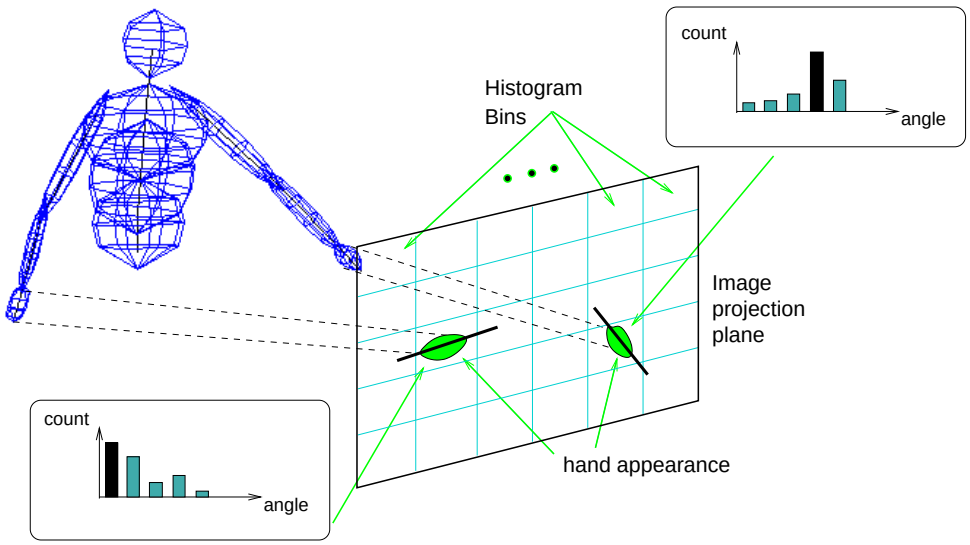


Fig. 3. Schematic diagram of the hand orientation histogram process.

2.2 Head Tracking using Mean Shift

The first two constraints to be exploited are that the head is generally larger than the hands in the image, and that head movement is significantly more stable and moderate than hand motion. We track the head directly using an iterated mean shift algorithm [3]. This method converges on the local mode of the skin probability distribution. Despite its simplicity, the algorithm is very robust to occlusion by hands. The head is modelled as a rectangular region containing skin pixels. A search region is defined such that it is centred on the head box but is slightly larger. Given an initial/previous position $(c_x(t), c_y(t))$, the algorithm is to iteratively calculate the spatial mean of skin pixels in the rectangular search region and shift the box to be centred on that estimated mean until it converges, as expounded in Figure 4.

After convergence, the size of the head box is set according to the following heuristic:

$$w = \sqrt{n_{skin}} \tag{1}$$

loop:
- $c_x(t-1) = c_x(t)$, $c_y(t-1) = c_y(t)$
- $c_x(t) = \frac{1}{n_{skin}} \sum_{p \in S} p_x$
  $c_y(t) = \frac{1}{n_{skin}} \sum_{p \in S} p_y$
  where $p = (p_x, p_y)$ is a pixel, S is the set of skin pixels in the search region and $n_{skin} = |S|$.
- Set search region centre to $(c_x(t), c_y(t))$.

until $c_x(t) = c_x(t-1)$ and $c_y(t) = c_y(t-1)$.

Fig. 4. The mean shift algorithm for tracking the head box.

$$h = 1.2w \tag{2}$$

Note that the search region must be slightly larger than the head rectangle to avoid continual shrinking of the box, and to allow significant movement of the head without loss of track.

2.3 Local Skin Colour Clusters

Under the assumption that the head and hands form the largest moving connected skin coloured regions in the image, tracking the hands reduces to matching the previous hand estimate to the skin clusters in the current frame. This association can be performed either at the pixel level or at a "cluster" level. At the pixel level, hands are tracked using local search via updating of spatial hand box means and variances (size). At the cluster level, a connected components algorithm is used to find all spatially connected sets of coloured pixels, which are subsequently treated as discrete entities. We have chosen to use the cluster representation for three reasons:

- The pixel-level approach requires estimation of spatial means and variances of pixels which are quite sensitive to outliers. Even if medians are used instead of means, the hand box sizes are very sensitive to noise.
- The local tracking approach requires heuristic search parameters, and is generally invalid for discontinuous motion since the hands may move a significant distance from one frame to the next.
- Reasoning about hand associations is easier using the higher-level cluster representation.

We used a connected components algorithm that has computational complexity linear in the number of skin pixels to obtain a list of skin clusters in the current frame. The components are drawn only from those portions of the region outside of an exclusion region defined by the head tracker box. The exclusion region is slightly larger than the head box due to protruding necklines or ears that can be mistaken for potential hand clusters. Clusters containing only a few pixels are assumed to be noise and removed. Finally the clusters are sorted in descending order of their skin pixel count for subsequent use.

2.4 Initialisation

Tracking is initialised by using skin colour to focus on areas of interest, then performing a multi-scale, multi-position identity-independent face search within these regions using a Support Vector Machine (SVM) [20]. An example is shown in Figure 5. The SVM has been trained only on frontal and near-frontal faces, so it is assumed that the subject is initially facing approximately towards the camera. The mean shift head tracker is then initialised on the detected face region. Since the hands tracker only uses temporal association as a secondary cue, full tracking of the body can begin immediately after this partial initialisation.

3 Reasoning about Body-Parts Association using Bayesian Inference

Given only the visual cues described in the previous section, the problem is now to determine the association of skin colour clusters to the left and right hands. One can consider this situation to be equivalent to watching a mime artist wearing a white face mask and white gloves in black clothing and a black background (see Figure 2(b)). Further, only discontinuous information is available as though a strobe light were operating, creating a "jerky" effect (see Figure 6(a)). Under these conditions explicit modelling of body dynamics

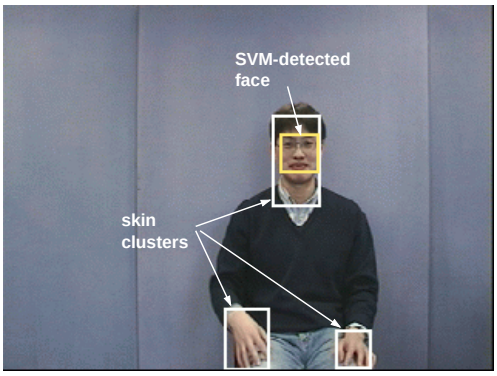


Fig. 5. Example of the tracker initialisation using an SVM.

inevitably makes too strong an assumption about image data. Rather, the tracking can be performed better and more robustly through a process of deduction. This requires full exploitation of both visual cues and high-level contextual knowledge. 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 Figures 6(b) and 6(c), or (3) it has disappeared from the image as in Figure 6(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 Figure 6(e).

Clearly a mechanism is required for reasoning about the situation. In the next section, Bayesian Belief Networks (BBNs) are introduced as a mechanism for performing inference, after which we describe how BBNs have been applied to our tracking problem.

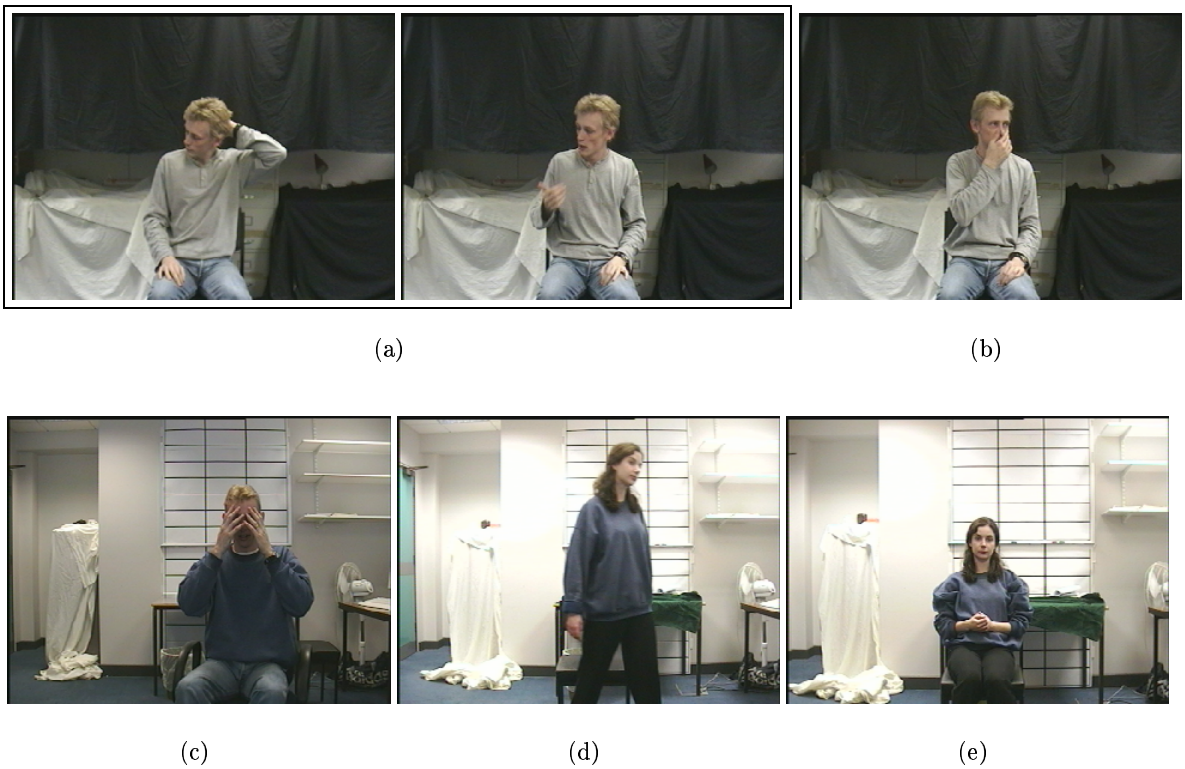


Fig. 6. Examples of the difficulties associated with tracking the body. (a) motion is discontinuous between frames; (b) one hand occludes the face; (c) both hands occlude the face; (d) a hand is invisible in the image; and (e) the hands occlude each other.

3.1 Bayesian Belief Networks

The obvious method of incorporating semantics into our tracking 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 rule-bases are notoriously sensitive to noise because once a decision has been made based on some fixed threshold, subsequent decision-making is isolated from the contending unchosen possibilities. Sensitivity to noise is undesirable in our situation since we are dealing with very noisy and uncertain image data. The rule-based approach can also suffer from global consistency problems because commitment to a single decision precludes feedback of higher-level knowledge to refine lower-level uncertain observations or beliefs.

An alternative approach to reasoning is based on soft, probabilistic decisions. Under such a framework all hypotheses are considered to some degree but with an associated probability. Bayesian Belief Networks provide a rigorous framework for combining semantic and sensor-level reasoning under conditions of uncertainty [21, 6, 8]. Given a set of variables $\mathbf{W}$ representing the scenario[1], 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 which reflect both prior and observed information:

$$BEL(x) = P(x|\mathbf{e}) \tag{3}$$

where $BEL(x)$ is the belief in the value of variable X given the evidence $\mathbf{e}$. Updating of beliefs occurs through a distributed message-passing process that is made possible via exploitation of local dependencies and global independencies. Hence dissemination of evidence to update currently-held beliefs can be performed in a tractable manner to arrive at a globally consistent evaluation of the situation.

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:

$$P(\mathbf{w}^*|\mathbf{e}) = \overset{\max}{\mathbf{w}} P(\mathbf{w}|\mathbf{e}) \tag{4}$$

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:

$$BEL^*(x) = \overset{\max}{\mathbf{w}'_X} P(x, \mathbf{w}'_X|\mathbf{e}) \tag{5}$$

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

3.2 Tracking by Inference

The BBN for tracking hands is shown in Figure 7. Abbreviations are: LH = left hand, RH = right hand, LS = left shoulder, RS = right shoulder, B1 = skin cluster 1, B2 = skin cluster 2. There are 19 variables, $\mathbf{W} = \{X_1, X_2, \ldots, X_{19}\}$. The first point to note is that some of the variables are conceptual, namely X_1, X_2, X_5 and X_9, while the remaining variables correspond to image-measurable quantities, $\mathbf{e} = \{X_3, X_4, X_6, X_7, X_8, X_{10}, \ldots, X_{19}\}$. 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. At each time step, all of the measurement variables are instantiated from

[1] Regarding notation, upper-case is used to denote a random variable, lower-case to denote its instantiation, and boldface is used to represent sets of variables.

[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.

observations. B1 and B2 refer to the two largest skin clusters in the image (apart from the head), obtained as per Section 2.3. Absence of clusters is handled by setting the variables X_5 and X_9 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:

$$P(\mathbf{w}^*|\{x_3, x_4, x_6, x_7, x_8, x_{10}, \ldots, x_{19}\}) = \overset{\max}{\mathbf{w}} P(\mathbf{w}|\{x_3, x_4, x_6, x_7, x_8, x_{10}, \ldots, x_{19}\}) \quad (6)$$

This yields the most likely joint values of X_1 and X_2, which can be used to set the left and hand box position.

Note that the network structure is not singly connected, due to the loops formed through X_1 and X_2. Consequently the simple belief revision algorithm of Pearl [21] cannot be used due to non-convergence. Instead, we apply the more general inference algorithm of Lauritzen and Spiegelhalter [15, 8]. 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 state space size amongst the cliques. The variables and their dependencies are now explained as follows.

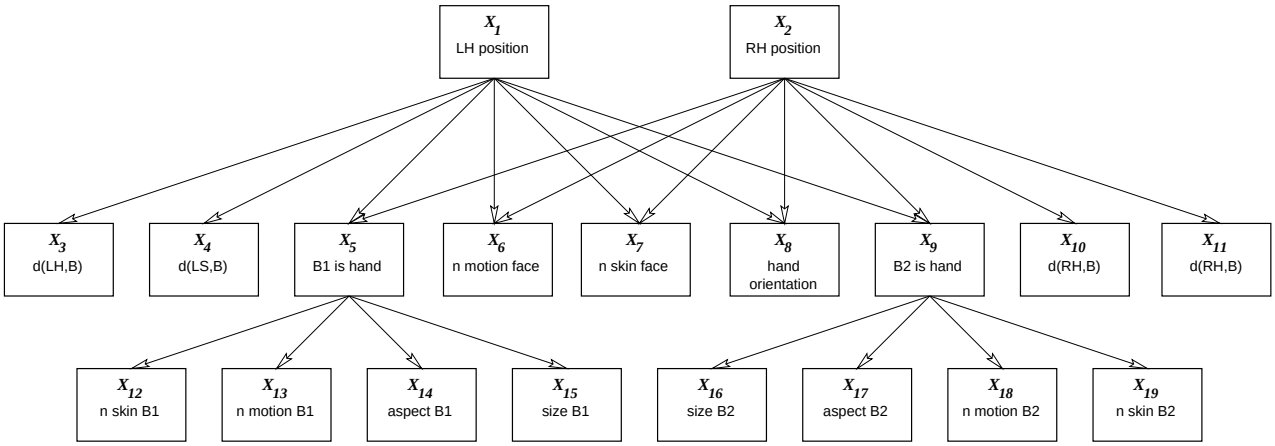


Fig. 7. A Bayesian Belief Network representing dependencies amongst variables in the human body-parts tracking scenario.

X_1 **and** X_2**:** 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. Note that disappearance of the hands is not modelled here for simplicity.

X_3**;** X_{10}**:** the distance in pixels of the previous left/right-hand box position from the currently hypothesised 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_4**;** X_{11}**:** the distance in pixels of the hypothesised cluster from the left/right shoulder. The shoulder position is estimated from the tracked head box. This dependency specifies that the hypothesised cluster should lie within a certain distance of the shoulder as defined by the length of the arm.

X_5, X_{12}, X_{13}, X_{14}, X_{15}; X_9, X_{16}, X_{17}, X_{18}, X_{19}**:** these variables determine whether each cluster is a hand. X_5 and X_9 are boolean variables specifying whether or not their respective clusters are hands or noise. The variables have an obvious dependency on X_1 and X_2: if either hand is a cluster, then that cluster must be a hand. The descendants of X_5 and X_9 provide evidence that the clusters are hands. X_{12} and X_{19} are the number of skin pixels in each cluster, which have some distribution depending on whether or not the cluster is a hand. X_{13} 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_{14} 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_{15} 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 exclusion box respectively. If either of the hands is hypothesised to occlude the head, we expect more skin pixels and some motion.

X_8**:** orientation of the respective hand, which depends to some extent on its spatial position in the screen relative to the head box. This orientation is calculated for each hypothesised hand position, and the histogram described in Section 2.1 is used to assign a conditional probability.

Under this framework, all of the visual cues can be considered *simultaneously* and consistently to arrive at a most probable explanation for the positions of both hands. BBNs lend the benefit of being able to "explain away" evidence, which can be of use in our network. 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_1 and X_2 and the fixed amount of probability attributable to any single piece of evidence. Hence probabilities are consistent and evidence is not "double counted" [21].

4 Experimental Evaluation

An experimental evaluation of the atemporal belief-based tracker is now presented. First, examples of the tracker's behaviour are given, then a comparison is performed between the BBN tracker and two other tracking methods. Note that to make our point about the difficulty of discontinuous motion more poignant, we captured all video data at a relatively high frame rate of 18 fps and used off-line processing.

4.1 Tracker Performance Examples

Selected frames from four different video sequences consisting of 141 to 367 frames per sequence are shown in Figure 8. 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, Figure 8(a), the hands are accurately tracked before, during and after mutual occlusion. In Figure 8(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 can be encountered. Although the frame rate was high, this discontinuity came about due to disk swapping during video capture. Nevertheless the tracker was able to correctly follow the hands. In Figure 8(c) the subject undergoes significant whole body motion to ensure that the tracker works while the head is constantly moving. With the hands alternately occluding each other and the face in a tumbling action, the tracker is still able to follow the body parts. In the third-to-last frame both hands simultaneously occlude the face. The example of Figure 8(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 tracker, occlusion with the head or the other hand is deduced. After these periods of disappearance, the hand is once again accurately tracked.

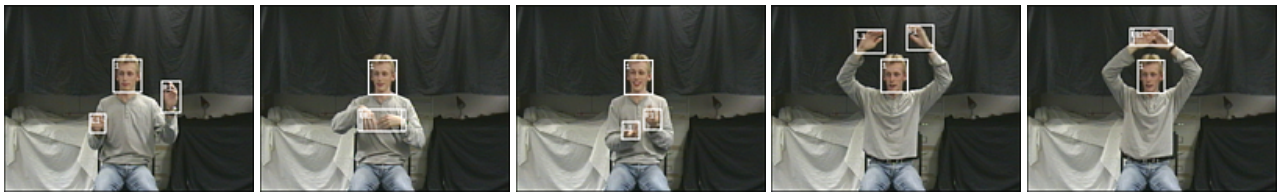
4.2 Comparison with Dynamic and Non-Contextual Trackers

We compared the atemporal belief-based tracker experimentally with two other tracking methods:

dynamic: 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 belief-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.

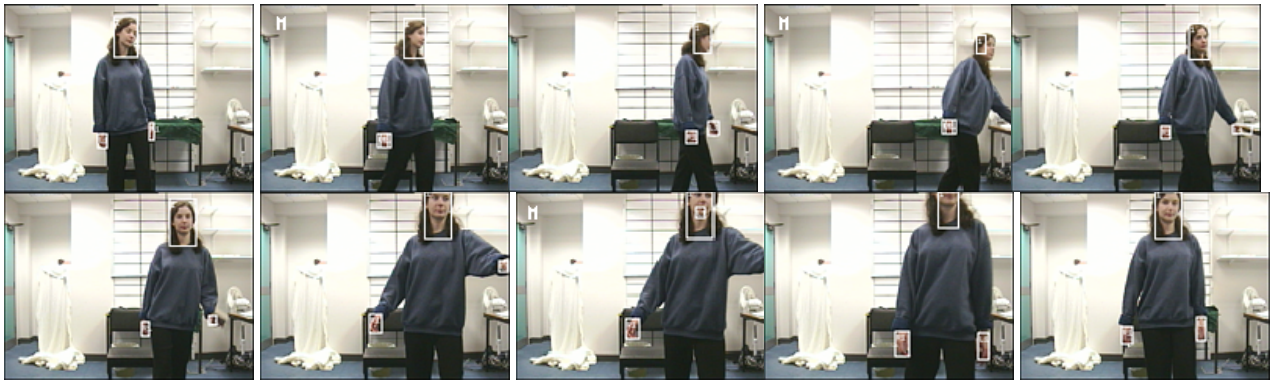
It is difficult to compare the tracking methods fairly in this context. Comparison of the average deviation from the true hand and head positions would be misleading because of the all-or-nothing nature of matching to discrete clusters. Another possible criterion is the number of frames until loss-of-track, 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 14 sequences containing two different people totalling 3300 frames.



(a)

(b)

(c)



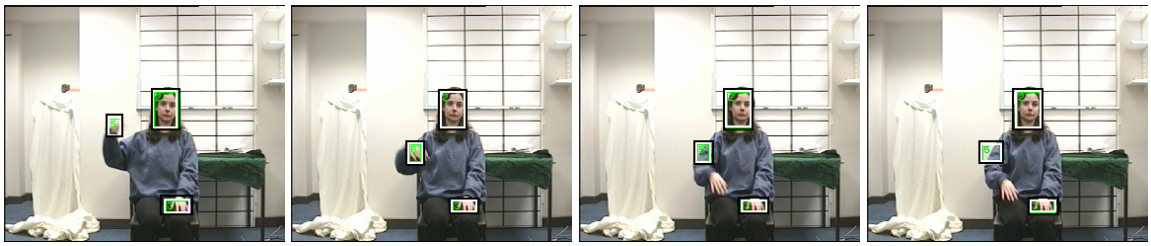
(d)

Fig. 8. Examples of discontinuous motion tracking.

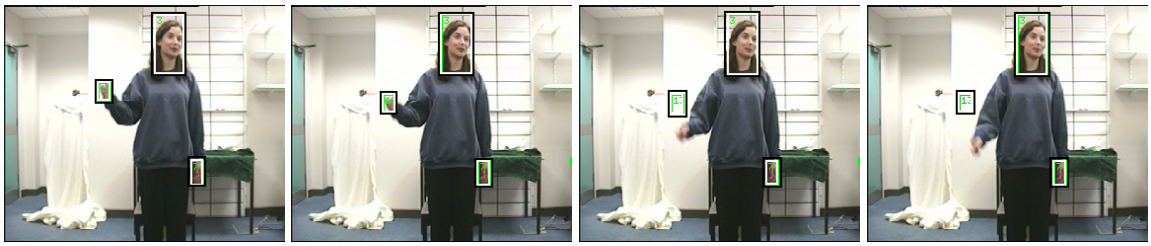
Table 1 shows the number of frames incorrectly tracked by each method, in absolute terms and as a percentage of the total number of frames. The belief-based tracker performs significantly better than the other two methods, even though the data was captured at a high frame rate. Therefore the benefits of using contextual knowledge to track discontinuous motion by inference rather than temporal continuity are significant. One would expect even better improvements if low frame-rate data were used. The most common failure modes for the belief-based and non-contextual trackers were incorrect assignment of the left and right hands to clusters, and locking on to background noise when one hand was occluded. The dynamic tracker often failed due to inaccurate temporal prediction of the hand position. Two examples of this failure are shown in consecutive frames in Figure 9. Although one could use more sophisticated dynamic models, it is 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 [25] switches in appropriate high-level models of behaviour for improved tracking, but the computational cost increases with the number of possible behaviours modelled. In terms of processing speed, all trackers had approximately the same performance. The average frame rate was about 4 fps on a PII 330 using 320x240 images, but with profiling the speed can be improved significantly.

method	**incorrect frames**	
	number	**%**
belief-based	439	13
dynamic	728	22
non-contextual	995	30

Table 1. Comparative results of the three tracking methods.



(a)



(b)

Fig. 9. Two examples of the failure of the dynamic Kalman filter tracker.

5 Conclusion

Observations of body motion in real-time systems can often be jerky and discontinuous. Contextual knowledge can be used to overcome ambiguities and uncertainties in measurement. We have presented a method for

tracking discontinuous motion of multiple occluding body parts of an individual from a single 2D view. Rather than modelling spatio-temporal dynamics, tracking is performed by reasoning about the observations using a Bayesian Belief Network. The BBN framework performs bottom-up and top-down message passing to fuse both conceptual and sensor-level quantities in a consistent manner. Hence the visual cues of skin colour, image motion and local intensity orientation are fused with contextual knowledge of the human body. The inference-based tracker was tested and compared with dynamic and non-contextual approaches. The results indicate that fusion of all available information at all levels significantly improves the robustness and consistency of tracking.

We wish to extend this work in two ways. First, the tracker can be made adaptive so that no parameters need to be changed when different people are tracked. Second, the current tracker assumes that there is only one person in the field of view, but we wish to use the tracker in scenes containing several people. We will investigate how trackers can be instantiated as people enter the scene, and how the tracker networks can be causally coupled so that skin clusters can be explained away by one network and not considered by the other networks. We intend to utilise the tracker as a component in a larger system for modelling the communicative behaviour of multiple people simultaneously. High-level interpretative information about behaviours can be fed back to the inference tracker as further evidence.

References

1. Yusuf Azoz, Lalitha Devi, and Rajeev Sharma. Tracking hand dynamics in unconstrained environments. In *Proc. 3rd IEEE Int. Conf. on Automatic Face and Gesture Recognition*, pages 247–279, Nara, Japan, 1998.
2. Andrew Blake and Michael Isard. *Active Contours*. Springer-Verlag, 1998.
3. Gary R. Bradski. Computer vision face tracking for use in a perceptual user interface. *Intel Technology Journal*, 2nd Quarter, 1998.
4. Hilary Buxton and Shaogang Gong. Visual surveillance in a dynamic and uncertain world. *Artificial Intelligence*, 78:431–459, 1995.
5. Tat-Jen Cham and James Rehg. Dynamic feature ordering for efficient registration. In *IEEE International Conference on Computer Vision*, volume 2, pages 1084–1091, Corfu, Greece, September 1999.
6. Eugene Charniak. Bayesian networks without tears. *AI Magazine*, 12(4):50–63, 1991.
7. James J. Clark and Alan L. Yuille. *Data Fusion for Sensory Information Processing Systems*. Kluwer Academic Publishers, 1990.
8. Robert G. Cowell, A. Philip Dawid, Steffen L. Lauritzen, and David J. Spiegelhalter. *Probabilistic Networks and Expert Systems*. Springer-Verlag, NY, 1999.
9. D. M. Gavrila. The visual analysis of human movement - a survey. *Computer Vision and Image Understanding*, 73(1), 1999.
10. D.M. Gavrila and L.S. Davis. 3-D model-based tracking of human motion in action. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 73–80, 1996.
11. Kazuyuki Imagawa, Shan Lu, and Seiji Igi. Color-based hands tracking system for sign language recognition. In *Proc. 3rd IEEE Int. Conf. on Automatic Face and Gesture Recognition*, pages 462–467, Nara, Japan, 1998.
12. Michael Isard and Andrew Blake. CONDENSATION – conditional density propagation for visual tracking. *International Journal of Computer Vision*, 29(1):5–28, 1998.
13. Cullen Jennings. Robust finger tracking with multiple cameras. In *Proceedings of the International Workshop on Recognition, Analysis and Tracking of Faces and Gestures in Real-Time Systems*, pages 152–160, Corfu, Greece, September 1999. IEEE Computer Society.
14. Nebojsa Jojic, Matthew Turk, and Thomas Huang. Tracking self-occluding articulated objects in dense disparity maps. In *IEEE International Conference on Computer Vision*, volume 1, pages 123–130, Corfu, Greece, September 1999.
15. Steffen L. Lauritzen and David J. Spiegelhalter. Local computations with probabilities on graphical structures and their application to expert systems. In Glenn Shafer and Judea Pearl, editors, *Readings in Uncertain Reasoning*, pages 415–448. Morgan Kauffmann, CA, 1990.
16. Jérôme Martin, Vincent Devin, and James Crowley. Active hand tracking. In *Proc. 3rd IEEE Int. Conf. on Automatic Face and Gesture Recognition*, pages 573–578, Nara, Japan, 1998.
17. Dimitris Metaxas. Deformable model and HMM-based tracking, analysis and recognition of gestures and faces. In *Proceedings of the International Workshop on Recognition, Analysis and Tracking of Faces and Gestures in Real-Time Systems*, pages 136–140, Corfu, Greece, September 1999. IEEE Computer Society.
18. Darnell J. Moore, Irfan A. Essa, and Monson H. Hayes III. Exploiting human actions and object context for recognition tasks. In *IEEE International Conference on Computer Vision*, volume 1, pages 80–86, Corfu, Greece, September 1999.
19. Eng-Jon Ong and Shaogang Gong. A dynamic human model using hybrid 2D-3D representations in hierarchical PCA space. In *British Machine Vision Conference*, volume 1, pages 33–42, Nottingham, UK, September 1999. BMVA.

20. E. Osuna, R. Freund, and F. Girosi. Training support vector machines: An application to face detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 130–136, 1997.
21. Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, San Mateo, CA, 1988.
22. Y. Raja, S. J. McKenna, and S. Gong. Tracking and segmenting people in varying lighting conditions using colour. In *Proc. 3rd IEEE Int. Conf. on Automatic Face and Gesture Recognition*, pages 228–233, Nara, Japan, 1998.
23. Herbert Simon. Rational choice and the structure of the environment. *Psychological Review*, 63:129–138, 1956.
24. Christopher Wren, Ali Azarbayejani, Trevor Darrell, and Alex Pentland. Pfinder: Real-time tracking of the human body. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):780–785, July 1997.
25. Christopher Wren and Alex Pentland. Understanding purposeful human motion. In *IEEE International Workshop on Modelling People*, pages 19–25, Corfu, Greece, September 1999.