Multimedia Systems 10: 164–179 (2004)
Digital Object Identifier (DOI) 10.1007/s00530-004-0148-1

Multimedia Systems

A hierarchical Bayesian network for event recognition of human actions and interactions

Sangho Park, J.K. Aggarwal

Department of Electrical and Computer Engineering, The University of Texas at Austin, Austin, TX 78712, USA
(e-mail: sangho@ece.utexas.edu, aggarwaljk@mail.utexas.edu)

Published online: 25 October 2004 –

Abstract. Recognizing human interactions is a challenging task due to the multiple body parts of interacting persons and the concomitant occlusions. This paper presents a method for the recognition of two-person interactions using a hierarchical Bayesian network (BN). The poses of simultaneously tracked body parts are estimated at the low level of the BN, and the overall body pose is estimated at the high level of the BN. The evolution of the poses of the multiple body parts are processed by a dynamic Bayesian network (DBN). The recognition of two-person interactions is expressed in terms of semantic verbal descriptions at multiple levels: individual body-part motions at low level, single-person actions at middle level, and two-person interactions at high level. Example sequences of interacting persons illustrate the success of the proposed framework.

Keywords: Surveillance – Event recognition – Human interaction – Motion – Bayesian network

1 Introduction

The recognition of human interaction has many applications in video surveillance, video-event annotation, virtual reality, human-computer interaction, and robotics. Recognizing human interactions, however, is a challenging task due to the ambiguity caused by body articulation, loose clothing, and mutual occlusion between body parts. This ambiguity makes it difficult to track moving body parts and to recognize their interaction. The recognition task depends on the reliable performance of low-level vision algorithms that include segmentation and tracking of salient image regions and extraction of object features. Involving more than one person makes the task more complicated since the individual tracking of multiple interacting body parts needs to be maintained along the image sequence.

In our previous paper [19], we presented a method to segment and track multiple body parts in two-person interactions. Our method is based on multilevel processing at pixel level, blob level, and object level. At the pixel level, individual pixels are classified into homogeneous blobs according to color. At the blob level, adjacent blobs are merged to form large blobs according to a blob similarity metric. At the object level, sets of multiple blobs are labeled as human body-part regions according to domain knowledge. The multiple body-part regions are tracked along the image sequence. As shown in Fig. 1, the body parts lack information about their poses, such as the orientation of the head, the hand position of the upper body, the foot position of the lower body, etc.

a **b**

Fig. 1. An example frame of the "hugging" sequence: the input image (**a**) and its tracked body parts indexed by different colors (**b**)

In this paper, we present a methodology that estimates body-part pose and recognizes different two-person interactions including pointing, punching, standing hand-in-hand, pushing, and hugging. The recognition algorithm is preceded by a feature extraction algorithm that extracts body-pose features from the segmented and tracked body-part regions. We use ellipses and convex hulls to represent body parts and build a hierarchical Bayesian network to estimate individual body poses at each frame. Including the dynamics of the body-pose changes along the sequence leads us to the recognition of two-person interactions.

Figure 2 shows the overall system diagram. Our system processes multiple levels of analysis. Body-part features about the ellipses and convex hulls are extracted from our already developed segmentation and tracking system. With the body features, we estimate body poses using a Bayesian network. The pose estimation results are then concatenated to form a sequence, and sequence classification is performed by a dynamic Bayesian network. Then we generate a user-friendly verbal semantic description of the interaction. In the body-pose-estimation step, the left panel shows a tree structure that represents the individual body part regions of a person; the root

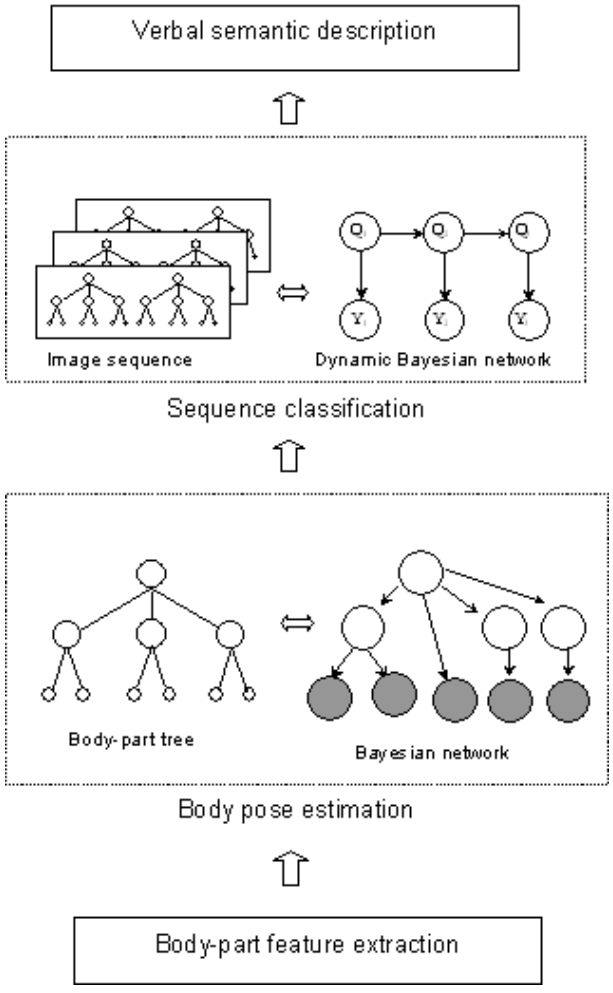


Fig. 2. System diagram

node represents the whole body region, and its children nodes represent head, upper body, and lower body, respectively. Each body part is subdivided into skin and nonskin parts such as face and hair, hand and torso, etc. The right panel shows a Bayesian network to estimate the overall body pose in each frame including the poses of torso, head, arm, and leg. In the sequence classification stage, the left panel shows the image sequence represented by the concatenation of two tree structures for two interacting persons at each frame. The right panel shows a dynamic Bayesian network that recognizes the two-person interactions. The result of recognition of the two-person interactions is provided in terms of semantic description with "subject + verb + object" format for a user-friendly verbal interface.

The contributions of this paper are as follows. (1) A new hierarchical framework is proposed for the recognition of two-person interactions at a detailed level using a hierarchical Bayesian network. (2) Ambiguity in human interaction due to occlusion is handled by inference with the Bayesian network. (3) A human-friendly vocabulary is generated for high-level event description. (4) A stochastic graphical model is proposed for the recognition of two-person interactions in terms of "subject + verb + object" semantics.

The rest of the paper is organized as follows. Section 2 summarizes research related to the representation of human body and the recognition paradigms for event recognition. Section 3 presents the formulation of hierarchical Bayesian network and the extraction of observed features. Section 4 shows the estimation of the individual body-part poses as well as the overall body pose. Section 5 presents the method of recognizing sequences using dynamic Bayesian network and verbal description of events. Section 6 addresses the need to process relative information about constraints between interacting persons. Results and conclusions follow in Sects. 7 and 8, respectively.

2 Previous work

In the last decade, the computer vision community has conducted extensive research on the analysis of human motion in general. The many approaches that have been developed for the analysis of human motion can be classified into two categories: model-based [4] and appearance-based [20]. See [1,8,15] for reviews. We will discuss appearance-based methods below.

Early research on human motion has focused on analyzing a single person in isolation [3,24] or on tracking only a subset of body parts such as head, torso, and hands [7,26], while research on analyzing the motion of multiple people has focused on silhouettes [10,16,25], stick figures [6], contours [18,28], or color [19].

In addition, research has used a coarse-level representation of the human body such as a moving bounding box or silhouette or has focused on the detection of specific interactions predefined in terms of head or hand velocity. However, a coarse-level representation of a human body using a bounding box is not powerful enough for detailed recognition of human interactions involving body parts. Silhouette-based or contour-based representation may be hampered by mutual occlusion caused by moving body parts in human interactions. Model-based representation is not robust due to ambiguities in body shape caused by loose clothing.

Many approaches have been proposed for behavior recognition using various methods including hidden Markov models, finite state automata, context-free grammar, etc. Oliver et al. [16] presented a coupled hidden Markov model (CHMM) for gross-level human interactions between two persons such as approach, meet, walk together, and change direction from bird's-eye-view sequences. Sato et al. [25] presented a method to use the spatiotemporal velocity of pedestrians to classify their interaction patterns. Hongeng et al. [11] proposed probabilistic finite state automata (FA) for gross-level human interactions. Their system utilized user-defined hierarchical multiple scenarios of human interaction. Park et al. [18] proposed a string-matching method using a nearest-neighbor classifier for detailed-level recognition of two-person interactions such as hand-shaking, pointing, and standing hand-in-hand. Wada et al. [29] used nondeterministic finite state automata (NFA) using state product space. Kojima et al. [14] presented a natural-language-based description of single-person activities. Park et al. [21] presented a recognition method that combines model-based tracking and deterministic finite state automata.

3 Hierarchical Bayesian network

3.1 Inference in a Bayesian network

Understanding events in video requires a representation of causality among random variables. Causal relations and conditional dependencies among the random variables are efficiently represented by a directed acyclic graph (DAG) in a Bayesian Network (BN) [5,12,13,22,27]. The network is composed of nodes and directed links. The nodes represent random variables, whereas the directed links point from causal to dependent variables. The conditioning variables and the dependent variables are called parent nodes and child nodes, respectively. For a link between two variables, $X \longrightarrow Y$, the overall joint distribution is specified by the product of the prior probability $P(X)$ and the conditional probability $P(Y \mid X)$. The dependencies are specified a priori and used to create the network structure. The distributions $P(x)$ and $P(y \mid x)$ must be specified beforehand to form the network from domain knowledge.

In general, given a set of N variables $H_{1:N} = H_1, \ldots, H_N$, the joint probability distribution $P(H_{1:N}) = P(H_1, H_2, \ldots, H_N)$ can be factored into a sparse set of conditional probabilities as follows according to the conditional independency:

$$P(H_{1:N}) = \prod_{i=1}^{N} P(H_i \mid pa(H_i)) \,, \tag{1}$$

where $pa(H_i)$ is the set of parent nodes of node H_i in the DAG.

A video sequence of human interactions will exhibit occlusion caused by articulated body parts. Occlusion may result in incomplete values for the random variables. Occlusion problems caused by human interaction are well suited to the BN, which can perform inference with a partially observed set e of variables referred to as "evidence". With the evidence provided, the set of beliefs is established and updated and established to reflect both prior and observed information depending on the evidence:

$$B(h) = P(h \mid e) \,, \tag{2}$$

where $B(h)$ is the belief in the value h of variable H given the evidence e.

The most probable explanation $h^\star$ of the hidden variable H given the evidence e is determined by:

$$\begin{aligned} h^\star &= \arg\max_h B(h) \\ &= \arg\max_h P(h \mid e) \,. \end{aligned} \tag{3}$$

The belief update is achieved by a message-passing process distributed along the network through the exploitation of local dependencies and global independencies of various variables. Therefore, the estimation of evidence to update the beliefs is performed to achieve a globally consistent evaluation of the situation.

We use the junction-tree algorithm [12, 13] for the inference of belief revision given evidence. This inference algorithm transforms the Bayesian network to a *join tree*, each node of which contains a subset of variables called a *clique*. The locally dependent joint probabilities of the network are represented by the clique potentials, and the global independency in the network is insured by distinct cliques in the join tree. The transformation to the join tree is performed only once offline, and the inference proceeds on the join tree via a message passing mechanism similar to the method proposed by Pearl [22].

3.2 Feature extraction for a Bayesian network

Understanding semantic events from video requires the incorporation of domain knowledge combined with image data. A human body model is introduced as the domain knowledge to group free-form blobs into meaningful human body parts: head, upper body, and lower body, each of which is recursively subdivided into skin and nonskin parts. The body model is appearance-based and represents image regions corresponding to each body part.

We need body-part features such as arm tip, leg tip, torso width, etc. to estimate the body poses. In this paper we propose a method to model human body parts by combining an ellipse representation and a convex hull-based polygonal representation of the interacting human body parts.

An ellipse is defined by the major and minor axes and their orientations in a two-dimensional (2D) space and optimally represents the dispersedness of an arbitrary 2D data (i.e., an image region in our case.) The direction and length of the major and minor axes are computed by principal component analysis (PCA) of the 2D data; the two eigenvectors of the image region define the center positions and directions of the axes, and the eigenvalues define the lengths of the axes. Suppose image pixel $X = [x_1, x_2]^T$ is distributed as $N_2(\mu, \Sigma)$. The density of X is described by an ellipsoid centered at μ as follows:

$$(x - \mu)^T \Sigma (x - \mu) = c^2 \,. \tag{4}$$

The major and minor directions $e = [e_a, e_b]$ and their extensions $\lambda = diag(\lambda_1, \lambda_2)$ of the ellipsoid are obtained by:

$$\Sigma e = \lambda e \,. \tag{5}$$

The principal axes $y = [y_a, y_b]$ of the ellipsoid are determined by:

$$y_a = e_a^T x \,, \tag{6}$$

$$y_b = e_b^T x \,. \tag{7}$$

A convex hull is defined as the minimum convex closure of an arbitrary image region specified by a set of vertex points and represents the minimum convex area containing the region. We use the set of contour points of the individual body-part region segmented in Fig. 1b to generate the convex hull of the corresponding body part. There are many efficient algorithms to compute convex hull. See [17] for details. We use Graham's algorithm [9].

The maximum curvature points of the convex hull indicate candidate positions of limb tips such as hand or foot and are used as extra observation evidence for the BN.

Note that the ellipse and the convex hull representations have tradeoffs. The advantage of the ellipse representation is its robustness against outlier pixels due to image noise; however, it is not sensitive enough to detect a hand position (i.e., the arm tip) located at an extreme position of the upper-body distribution, resulting in false negative detection. In contrast, the convex hull representation of the upper body effectively detects a candidate arm tip as one of the maximum curvature

a **b**

Fig. 3. Body-part parameterization for Fig. 1b. Each body part is parameterized by both an ellipse (**a**) and a convex hull (**b**)

points of the convex hull. However, the convex hull representation is prone to false positive detection of the arm tip when the hand is nearer than the elbow to the trunk or when a thin long region is attached to the upper body due to image noise. In order to cope with the problem, we choose as the candidate hand positions the maximum curvature points that coincide with a skin blob. We use the skin detector described in [19].

We introduce a hierarchical Bayesian network that infers body pose. The Bayesian network estimates the configuration of the head, hand, leg, and torso in a hierarchical way by utilizing partial evidence (i.e., the ellipse parameters and convex hull parameters) observed in each frame. Our method uses color- and region-based segmentation of body parts and is effective in handling mutual occlusion. Mutual occlusion inevitably involves uncertainty in estimation of body pose.

All the observation node states are initially continuous, but they are normalized and discretized to generate codebook vectors by vector quantization, as follows:

A freely moving human figure is captured by a camera and the image is divided horizontally and vertically into a grid of L histogram bins as shown in Fig. 4. The individual configuration of the head, hand, leg, and torso is recorded in terms of a random variable v in each dimension of the image coordinates. Let $v \in R$ be a random variable with a uniform prior probability distribution in the range $[v_a, v_b]$. Assume that $V = \{1, \ldots, L\}$ is the discretized space of R and has L clusters. Then v is converted to a codebook vector $v_k \in V$ by multiple thresholding as follows:

$$v_k = \arg\min_i \left\{ \frac{v - v_a}{v_b - v_a} - \frac{i}{L} \right\} \quad \text{for all } i \in [1, L]\,. \tag{8}$$

The frequency in each histogram bin for an individual body part is accumulated along the overall sequences of the training data. The histogram process is illustrated in Fig. 4.

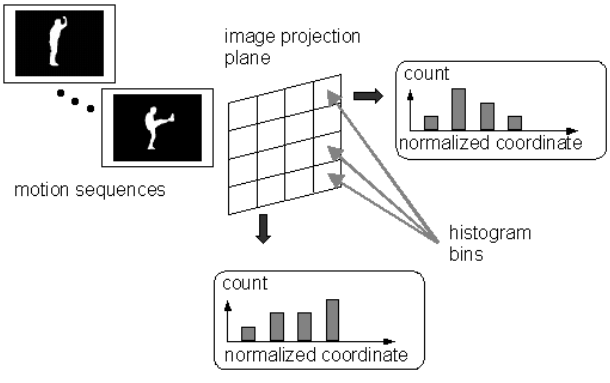


Fig. 4. Diagram of histogram process

The vector quantization leads to a Bayesian network with discrete hidden and observation nodes. The discrete version of the BN uses conditional probability tables as an approximation of conditional probability distributions. We assume that the conditional probability distribution of each of the observation variables is a triangular distribution centered at the most probable state. The most probable state is determined by training data.

A two-person interaction is represented in terms of the configuration changes of the individual body parts and is recognized by a dynamic Bayesian network (DBN) that utilizes the hierarchical information of the body pose. The recognition results are expressed in terms of semantic motion descriptions. Example sequences illustrate the success of the proposed framework. The body parts are represented by ellipses and convex hulls (Fig. 3) and then processed by a hierarchical Bayesian network. The hierarchical Bayesian network estimates the configuration of the head, hand, leg, and torso by utilizing partial evidence observed in the frame, analyzes temporal changes of the body poses along the sequence, and interprets human interactions.

4 Pose estimation using a hierarchical Bayesian network

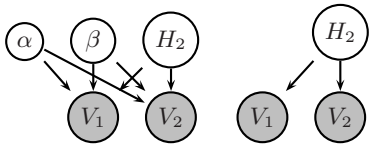
4.1 Head-pose estimation

A Bayesian network for estimating the head pose of a person is constructed as shown in Fig. 5a with a geometric transformation parameter α, an environmental setup parameter β, head pose H_2 as a hidden node, and head appearance V_1 and V_2 as observation nodes. The geometric transformation, which may include camera geometry, determines the imaging process. The environmental setup, which may include lighting conditions, determines the reflectance of light from the head. The head pose, which may include the head's three-dimensional rotation angles, determines which part of the head is visible.

The legend for Fig. 5 is as follows:

α: Geometric transformation parameter such as camera model
β: Environmental setup parameter such as lighting condition
H_2: Head pose
V_1: Index for angle of vector from head center to face center
V_2: Index for ratio of face area to head area

An example of the observations (i.e., evidence) of the head appearance in Fig. 3a is shown in Fig. 6. For each head, the large ellipse represents the overall head region and the small ellipse represents the face region. Arrows represent the vector from the head center to the face center. The angle of the vector corresponds to the value of observation node V_1, and the



a **b**

Fig. 5. Bayesian network for head-pose estimation of a person. See legend below for node labels

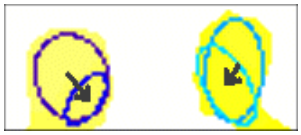


Fig. 6. Observations for head pose estimation. For each head, the *large ellipse* represents the overall head region and the *small ellipse* represents the face region. *Arrows* represent the vector from head center to face center

ratio of the two ellipses in a head corresponds to the value of observation node V_2 in Fig. 5.

The observations of the head appearance are determined by the generative model parameters α, β, and H_2. The causal relations between the generative model parameters and observations are represented by the arrows between the hidden nodes and the observation nodes in Fig. 5.

In this paper we assume, for simplicity, that the geometric transformation parameter α and the environmental setup parameter β are fixed and known. This assumption is based on our experience that the detailed recognition of human interactions requires that body-part information be available or able to be inferred from image data. That is, we assume that a fixed camera is used with a viewing axis parallel to the ground (i.e., horizon) and that the people are in an indoor setting with constant lighting conditions. (Note that this assumption may be relaxed by learning the parameters α and β, with enough training data.) Currently, this assumption allows the nodes α and β to be omitted from computation and simplifies the structure of the Bayesian network, as shown in Fig. 5b.

The hidden node's states are defined as

H_2 = {front view, left view, right view, rear view}.

The visible nodes' state labels are defined as follows:

V_1 = {$-45°$, $-90°$, $-135°$, null}
V_2 = {0, 1/2, 1}

where null angle in V_1 and 0 ratio in V_2 indicate that no face ellipse was detected by the occlusion.

The joint probability of the BN in Fig. 5b is factored into conditional probabilities and prior probabilities according to (1) as follows:

$$P(V_{1:2}, H_2) = P(V_{1:2}|H_2)P(H_2) = P(V_1|H_2)P(V_2|H_2)P(H_2) \quad (9)$$

Our task is to estimate the *belief* of the state of the hidden node H_2 given the evidence $V_{1:2}$:

$$P(H_2|V_{1:2}) = \frac{P(V_{1:2}, H_2)}{P(V_{1:2})}, \quad (10)$$

$$= \frac{P(V_{1:2}, H_2)}{\sum_{all H_{2,n}} P(V_{1:2}, H_2)}, \quad (11)$$

$$= \frac{P(V_1|H_2)P(V_2|H_2)P(H_2)}{\sum_{all H_{2,n}} P(V_1|H_2)P(V_2|H_2)P(H_2)},$$

where the summation is over all possible configurations of values on the parent node H_2. Here $H_{2,n}$ denotes a particular value for state n on node H_2.

The factors of Eq. 10, i.e., the conditional probability table (CPT) and the prior probability table (PPT) for Fig. 5b, are trained by the histogram procedure in Fig. 4 and specified in Tables 1–3.

Table 1. Conditional probability table (CPT) for $P(V_1|H_2)$

	$P(V_1\|H_2)$			
	A	B	C	D
A	0.11	0.09	0.7	0.1
B	0.05	0.72	0.18	0.05
C	0.05	0.05	0.21	0.69
D	0.8	0.1	0.02	0.08

Table 2. CPT for $P(V_2|H_2)$

	$P(V_2\|H_2)$		
	A	B	C
A	0.12	0.08	0.8
B	0.06	0.8	0.14
C	0.03	0.8	0.17
D	0.78	0.12	0.1

Table 3. Prior probability table (PPT) for $P(H_2)$

$P(H_2)$			
A	B	C	D
0.25	0.25	0.25	0.25

Each row of the tables in this paper represents a state of a conditional variable expressed in terms of the alphabetical index (i.e., A, B, C, etc.), and each column of the tables represent a state of a dependent variable expressed in terms of another alphabetical index. For example, in Table 1 the cell value 0.18 for row B and column C represents $P(V_1 = C|H_2 = B) = 0.18$. Here, the distinct indices B and C of the distinct variables V_1 and H_2 are independent of each other. For example, given that we are in a certain head pose (i.e., at a certain row), we must have some observation (i.e., at a certain column). Each row in a CPT corresponds to a cause and each column to an effect; thus, each row sums to 1.

4.2 Arm pose estimation

Another Bayesian network is built for estimating the pose of the salient arm. The *salient* arm is defined to be the arm that is involved in human interaction. Usually, the salient arm is the outermost arm stretching toward the opposite person. The BN for the arm pose is shown in Fig. 7.

The geometric transformation parameter α and the environment-setup parameter β are dropped from computation, as in Sect. 4.1.

The legend for Fig. 7 is as follows:

H_3: Vertical pose of outermost arm

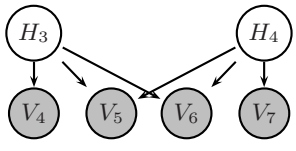
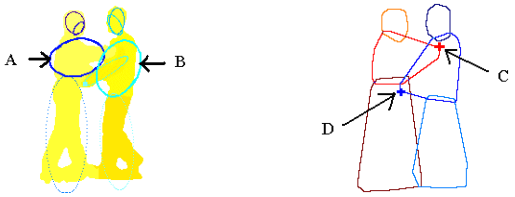


Fig. 7. Bayesian network for arm-pose estimation of a person. See legend below for node labels



a **b**

Fig. 8. Observation for arm-pose estimation. The individual upper body is represented by *ellipse* A or B in **a** and the corresponding *convex hull* in **b**. The maximum curvature points C and D in **b** are detected as candidate hand positions

H_4: Horizontal pose of outermost arm
V_4: Index for the vertical position of maximum curvature point of upper-body convex hull
V_5: Index for torso ellipse aspect ratio of major to minor axes
V_6: Index for torso ellipse rotation on image plane
V_7: Index for horizontal position of maximum curvature point of upper-body convex hull

An example of the observation data for estimating the arm pose is shown in Fig. 8. The ellipse and the convex hull representations of the torso region are used for each person.

Ellipse A in Fig. 8a represents the degree of spatial distribution of the upper-body region of the left person and is summarized in terms of the major and minor axes and their orientations computed from principal component analysis (PCA). The torso ellipse's aspect ratio and rotation angle on the image plane are used as observation evidence V_5 and V_6 of the BN. A convex hull represents the maximal area of the region. The image coordinate of the maximum curvature point C of the convex hull of the left person in Fig. 8b is detected as a candidate hand position and is used as extra observation evidence V_4 and V_7 for the BN. The same computations apply to the right person in Fig. 8 with ellipse B and maximum curvature point D. A separate BN is used for the right person.

If multiple maximum curvature points exist for a candidate hand position, we choose the maximum curvature point that coincides with a skin blob detected by the skin detector described in [19]. However, the skin blob can be occluded by articulation, and it is not guaranteed that the maximum curvature point of a convex hull actually represents an arm tip. Our solution to this ambiguity is to estimate the arm pose based on the observations using the BN in Fig. 7.

The hidden node's states are defined below:

H_3 = {high, mid-high,mid-low, low}
H_4 = {withdrawn, intermediate, stretching}

The visible nodes' state labels are defined as follows:

V_4 = {0, 0.2, 0.35,0.5}
V_5 = {0.3, 0.6, 0.9}
V_6 = {30°, 60°, 90°}
V_7 = {0, 0.1, 0.15, 0.23}

Table 4. Conditional probability table for $P(V_4|H_3)$

	$P(V_4\|H_3)$			
	A	B	C	D
A	0.89	0.05	0.03	0.03
B	0.03	0.8	0.13	0.04
C	0.04	0.02	0.91	0.03
D	0.03	0.04	0.03	0.9

The joint probability of the BN in Fig. 7 is factored into conditional probabilities and prior probabilities as follows:

$$\begin{aligned} P(V_{4:7}, H_{3:4}) &= P(V_{4:7}|H_{3:4})P(H_{3:4})\,, \\ &= P(V_{4:7}|H_{3:4})P(H_3)P(H_4)\,, \\ &= P(V_4|H_{3:4})P(V_5|H_{3:4})P(V_6|H_{3:4})\,, \\ &\quad \times P(V_7|H_{3:4})P(H_3)P(H_4)\,, \\ &= P(V_4|H_3)P(V_5|H_{3:4})P(V_6|H_{3:4})\,, \\ &\quad \times P(V_7|H_4)P(H_3)P(H_4)\,. \end{aligned} \tag{12}$$

In this case our goal is to estimate the *belief* of the states of the hidden nodes $H_{3:4}$ given the evidence $V_{4:7}$:

$$P(H_{3:4}|V_{4:7}) = \frac{P(V_{4:7}, H_{3:4})}{P(V_{4:7})}\,, \tag{13}$$

$$= \frac{P(V_{4:7}, H_{3:4})}{\sum_{allH_{3,m}} \sum_{allH_{4,n}} P(V_{4:7}, H_{3:4})}\,, \tag{14}$$

$$= \frac{Numer_2}{Denom_2}\,,$$

where the numerator $Numer_2$ is

$$\begin{aligned} Numer_2 = {} & P(V_4|H_3)P(V_5|H_{3:4})P(V_6|H_{3:4}) \\ & \times P(V_7|H_4)P(H_3)P(H_4) \end{aligned} \tag{15}$$

and the denominator $Denom_2$ is

$$\begin{aligned} Denom_2 = {} & \sum_{allH_{3,m}} \sum_{allH_{4,n}} [P(V_4|H_3)P(V_5|H_{3:4}) \\ & \times P(V_6|H_{3:4})P(V_7|H_4)P(H_3)P(H_4)]\,. \end{aligned} \tag{16}$$

The summations are over all possible configurations of values on the parent nodes H_3 and H_4. Here $H_{3,m}$ denotes a particular value for state m on node H_3, and $H_{4,n}$ denotes a particular value for state n on node H_4.

The factors of Eq. 13, i.e., the conditional probability table (CPT) and the prior probability table (PPT) for Fig. 7, are specified in Tables 4–7. In the left panel of Table 5, the cell value 0.34 of row CA and column B represents $P(V_5 = B|H_3 = C, H_4 = A) = 0.34$.

4.3 Leg-pose estimation

A Bayesian network similar to that in Sect. 4.2 is built for estimating the leg pose of a person as shown in Fig. 9, with hidden nodes H_5 and H_6 and observation nodes $V_8 - V_{11}$.

The legend for Fig. 9 is as follows:

H_5: Vertical pose of outermost leg
H_6: Horizontal pose of outermost leg
V_8: Index for vertical position of maximum curvature point of lower-body convex hull

Table 5. CPT for $P(V_5|H_{3:4})$ and $P(V_6|H_{3:4})$

	$P(V_5\|H_{3:4})$		
	A	B	C
AA	0.45	0.35	0.2
BA	0.17	0.43	0.4
CA	0.26	0.34	0.4
DA	0.8	0.09	0.11
AB	0.3	0.5	0.2
BB	0.31	0.29	0.4
CB	0.24	0.35	0.41
DB	0.27	0.33	0.4
AC	0.31	0.25	0.44
BC	0.3	0.31	0.39
CC	0.31	0.28	0.41
DC	0.32	0.27	0.41

	$P(V_6\|H_{3:4})$		
	A	B	C
AA	0.45	0.33	0.22
BA	0.21	0.39	0.4
CA	0.33	0.27	0.4
DA	0.33	0.29	0.38
AB	0.31	0.3	0.39
BB	0.32	0.28	0.4
CB	0.31	0.3	0.39
DB	0.28	0.31	0.41
AC	0.31	0.31	0.38
BC	0.32	0.27	0.41
CC	0.25	0.34	0.41
DC	0.31	0.29	0.4

Table 6. CPT for $P(V_7|H_4)$

	$P(V_7\|H_4)$			
	A	B	C	D
A	0.39	0.41	0.11	0.09
B	0.12	0.43	0.37	0.08
C	0.1	0.1	0.42	0.38

Table 7. PPT for $P(H_3)$ and $P(H_4)$

$P(H_3)$			
A	B	C	D
0.2	0.1	0.3	0.4

$P(H_4)$		
A	B	C
0.6	0.25	0.15

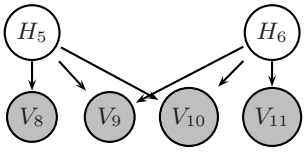


Fig. 9. Bayesian network for pose estimation of an interacting person. The BN is composed of 6 hidden nodes $H_{1:6}$ and 11 observation nodes $V_{1:11}$. See legend below for node labels

V_9: Index for lower-body ellipse aspect ratio of major to minor axes
V_{10}: Index for lower-body ellipse rotation on image plane
V_{11}: Index for horizontal position of maximum curvature point of lower-body convex hull

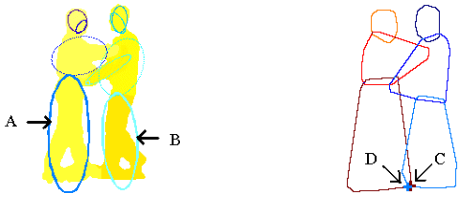
The hidden node's states are defined below:

H_5 = {low, middle, high}
H_6 = {withdrawn, intermediate, out-reached}

The visible nodes' state labels are defined as follows:

V_8 = {0.5, 0.75, 1}
V_9 = {0.3, 0.6, 0.9}
V_{10} = {30°, 60°, 90°}
V_{11} = {0, 0.15, 0.3}

The lower body of each person represented by an ellipse and a convex hull is shown in Fig. 10. The BN for leg pose has a similar structure to the BN for arm pose, but the prior probability distribution and the conditional probability distribution may differ significantly.



a **b**

Fig. 10. Observation for leg-pose estimation. The individual lower body is represented by *ellipse* A or B in **a** and corresponding *convex hull* in **b**. The maximum curvature *points* C and D in **b** are detected as candidate foot positions

Table 8. CPT for $P(V_8|H_5)$ and $P(V_{11}|H_6)$

	$P(V_8\|H_5)$		
	A	B	C
A	0.12	0.13	0.75
B	0.11	0.74	0.15
C	0.81	0.09	0.1

	$P(V_{11}\|H_6)$		
	A	B	C
A	0.7	0.2	0.1
B	0.1	0.6	0.3
C	0.09	0.21	0.7

The joint probability of the BN in Fig. 9 is factored into conditional probabilities and prior probabilities as follows:

$$\begin{aligned} P(V_{8:11}, H_{5:6}) &= P(V_{8:11}|H_{5:6})P(H_{5:6}) \\ &= P(V_{8:11}|H_{5:6})P(H_5)P(H_6) \\ &= P(V_8|H_{5:6})P(V_9|H_{5:6}) \\ &\quad \times P(V_{10}|H_{5:6})P(V_{11}|H_{5:6})P(H_5)P(H_6) \\ &= P(V_8|H_5)P(V_9|H_{5:6}) \\ &\quad \times P(V_{10}|H_{5:6})P(V_{11}|H_6)P(H_5)P(H_6) \end{aligned} \quad (17)$$

In this case our goal is to estimate the *belief* of the states of the hidden nodes $H_{5:6}$ given the evidence $V_{8:11}$:

$$\begin{aligned} P(H_{5:6}|V_{8:11}) &= \frac{P(V_{8:11}, H_{5:6})}{P(V_{8:11})} \\ &= \frac{P(V_{8:11}, H_{5:6})}{\sum_{allH_{5,m}} \sum_{allH_{6,n}} P(V_{8:11}, H_{5:6})} \\ &= \frac{Numer_3}{Denom_3}, \end{aligned} \quad (18)$$

where the numerator $Numer_3$ is

$$\begin{aligned} Numer_3 = P(V_8|H_5)P(V_9|H_{5:6})P(V_{10}|H_{5:6}) \\ \times P(V_{11}|H_6)P(H_5)P(H_6) \end{aligned} \quad (19)$$

and the denominator $Denom_3$ is

$$\begin{aligned} Denom_3 = \sum_{allH_{5,m}} \sum_{allH_{6,n}} [P(V_8|H_5)P(V_9|H_{5:6}) \\ \times P(V_{10}|H_{5:6})P(V_{11}|H_6)P(H_5)P(H_6)] . \end{aligned} \quad (20)$$

The summations are over all possible configurations of values on the parent nodes H_5 and H_6. Here $H_{5,m}$ denotes a particular value for state m on node H_5, and $H_{6,n}$ denotes a particular value for state n on node H_6.

The conditional probability table (CPT) and the prior probability table (PPT) for Fig. 9 are specified in Tables 8–10.

4.4 Overall body-pose estimation

The individual Bayesian networks are integrated in a hierarchy to estimate the overall body pose of a person, as shown

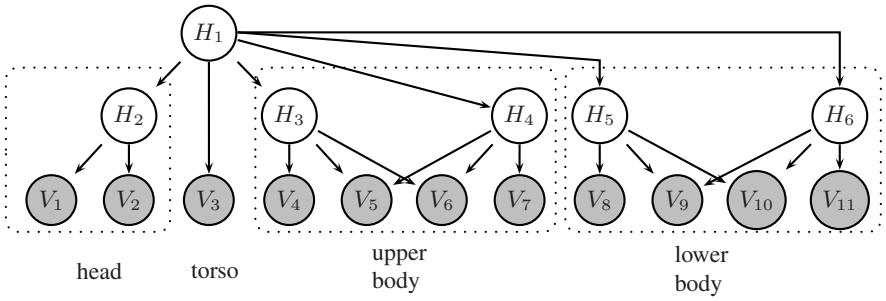


Fig. 11. Hierarchical composition of the Bayesian networks for pose estimation of a person. The BN is composed of 6 hidden nodes $H_{1:6}$ and 11 visible nodes $V_{1:11}$. See legend below for node labels

Table 9. CPT for $P(V_9|H_{5:6})$ and $P(V_{10}|H_{5:6})$

	$P(V_9\|H_{5:6})$		
	A	B	C
AA	0.77	0.11	0.12
AB	0.19	0.41	0.4
AC	0.21	0.4	0.39
BA	0.11	0.3	0.59
BB	0.1	0.5	0.4
BC	0.31	0.29	0.4
CA	0.28	0.31	0.41
CB	0.31	0.32	0.37
CC	0.09	0.31	0.6

	$P(V_{10}\|H_{5:6})$		
	A	B	C
AA	0.42	0.38	0.2
AB	0.23	0.38	0.39
AC	0.31	0.3	0.39
BA	0.29	0.31	0.4
BB	0.3	0.3	0.4
BC	0.32	0.28	0.4
CA	0.31	0.3	0.39
CB	0.3	0.3	0.4
CC	0.28	0.32	0.4

Table 10. PPT for $P(H_5)$ and $P(H_6)$

$P(H_5)$		
A	B	C
0.6	0.25	0.15

$P(H_6)$		
A	B	C
0.59	0.26	0.15

in Fig. 11. The proper hierarchy depends on domain-specific knowledge. It is also possible to learn the structure of the BN given enough training data. However, in the video surveillance domain, there is usually not enough training data available. Therefore, we manually constructed the BN.

The overall BN is specified by the prior probability distribution of H_1 and the conditional probability distributions of the rest of the nodes. The prior probabilities $P(H_2), P(H_3), P(H_4)$, and $P(H_5)$ are replaced by the corresponding conditional probabilities $P(H_i|pa(H_i))$ in the overall BN, where $pa(H_i)$ is the set of parent nodes of node H_i in the directed acyclic graph (DAG). The BN has discrete hidden nodes and discrete observation nodes that are discretized by vector quantization in Sect. 3. The probability distributions are approximated by discrete tabular form and trained by counting the frequency of co-occurrence of states in a node and its parent nodes using training sequences.

We introduce an additional hidden node, H_1, for the torso pose as a root node of the hierarchical BN and an additional observation node, V_3, the median width of the torso in an image. The torso pose is inferred from the poses of the head, arm, and leg, as well as the additional observation of the median width of the torso.

The states of the hidden node H_1 are defined as H_1 = {front view, left view, right view, rear view}.

The joint probability of the BN in Fig. 11 is factored into conditional probabilities and prior probabilities as follows:

$$\begin{aligned} P(V_{1:11}, H_{1:6}) &= \prod_{i=1}^{N} P(X_i|pa(X_i)), \\ &\quad X \in \{H, V\}, N = 6 + 11 \\ &= [\prod_{i=1}^{11} P(V_i|pa(V_i))] \times [\prod_{j=1}^{6} P(H_j|pa(H_j))] \\ &= Factors_V \times Factors_H \end{aligned} \tag{21}$$

where

$$\begin{aligned} Factors_V &= [\prod_{i=1}^{11} P(V_i|pa(V_i))] \\ &= P(V_1|H_2)P(V_2|H_2)P(V_3|H_1)P(V_4|H_3) \\ &\quad \times P(V_5|H_{3:4})P(V_6|H_{3:4})P(V_7|V_4)P(V_8|H_5) \\ &\quad \times P(V_9|H_{5:6})P(V_{10}|H_{5:6})P(V_{11}|H_6) \end{aligned} \tag{22}$$

and

$$\begin{aligned} Factors_H &= [\prod_{j=1}^{6} P(H_j|pa(H_j))] \\ &= P(H_2|H_1)P(H_3|H_1)P(H_4|H_1) \\ &\quad \times P(H_5|H_1)P(H_6|H_1)P(H_1)\,. \end{aligned} \tag{23}$$

Our overall goal is to estimate the *belief* of the states of the hidden nodes $H_{1:6}$ given the evidence $V_{1:11}$:

$$\begin{aligned} P(H_{1:6}|V_{1:11}) &= \frac{P(V_{1:11}, H_{1:6})}{P(V_{1:11})} \\ &= \frac{Numer_4}{Denom_4}\,, \end{aligned} \tag{24}$$

where the denominator $Denom_4$ is the marginalized version of Eq. 21 along $H_{1:6}$:

$$Denom_4 = \sum_{H_1}\sum_{H_2}\sum_{H_3}\sum_{H_4}\sum_{H_5}\sum_{H_6} P(V_{1:11}, H_{1:6})\,. \tag{25}$$

Some of the factors of Eq. 24 were presented in Sects. 4.1–4.3, and the others are specified in Tables 11–14.

Our hierarchical Bayesian network efficiently decomposes the overall task of estimating the beliefs of the overalll body pose into individual beliefs, as described in Sects. 4.1 through 4.4. Depending on the available evidence values of the observation nodes, the actual computation for the belief estimation varies accordingly.

At this stage, we have the pose of a person at a given frame represented in terms of the hidden state indices of $H_{1:6}$. For

Table 11. CPT for $P(H_2|H_1)$ and $P(V_3|H_1)$

$P(H_2\|H_1)$	A	B	C	D
A	0.7	0.16	0.14	0
B	0.13	0.72	0	0.15
C	0.17	0	0.69	0.14
D	0	0.15	0.14	0.71

$P(V_3\|H_1)$	A	B
A	0.1	0.9
B	0.6	0.4
C	0.58	0.42
D	0.11	0.89

Table 12. CPT for $P(H_3|H_1)$ and $P(H_4|H_1)$

$P(H_3\|H_1)$	A	B	C	D
A	0.1	0.2	0.2	0.5
B	0.21	0.31	0.29	0.19
C	0.22	0.28	0.3	0.2
D	0.11	0.2	0.19	0.5

$P(H_4\|H_1)$	A	B	C
A	0.7	0.21	0.09
B	0.5	0.3	0.2
C	0.5	0.28	0.22
D	0.71	0.19	0.1

Table 13. CPT for $P(H_5|H_1)$ and $P(H_6|H_1)$

$P(H_5\|H_1)$	A	B	C
A	0.8	0.09	0.11
B	0.5	0.31	0.19
C	0.52	0.28	0.2
D	0.8	0.08	0.12

$P(H_6\|H_1)$	A	B	C
A	0.78	0.12	0.1
B	0.49	0.31	0.2
C	0.51	0.28	0.21
D	0.81	0.09	0.1

Table 14. Prior probability table for $P(H_1)$

$P(H_1)$			
A	B	C	D
0.25	0.25	0.25	0.25

example, given the evidence of observation e_t^1 of the first (i.e., left) person at frame t, the body pose of the person at frame t in Fig. 1 is represented as the n-tuple ($n = 6$) of the alphabetical indices of the hidden state values ϕ_t^1 as:

$$\begin{aligned} e_t^1 &= [v_{1:11}] \\ &= [DBACDBAACCB]^T , \\ \phi_t^1 &= [H_{1:6}] \\ &= [CCCCAA]^T . \end{aligned} \tag{26}$$

whereas given the evidence of observation e_t^2 of the second (i.e., right) person at frame t, the body pose of the person at frame t is represented as ϕ_t^2:

$$\begin{aligned} e_t^2 &= [V_{1:11}] \\ &= [CBADCBBBCCB]^T , \\ \phi_t^2 &= [H_{1:6}] \\ &= [BBDBAB]^T , \end{aligned} \tag{27}$$

5 Recognition by DBN

We model human interaction as a sequence of state changes that represents the configuration and movement of individual body parts (i.e., torso, arms, and legs) in the spatiotemporal domain. This model requires a representation of 6 interacting processes (3 body parts $\times$ 2 persons) for a two-person interaction. Involving multiple people and multiple body parts makes it difficult to apply a simple HMM model that has a state space composed of a single random variable. Because the overall state space is very large, the joint probability distribution becomes intractable, and a prohibitively large amount of data is required to train the model. It is well known that the exact solution of extensions of the basic HMM to three or more processes is intractable [16].

Our solution to overcoming the exponential growth of the parameter space is to represent human behavior at multiple levels in a hierarchy. We utilize the hierarchy of our BN and abstract the states and the events at multiple levels. Individual body-part motions are analyzed at the low level, and the whole-body motion of a person is processed at the middle level by combining the individual body-part motions. The two-person interaction patterns are recognized at the high level by incorporating the whole-body motion and spatiotemporal constraints about relative positions and causal relations between the two persons.

5.1 Body-part pose evolution

Our dynamic Bayesian network (DBN) is constructed by establishing temporal links between the identical hidden nodes of the BN in Fig. 11 at time t and $t + 1$. This means that we model the hidden state evolutions based on the first-order Markov process. We use a DBN equivalent of HMM for each body part – legs, torso, and arms – as shown in Fig. 12. For simplicity, we assume that the evolution of H_i is independent of the evolution of H_j if $i \neq j$. The difference between a DBN and an HMM is that a DBN represents the hidden state in terms of a set of random variables, ${}^1q_t, ..., {}^Nq_t$, i.e., it uses

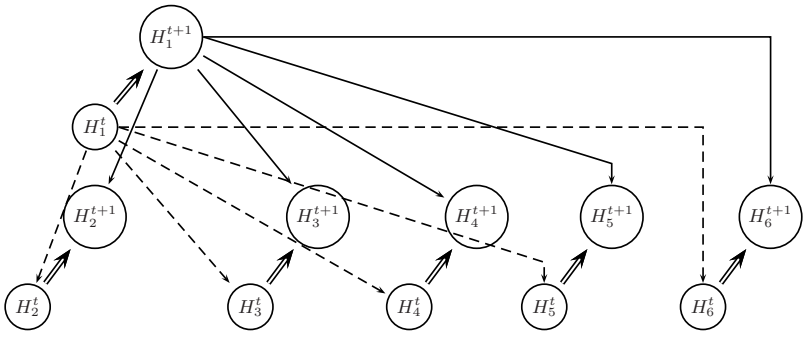


Fig. 12. Hidden nodes' evolution from frame t to frame $t + 1$. The BN structure is specified by *dotted arrows* at frame t and by *solid arrows* at frame $t + 1$. The evolution of each node state is represented by *double solid arrows*

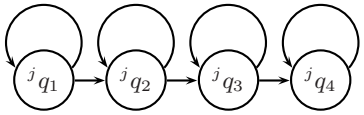


Fig. 13. DBN equivalent of HMM for jth body part unrolled for $T = 4$. *Nodes* represent states, and *arrows* represent allowable transitions, i.e., transitions with nonzero probability. The self-loop on state i means $P(q_t = i|q_{t-1} = i) = A(i, i) > 0$

a distributed representation of state. By contrast, in an HMM, the state space consists of a single random variable q_t. The DBN is a stochastic finite state machine defined by initial state distribution π, observation probability B, and state transition probability A:

$$A = \{a_{ij} \mid a_{ij} = P(q_{t+1} = j \mid q_t = i)\},$$
$$B = \{b_j(k) \mid b_j(k) = P(v_t = k \mid q_t = j)\},$$
$$\pi = \{\pi_i \mid \pi_i = P(q_1 = i)\},$$

where q_t is the state value of the hidden node H at time t and v_t is the observation value of the observation node V at time t.

The observation probability B and the initial state distribution π specify the DBN at a given time t and are already established in our Bayesian network, as described in Sect. 3. Therefore, our DBN only needs the state transition probability A that specifies the state evolution from time t to time $t + 1$. Figure 13 shows a DBN equivalent of HMM unrolled for $T = 4$ time slices, where ${}^jq_t$ are the hidden node states for the jth body part at time t: ${}^1q$ for the leg, ${}^2q$ for the torso, and ${}^3q$ for the arm. We use the Baum-Welch algorithm for training the DBN and the Viterbi algorithm for decoding the network [23].

The DBN hidden states correspond to the progress of the gesture with time.

${}^1Q$ is the set of DBNs for legs: ${}^1Q$ = {"both legs are together on the ground", "both legs are spread on the ground", and "one foot is moving in the air while the other is on the ground".}

Similarly, ${}^2Q$ is the set of DBNs for the torso: ${}^2Q$ = {"stationary", "moving forward", and "moving backward".}

${}^3Q$ is the set of DBNs for arms: ${}^3Q$ = {"both arms stay down", "at least one arm stretches out", and "at least one arm gets withdrawn".}

The assumption of independence between the individual DBNs in our network structure drastically reduces the size of the overall state space and the number of relevant joint probability distributions.

5.2 Whole-body pose evolution

The evolution of a person's body pose is defined by the three results, ${}^1q$, ${}^2q$, and ${}^3q$, of the DBNs in Sect. 5.1 and forms the description of the overall body pose evolution of the person. Examples of the description include {"stand still with arms down", "move forward with arm(s) stretched outward", "move backward with arm(s) raised up", "stand stationary while kicking with leg(s) raised up", etc.}

Instead of using an exponentially increasing number of all possible combinations of the joint states of the DBN, we focus only on a subset of all states to register the semantically meaningful body motions involved in two-person interactions. We observe that a typical two-person interaction usually involves a significant movement of either arm(s) or leg(s), but not both. This means that we assume the joint probability distribution of meaningless combinations of ${}^1q$, ${}^2q$, and ${}^3q$ has a probability virtually equal to 0 in the usual body-part pose combinations.

5.3 Two-person interaction

One of the ultimate goals of recognizing visual events is the automatic generation of user-friendly descriptions of the events. At the high level, we consider the recognition of human behavior from the viewpoint of language understanding in terms of "*subject* + *verb* + (*object*)". The subject corresponds to the person of interest in the image, the verb to the motion of the subject, and the object to the optional target of the motion (i.e., usually the other person's body part). The proximal target of the arm motion (i.e., the torso or head) is regarded as an object term in our semantic verbal description of the interaction. Our body-part-segmentation module provided a set of meaningful body-part labels as a vocabulary for the object term.

Subject = {torso, arm, leg}
Verb = {raise, lower, stretch, withdraw, stay, move forward, move backward}
Object = {head, upper body, hand, lower body}

Our language-based framework is similar to that of [14], but our interest is in the interaction between two autonomous persons.

To recognize an interaction pattern between two persons, we incorporate the relative poses of the two persons. We juxtapose the results of the mid-level description of the individual person along a common time line and add spatial and temporal constraints of a specific interaction type based on domain knowledge about causality in interactions.

6 Relative constraints

The pose estimation of individual body parts is based on the single person's posture in terms of an object-centered coordinate system. However, knowing the body poses is not enough for the recognition of *interaction* between people. Recognizing two-person interactions requires information about the *relative* positions of the two persons' individual body parts. For example, in order to recognize "approaching", we need to know the relative distance and direction of torso movement with respect to that person. This fact leads us to the notion of the dynamics of body movements: the evolution of body-part motion along the sequence under spatiotemporal constraints.

6.1 Spatial constraints

The spatial constraints are defined in terms of the relative position and orientation. For example, "standing hand-in-hand" requires that the torsos of the two persons be side by side and facing in the same direction, while "pointing at the opposite person" requires that the torsos face one another. The "pointing

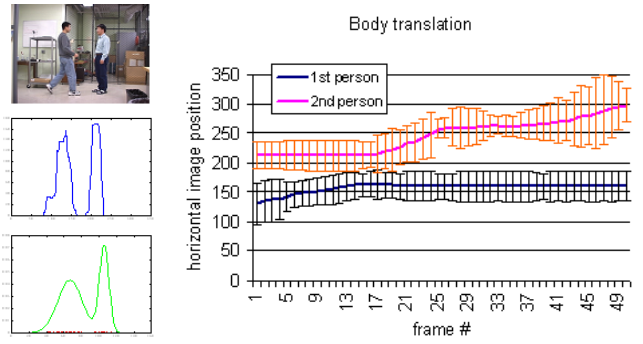


Fig. 14. Gross-level estimation of proximity between two persons using a Gaussian-based tracker [19]. **a** Input image, its 1D projection, and 1D Gaussian approximation from *top* to *bottom*, respectively. **b** Estimated horizontal positions of the torso centers of two persons

at the opposite person" interaction also involves a systematic increase of proximity between a person's moving arm and the torso or head of the opposite person.

We represent the interacting persons' relative position at multiple levels of detail: gross level, intermediate level, and detailed level. The gross-level estimation of proximity between two persons in the "pushing" sequence shown in Fig. 1 are shown in Fig. 14. Each person's foreground silhouette is vertically projected and modeled by a one-dimensional (1D) Gaussian (Fig. 14a). A mixture of the 1D Gaussians is used to model multiple people. Two Gaussians are fitted, each of which models a single person without occlusion. Frame-to-frame update of these Gaussian parameters along the sequence (Fig. 14b) amounts to tracking the whole-body translation of each person in the horizontal image dimension. The plot in Fig. 14b shows the estimated horizontal position of torso center with vertical bars representing uncertainty in terms of standard deviation of the Gaussian tracker. The x-axis represents the frame number and the y-axis the pixel coordinate of the horizontal image dimension. We can see that the left person (i.e., *the first person*) approaches the right person (i.e., *the second person*), who stands still up to frame 17; then the second person moves back while the first person stands still. However, we do not know what kind of event happens between the two persons; does the first person push the second person, or does the first person meet the second person and the second person turn back and depart?

At the intermediate level of representation for relative position, we determine the relative orientations of the torso poses between the two persons. Examples of different relative torso poses are shown in Fig. 15; the top panel shows the camera setup viewing individual persons from distance, and the bottom panel shows the corresponding appearance. An interaction between two persons facing each other (Fig. 15a) may have a very different connotation from a similar interaction in which one faces the other's back (Fig. 15c). The relative orientations may constrain possible interactions; that is, it is unlikely that the two persons in Fig. 15b shake hands with each other. In the example, the left person is in front view (i.e., $azimuth_1 = 0°$), and the right person is in left view (i.e., $azimuth_2 = 90°$). The relative azimuth R_a defines the relative torso poses between the two persons as follows:

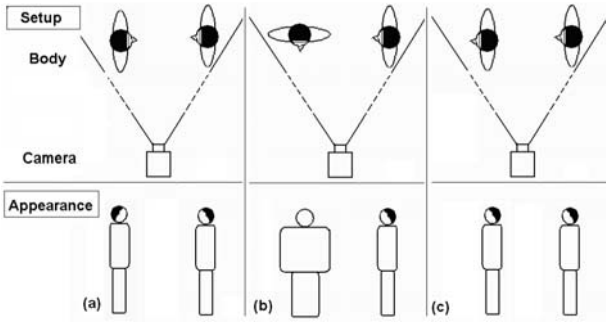


Fig. 15. Diagram for the intermediate-level estimation of relative poses between two persons. **a** Seeing the face of the left person. **b** Seeing the side of the left person. **c** Seeing the back of the left person. *Upper panels* depict geometric setup of camera and persons, and *lower panels* show appearance of the persons

$$R_a = azimuth_2 - azimuth_1 \,. \tag{28}$$

At the detailed level of representation for relative position, we analyze further the relative configuration of individual body parts in order to recognize the specific interaction. An example of detailed information about the relative position of the interacting body parts in Fig. 8 is shown in Fig. 16a. It represents the position of the estimated hand, C, of the left person and the estimated ellipse, B, of the upper body of the right person. The proximity between the estimated arm tip and the estimated upper body is measured. The proximity relations are computed for all combinations of the body parts (i.e., the head, arm, leg, and torso) of the left person with respect to those of the right person, resulting in a 4×4 matrix format. The most significant combination is chosen as the salient proximity in a given frame of a specific interaction type. For example, in the pushing interaction, the proximity relation of the pushing person's arm with respect to the nearest body part of the pushed person is chosen as the salient proximity in the interaction.

6.2 Temporal constraints

The temporal constraints of two-person interactions are defined by two events in terms of causal and coincident relations of the two persons' body-pose changes.

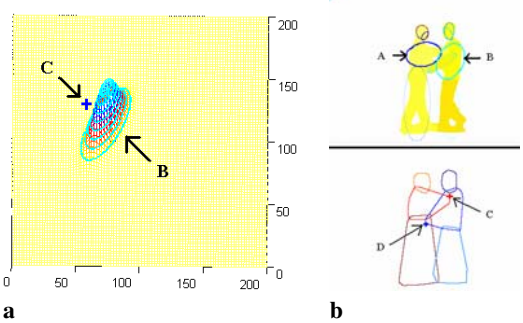


Fig. 16. Detailed-level estimation of proximity between body parts. **a** Relative distance between the left person's hand C and the right person's upper-body ellipse B of Fig. 8. **b** Ellipse-based representation and convex-hull-based representation of the body parts

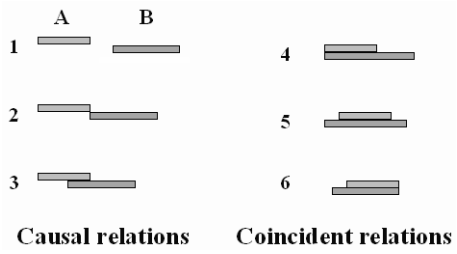


Fig. 17. Temporal constraints between events A and B for **a** causal relations (1: *before*, 2: *meet*, 3: *overlap*) and **b** coincident relations (4: *start*, 5: *during*, 6: *finish*)

Fig. 18. Example frames of the "pointing" sequence. **a** Input image frame. **b** Segmented and tracked body parts. **c** Ellipse parameterization of each body part

We adopt Allen's interval temporal logic [2] to represent the causal and coincident relations in the temporal domain (i.e., *before, meet, overlap, start, during,* and *finish* etc.), as shown in Fig. 17. For example, a pushing interaction involves event A of "a person moving forward with arm(s) stretched outward toward the second person" followed by event B of "move-backward of the second person" as a result of pushing.

We introduce appropriate spatial/temporal constraints for each of the various two-person interaction patterns as domain knowledge. The satisfaction of specific spatial/temporal constraints gates the high-level recognition of the interaction. Therefore, the mid-level and high-level recognitions are characterized by the integration of domain-specific knowledge, whereas the low-level recognition is more closely related to the pure motion of a body part.

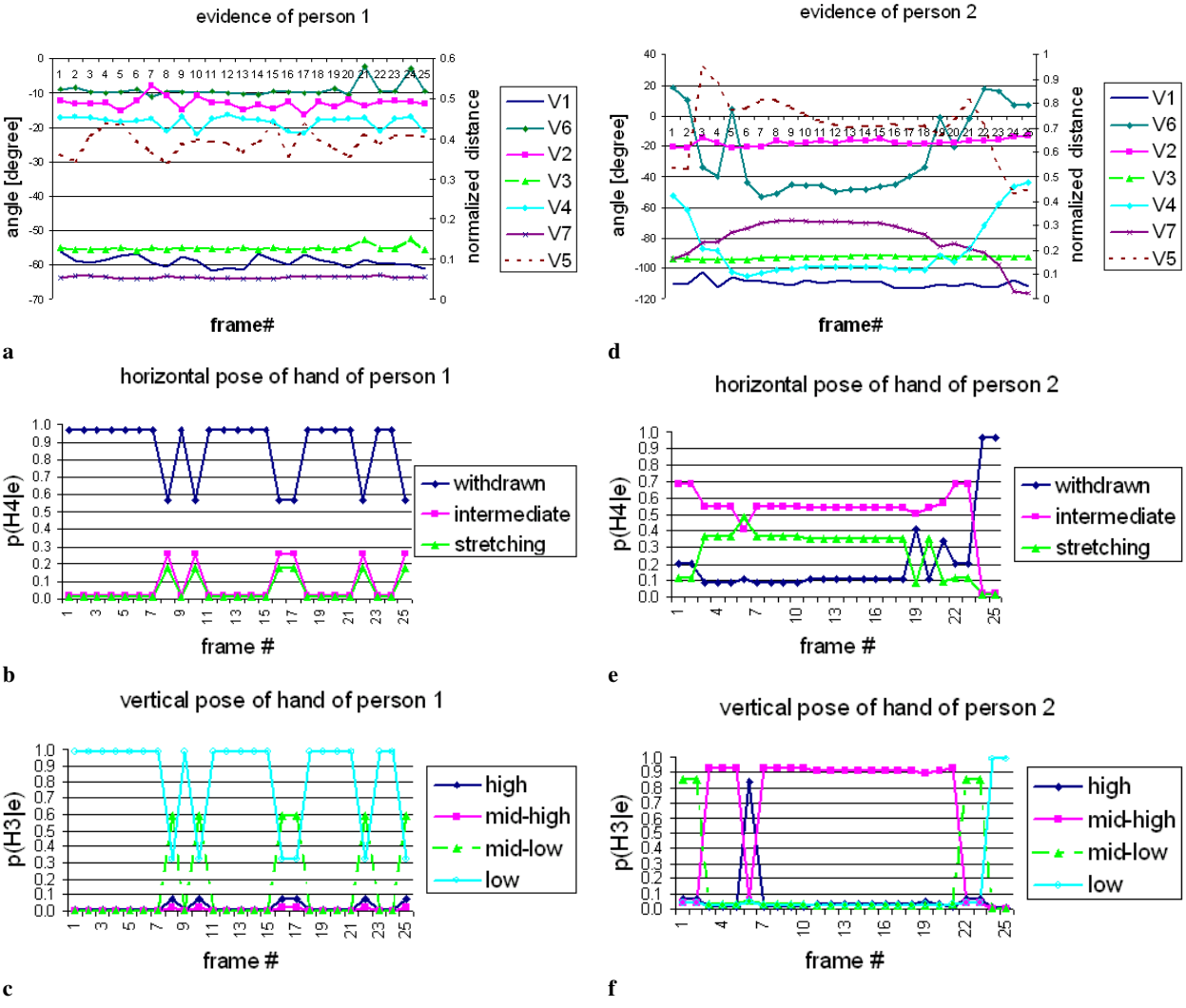


Fig. 19. BN's beliefs about the arm poses of the two persons in the pointing sequence of Fig. 22a. The plots in the *left* and *right columns* show the results for the left and right person in the sequence, respectively. See text for details

7 Experimental results

We have tested our system for various human interaction types including (1) approaching, (2) departing, (3) pointing, (4) standing hand-in-hand, (5) shaking hands, (6) hugging, (7) punching, (8) kicking, and (9) pushing. The images used in this work are 320 $\times$240 pixels in size, obtained at a rate of 15 frames/s. Six pairs of different interacting people with various clothing were used to obtain the total 56 sequences (9 interactions $\times$ 6 pairs of people), in which a total of 285, 293, 232, 372, 268, 281, 220, 230, 264 frames are contained in each of the above interaction types (1)–(9), respectively.

We first evaluated the performance of the BN using the test sequences of the pointing interaction (Fig. 18). Figure 22a shows more frames of the pointing interaction sequence, where the left person stands stationary while the right person raises and lowers his arm to point at the left person. The desired performance of the BN is that stationary posture should be recognized as "stationary" and moving posture should be recognized as "moving". Our BN showed the corresponding results (Fig. 19).

Figure 19 shows the values of the observation nodes V_1–V_7 before quantization and the corresponding beliefs (i.e., the posterior probability) for the two persons in Fig. 22a. In Figs. 19a and d, the angle of vector from the head ellipse center to the face ellipse center (V_1) and the angle of rotation of the torso ellipse (V_6) refer to the left y-axis, while all the other values of the visible nodes are normalized in terms of the height of the person and refer to the right y-axis. The normalization makes the system robust against the height variation of different people. We can incorporate a validation procedure in the Bayesian network in a straightforward manner to check whether an observation is valid in a given situation. For example, if the first person's hand is occluded, its observation is nullified and removed from the evidence.

Figures 19b and c show the BN's belief about each state of the hidden nodes for the horizontal and vertical poses of the salient arm for the left persons in the sequence, given the evidence in Fig. 19a. The BN's belief shows that the horizontal pose of the arm of the left person stays in the *withdrawn* state and the vertical pose of the arm stays in the *low* or the *mid-low* state. These results demonstrate that the belief of the BN for the left person is stable for stationary poses in the sequence.

Corresponding data for the right person are shown in Figs. 19d–f. Figure 19d shows the systematic change of hand position (V_4, V_7) and torso ellipse (V_5, V_6) as the second person points at the first person. Figures 19e and f show the BN's belief about each state of the hidden nodes for the horizontal and vertical poses of the salient arm for the right persons in the sequence, given the evidence in Fig. 19d. The BN's belief shows that the horizontal pose of the arm of the right person changes from *intermediate* $\rightarrow$ *stretching* $\rightarrow$ *intermediate* $\rightarrow$ *withdrawn* states, and that the vertical pose of the arm changes from *mid-low* $\rightarrow$ *mid-high* $\rightarrow$ *high* $\rightarrow$ *mid-high* $\rightarrow$ *mid-low* $\rightarrow$ *low* as he moves his arm. This result corresponds well to the human interpretation of the input sequence. These results demonstrate that the belief of the BN for the right person properly detects the state changes as the person moves his arm for the interaction.

Figure 20 shows other examplar poses of positive interactions: standing hand-in-hand, shaking hands, hugging; and negative interactions: punching, kicking, pushing. The degree of occlusion increases from the *left* to the *right* examples in each *row*

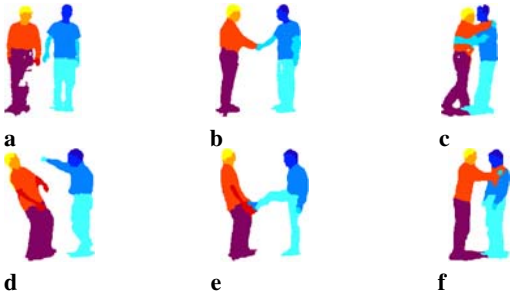


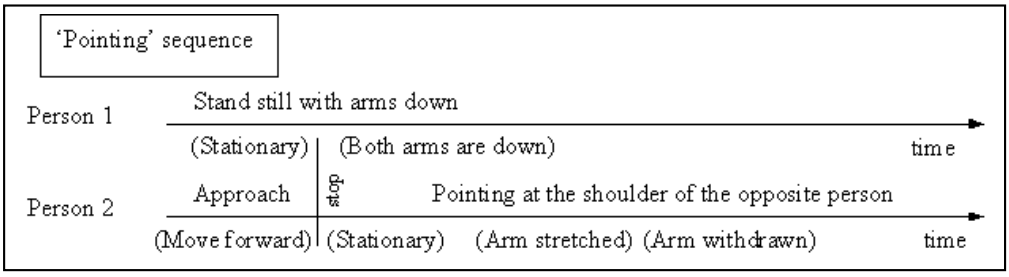
Fig. 20. Example frames of different interactions including positive interactions: **a** standing hand-in-hand, **b** shaking hands, **c** hugging; and negative interactions: **d** punching, **e** kicking, **f** pushing. Degree of occlusion increases from the *left* to the *right* examples in each *row*

The results of the BN of the individual sequences are processed by the dynamic Bayesian network for the sequence interpretations. The dynamic BN generates verbal descriptions of the interactions at the body-part level, the single-person level, and the mutual interaction level. The overall description is plotted over a time line.

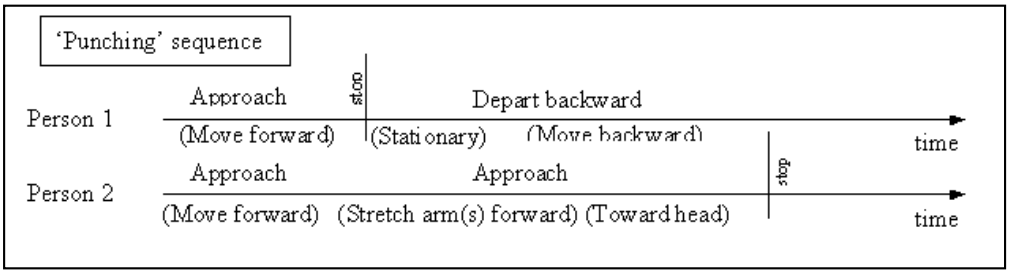
A cross-validation procedure is used to classify the 6 sequences for each interaction type; that is, among the 6 sequences obtained from different pairs of people, 5 sequences were used as training data to test the other one sequence. All 6 sequences were tested in the same way for each of the 9 interaction types. The accuracies of the sequence classification are 100, 100, 67, 83, 100, 50, 67, 83, 50% for the above interaction types (1)-(9), respectively. The overall average accuracy is 78%. The low recognition accuracy of the *hugging* interactions is due to the occlusion involved in the interaction. The low recognition accuracies of the *pointing*, *punching*, and *pushing* interactions are due to the similarity of those interactions. For example, we observed that none of the subjects in the *punching* interaction actually hit the opponent person. The opponent person escapes *before* the actual hitting occurs. It makes the *punching* interaction appear more similar to the *pointing* or *pushing* interactions. The subjects in the experiments were aware of the interactions they were supposed to perform, which makes the escape behavior possible. It is expected that in real situations people would behave in the manner that conforms to the definition of the interactions. Figure 21 shows examples of the time line of the interaction behaviors pointing, punching, standing hand-in-hand, pushing, and hugging, represented in terms of events and simple behaviors. Figure 22 shows subsampled frames of various interaction sequences. Note that the degree of occlusion increases from top to bottom rows.

8 Conclusion

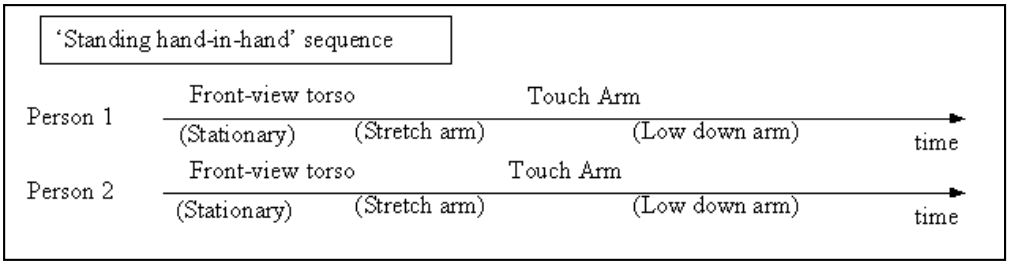
This paper has presented a method for the recognition of two-person interactions using a hierarchical Bayesian network (BN). The poses of simultaneously tracked body parts are estimated at the low level, and the overall body pose is estimated at the high level. The evolution of the poses of the multiple



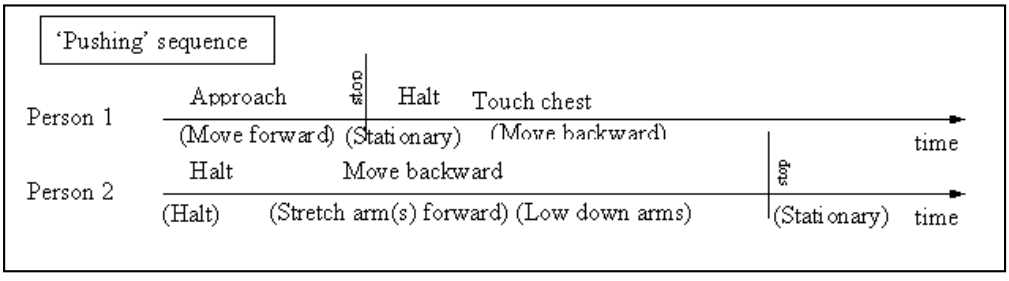
a



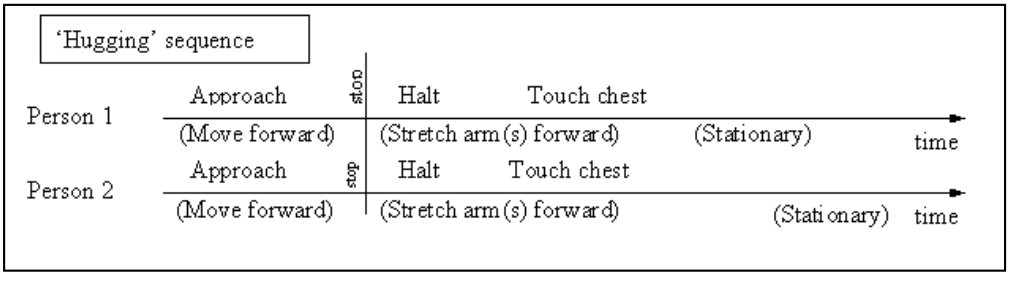
b



c



d



e

Fig. 21. Semantic interpretation of two-person interactions: **a** pointing, **b** punching, **c** standing hand-in-hand, **d** pushing, and **e** hugging

Fig. 22. Subsampled frames of various interaction sequences: pointing, punching, standing hand-in-hand, pushing, and hugging interactions from *top* to *bottom rows*

body parts during the interaction is analyzed by a dynamic Bayesian network. The recognition of two-person interactions is achieved by incorporating domain knowledge about relative poses and event causality.

The key contributions of this paper are as follows. (1) A new hierarchical framework is proposed for the recognition of two-person interactions at a detailed level using a hierarchical Bayesian network. (2) Ambiguity in human interaction due to occlusion is handled by inference with the Bayesian network. (3) A human-friendly vocabulary is generated for high-level event description. (4) A stochastic graphical model is proposed for the recognition of two-person interactions in terms of "subject + verb + object" semantics.

Our experiments show that the proposed method efficiently recognizes detailed interactions that involve the motions of the multiple body parts. Our plans for future research include the exploration of the following aspects: (1) extending the method to crowd behavior recognition, (2) incorporating various camera-view points, and (3) recognizing more diverse interaction patterns.

References

1. Aggarwal JK, Cai Q (1999) Human motion analysis: a review. Comput Vis Image Understand 73(3):295–304
2. Allen JF, Ferguson G (1994) Actions and events in interval temporal logic. J Logic Comput 4(5):531–579
3. Bakowski A, Jones G (1999) Video surveillance tracking using color region adjacency graphs. In: 7th international conference on image processing and its applications, 13–15 July 1999, University of Manchester, UK, pp 794–798
4. Barron C, Kakadiaris I (2003) A convex penalty method for optical human motion tracking. In: ACM international workshop on video surveillance (IWVS), Berkeley, CA, November 2003, pp 1–10
5. Cowell RG, Dawid AP, Lauritzen SL, Spiegelhalter DJ (1999) Probabilistic networks and expert systems. Springer, Berlin Heidelberg New York
6. Data A, Shah M, Lobo N (2002) Person-on-person violence detection in video data. In: Proceedings of the international conference on pattern recognition, Quebec City, Canada, 1:433–438
7. Elgammal AM, Davis L (2001) Probabilistic framework for segmenting people under occlusion. In: International conference on computer vision, Vancouver, Canada, 2:145–152
8. Gavrila D (1999) The visual analysis of human movement: a survey. Comput Vis Image Understand 73(1):82–98
9. Graham RL (1972) An efficient algorithm for determining the convex hull of a finite planar set. Inf Process Lett 1:132–133
10. Haritaoglu I, Harwood D, Davis LS (2000) W4: Real-time surveillance of people and their activities. IEEE Trans Pattern Anal Mach Intell 22(8):797–808
11. Hongeng S, Bremond F, Nevatia R (2000) Representation and optimal recognition of human activities. In: IEEE conference on computer vision and pattern recognition, 1:818–825
12. Huang C, Darwiche A (1996) Inference in belief networks: a procedural guide. Int J Approx Reason 15(3):225–263
13. Jensen FV, Jensen F (1994) Optimal junction trees. In: Conference on uncertainty in artificial intelligence, Seattle, July 1994
14. Kojima A, Tamura T, Fukunaga K (2002) Natural language description of human activities from video images based on concept hierarchy of actions. Int J Comput Vis 50(2):171–184
15. Moeslund T, Granum E (2001) A survey of computer vision-based human motion capture. Comput Vis Image Understand 81(3):231–268
16. Oliver NM, Rosario B, Pentland AP (2000) A Bayesian computer vision system for modeling human interactions. IEEE Trans Pattern Anal Mach Intell 22(8):831–843
17. O'Rourke J (1994) Computational geometry in C. Cambridge University Press, Cambridge, UK, pp 70–112
18. Park S, Aggarwal JK (2000) Recognition of human interaction using multiple features in grayscale images. In: Proceedings of the internaitonal conference on pattern recognition, Barcelona, Spain, September 2000, 1:51–54
19. Park S, Aggarwal JK (2002) Segmentation and tracking of interacting human body parts under occlusion and shadowing. In: IEEE workshop on motion and video computing, Orlando, FL, pp 105–111
20. Park S, Aggarwal JK (2003) Recognition of two-person interactions using a hierarchical Bayesian network. In: ACM international workshop on video surveillance, Berkeley, CA, pp 65–76
21. Park S, Park J, Aggarwal JK (2003) Video retrieval of human interactions using model-based motion tracking and multi-layer finite state automata. In: Lecture notes in computer science, vol 2728. Springer, Berlin Heidelberg New York, pp 394–403
22. Pearl J (1988) Probabilistic reasoning in intelligent systems: networks of plausible inference. Morgan Kaufmann, San Mateo, CA, pp 337–340
23. Rabiner L (1989) A tutorial on hidden markov models and selected applications in speech recognition. Proc IEEE 77(2):257–286
24. Rosales R, Sclaroff S (2000) Inferring body pose without tracking body parts. In: Computer vision and pattern recognition, Hilton Head Island, SC, pp 721–727
25. Sato K, Aggarwal JK (2001) Recognizing two-person interactions in outdoor image sequences. In: IEEE workshop on multi-object tracking, Vancouver, CA
26. Sherrah J, Gong S (2000) Resolving visual uncertainty and occlusion through probabilistic reasoning. In: British machine vision conference, Bristol, UK, pp 252–261
27. Sherrah J, Gong S (2000) Tracking discontinuous motion using bayesian inference. In: 6th European conference on computer vision, pp 150–166
28. Siebel N, Maybank S (2001) Real-time tracking of pedestrians and vehicles. In: IEEE workshop on PETS, Kauai, HI
29. Wada T, Matsuyama T (2000) Multiobject behavior recognition by event driven selective attention method. IEEE Trans Pattern Anal Mach Intell 22(8):873–887