Semantic-level Understanding of Human Actions and Interactions using Event Hierarchy

Sangho Park
Electrical and Computer Engineering
The University of Texas at Austin
Austin, TX 78712, USA
sangho@ece.utexas.edu

J.K. Aggarwal
Electrical and Computer Engineering
The University of Texas at Austin
Austin, TX 78712, USA
aggarwaljk@mail.utexas.edu

Abstract

Understanding human behavior in video data is essential in numerous applications including surveillance, video annotation/retrieval, and human-computer interfaces. This paper describes a framework for recognizing human actions and interactions in video by using three levels of abstraction. At low level, the poses of individual body parts including head, torso, arms and legs are recognized using individual Bayesian networks (BNs), which are then integrated to obtain an overall body pose. At mid level, the actions of a single person are modeled using a dynamic Bayesian network (DBN) with temporal links between identical states of the Bayesian network at time t and $t+1$. At high level, the results of mid-level descriptions for each person are juxtaposed along a common time line to identify an interaction between two persons. The linguistic 'verb argument structure' is used to represent human action in terms of <agent-motion-target> triplets. Spatial and temporal constraints are used for a decision tree to recognize specific interactions. A meaningful semantic description in terms of subject-verb-object is obtained. Our method provides a user-friendly natural-language description of several human interactions, and correctly describes positive, neutral, and negative interactions occurring between two persons. Example sequences of real persons are presented to illustrate the paradigm.

1. Introduction

Understanding human behavior is essential in applications including automated surveillance, video archival/retrieval, medical diagnosis, and human-computer interaction. One of the goals in video understanding is to describe the actions and interactions between persons at an event level. At the event level, the focus of computing is to achieve semantic understanding of video imagery. The semantic-level understanding involves semantic description of video events; we call it *event semantics*. A description gap exists between geometric information obtained from images and semantic information contained in natural language [3]. It is necessary to associate visual features with natural language verbs and symbols to build the event semantics of two-person interactions. 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). Designing a natural language-based human-computer interface is highly desirable because of the rich structure of syntax and semantics representing domain-specific rules and contexts.

The notion of describing interaction between human and inanimate objects has been pursued by several researchers. Early study of conceptual dependency in understanding human behavior was proposed in [11] from the perspective of artificial intelligence. A single person's hand manipulation was interpreted in terms of the laws of physics in [4] and in terms of sign gesture in [12]. Single-person activities in an office environment were described in [3, 2]. Event description of remote scenes in outdoor surveillance was presented in [5, 6]. Recognition of interactions between two pedestrians at blob level was presented in [10]. Most of the research has been aimed either at understanding single person actions with inanimate objects such as office gadgets or at understanding multiple-person interactions in remote scenes at a coarse level, with each person represented as a simple moving box. Description and understanding of person-to-person interactions at a detailed level with the information about individual body parts has not been addressed.

Recognizing human interactions is a challenging task due to the ambiguity caused by nonrigid body articulation,
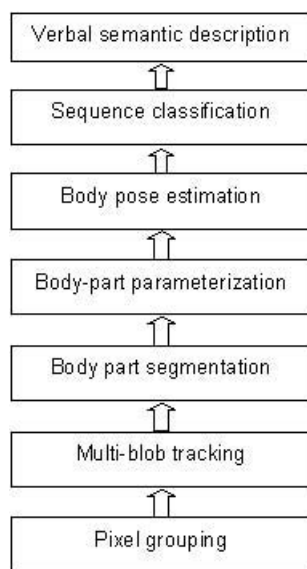
Figure 1. System diagram.

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 this paper, we present a methodology that estimates body-part pose and recognizes different two-person interactions. The following nine interaction types are considered in this paper: the *neutral* interactions include (1) *approaching each other*, (2) *departing each other*, and (3) *pointing*, and the *positive* interactions include (4) *shaking hands*, (5) *hugging*, and (6) *standing hand-in-hand*, and the *negative* interactions include (7) *punching*, (8) *pushing*, and (9) *kicking*.

Detailed recognition of human interactions requires that body-part information be available or able to be inferred from image data. To meet this requirement, we assume, for simplicity, the use of a single fixed camera with a viewing axis parallel to the ground (i.e., horizon), and stable ambient illumination in an indoor environment. The overall system diagram is shown in Fig. 1.   The recognition algorithm is preceded by a feature extraction algorithm that extracts body-pose features from the segmented and tracked body-part regions. At low levels, individual pixels are grouped into blobs according to the pixel color and position, and multiple blobs are tracked along the sequence. A simple human body model is incorporated to segment body parts, and the poses of individual body parts including head, torso,

arms and legs are recognized using individual Bayesian networks, which are then integrated to obtain an overall body pose. At mid level, the actions of a single person are modeled. For this, a concept of dynamic Bayesian network is introduced, which uses temporal links between identical states of the original Bayesian network at time t and $t + 1$. For high level interactions between two persons, the results of mid-level descriptions for each person are juxtaposed along a common time line. At this level, 'verb argument structure' in linguistics is used to represent human action in terms of <agent-motion-target> triplets. Spatial and temporal constraints are used for a decision tree to recognize the specific interactions. In this framework, human action is automatically represented in terms of verbal description according to *subject + verb + object* syntax, and human interaction is represented in terms of *cause + effect* semantics between the human actions. Our framework can recognize person-to-person interactions of various types (i.e., positive, neutral, and negative types) at a detailed semantic level in which multiple body-part motions are involved.

The rest of the paper is organized as follows: Section 2 summarizes our previous work related to the stages of feature extraction, pose estimation, and sequence classification in our system in Fig. 1. Section 3 presents the formulations of interaction hierarchy, verb argument structure, and vocabulary lists. Sections 4 and 5 describe the interrelations of multiple action events involved in two-person interactions. Section 6 shows the definition of interaction types and Section 7 presents the rule-based classification procedure. Experimental results and conclusions follow in Section 8.

2. From image sequence to verb phrase

In our previous work, we presented a method to segment and track multiple body parts in two-person interactions [7], and a method to estimate body poses and gestures using the ellipses and convex hulls of individual body parts [8] as shown in Fig. 2.

Our method is based on multi-level processing at pixel-level, blob-level, and object-level. At pixel level, individual pixels of the foreground image are classified into homogeneous blobs according to color (Fig. 2(a).) At blob level, adjacent blobs are merged to form large blobs according to a blob similarity metric. At object level, sets of multiple blobs are labeled as human body-part regions according to domain knowledge. The segmented body parts include head, upper body, arm(s), lower body, and leg(s) (Fig. 2(b).) The multiple body part regions are tracked along the image sequence. In [8] we presented a Bayesian network to estimate body poses and gestures by parameterizing the body parts in terms of ellipses and convex hulls (Fig. 2(c) and (d).) The Bayesian network estimates instantaneous body

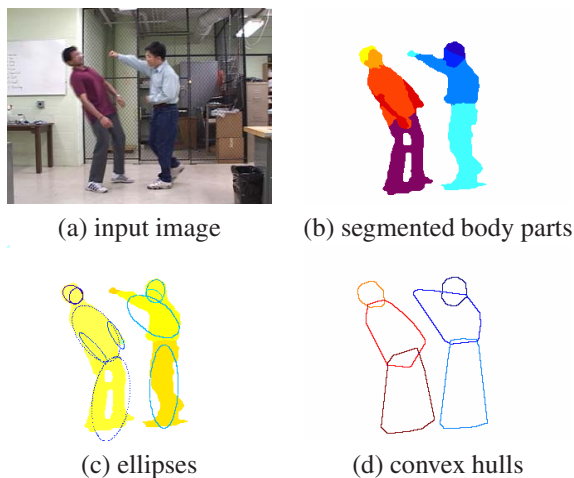|	
(a) input image	(b) segmented body parts
(c) ellipses	(d) convex hulls

Figure 2. An example of estimating pose and gesture from the 'punching' sequence. Ellipses and convex hulls are used for pose estimation by the Bayesian network.

poses, such as the orientation of the head, the hand position, the foot position, etc. Then, a dynamic Bayesian network is constructed to estimate the temporal evolution of the poses along the sequence to recognize the dynamic gestures of the body parts.

The Bayesian network converts the quantitative image features to qualitative information about poses and gestures. For high-level understanding of human actions and interactions, it is desired to convert the qualitative aspects of poses and gestures to the highly organized structure of semantics. We present a new framework for building event hierarchy in the next section.

3. Event hierarchy

Our representation of human interaction is based on the notion of hierarchy; two-person *interaction* is a combination of single-person actions, and the single-person *action* made up of multiple body-part *gestures* such as torso motion and arm/leg motion. Each body-part gesture is an *elementary event* of motion and is composed of a sequence of instantaneous *poses* at each frame. (See Fig. 3.) The human body is represented as both the autonomous *subject* and *object* in the two-person interaction. Each person is considered to be an autonomous subject, with the interacting person regarded as the object of that subject. Thus each person in a two-person interaction is both subject and object.

We conceptualize human actions in terms of an *operation triplet* defined as <agent−motion−target> according

Interaction hierarchy: interaction − action − gesture − pose
Human *interaction*   = combination of two single-person actions
Single-person *action* = torso gesture + arm/leg gesture
Torso-*gesture* :
      constrains possible configuration of body-part *poses*
      associated with specific interactions
Arm-/leg-*gesture* :
      constitutes action-units characterized by trajectory.
Instantaneous *pose* :
      basic building block of human interaction hierarchy

Figure 3. Human interaction hierarchy

to the linguistic theory of 'verb argument structure' [9]. The argument structure of a verb allows us to predict the relationship between the syntactic arguments of a verb and their role in the underlying lexical semantics of the verb. (See Fig. 4.)

Set notation for human action:
The universe set of human action: U
    U = { action | action = <agent−motion−target> }
agent set: S
    S = { s_i | s_i = various body parts as agent term}
      = {*head, torso, arm, leg*}
motion set: V
    V = { v_j | v_j = movement of the body part}
      = {*stay, move left, move right, raise, lower, stretch, withdraw*}
target set: O
    O = { o_k | o_k = the other person's body parts}
      = {*head, torso, chest, abdomen, arm, leg, null*}

Figure 4. Human action as 'operation triplet' and corresponding vocabulary sets.

The *operation triplet* represents the goal-oriented motion of an agent (i.e., a body part) directed toward an optional target. The *agent* set contains 'head', 'torso', 'arm' and 'leg' as vocabulary for possible body parts. The *motion* set contains basic 'action-atoms' such as 'stay', 'move left', 'move right', 'raise', 'lower', 'stretch' and 'withdraw' as vocabulary for possible motion of the body parts. The *target* set contains 'head', 'torso', 'chest', 'abdomen', 'arm', 'leg' and 'null' as vocabulary for possible target of the motion, where 'null' indicates no target is involved.

Event understanding is achieved by transforming a video sequence to a verbal description using various *operation triplets* filled with the appropriate vocabulary terms in

Fig. 4. The transformation rules are determined by domain-specific knowledge about human interactions and human-body kinematics. The universe of the total operational triplets are summarized in Fig. 5.

α_1: [Torso$_i$ Move-Forward Torso$_j$]
α_2: [Torso$_i$ Move-Forward Null$_j$]
α_3: [Torso$_i$ Move-Backward Torso$_j$]
α_4: [Torso$_i$ Move-Backward Null$_j$]
α_5: [Torso$_i$ Stay-Stationary Null$_j$]
α_6: [Arm$_i$ Stay-Stationary Head$_j$]
α_7: [Arm$_i$ Stay-Stationary Torso$_j$]
α_8: [Arm$_i$ Stay-Stationary Chest$_j$]
α_9: [Arm$_i$ Stay-Stationary Abdomen$_j$]
α_{10}: [Arm$_i$ Stay-Stationary Arm$_j$]
α_{11}: [Arm$_i$ Stay-Stationary Leg$_j$]
α_{12}: [Arm$_i$ Stay-Stationary Null$_j$]
α_{13}: [Arm$_i$ Raise Head$_j$]
α_{14}: [Arm$_i$ Raise Torso$_j$]
α_{15}: [Arm$_i$ Raise Chest$_j$]
α_{16}: [Arm$_i$ Raise Abdomen$_j$]
α_{17}: [Arm$_i$ Raise Arm$_j$]
α_{18}: [Arm$_i$ Raise Leg$_j$]
α_{19}: [Arm$_i$ Raise Null$_j$]
α_{20}: [Arm$_i$ Lower Head$_j$]
α_{21}: [Arm$_i$ Lower Torso$_j$]
α_{22}: [Arm$_i$ Lower Chest$_j$]
α_{23}: [Arm$_i$ Lower Abdomen$_j$]
α_{24}: [Arm$_i$ Lower Arm$_j$]
α_{25}: [Arm$_i$ Lower Leg$_j$]
α_{26}: [Arm$_i$ Lower Null$_j$]
α_{27}: [Arm$_i$ Stretch Head$_j$]
α_{28}: [Arm$_i$ Stretch Torso$_j$]
α_{29}: [Arm$_i$ Stretch Chest$_j$]
α_{30}: [Arm$_i$ Stretch Abdomen$_j$]
α_{31}: [Arm$_i$ Stretch Arm$_j$]
α_{32}: [Arm$_i$ Stretch Leg$_j$]
α_{33}: [Arm$_i$ Stretch Null$_j$]
α_{34}: [Arm$_i$ Withdraw Null$_j$]
α_{35}: [Leg$_i$ Stay-Stationary Head$_j$]
α_{36}: [Leg$_i$ Stay-Stationary Torso$_j$]
. . .
α_{62}: [Leg$_i$ Stretch Null$_j$]
α_{63}: [Leg$_i$ Withdraw Null$_j$]
α_0: [Null$_i$ Null Null$_j$]

Figure 5. The universe of the operational triplets. $\alpha_{37} - \alpha_{61}$ **are similar to** $\alpha_8 - \alpha_{32}$ **with exception that agent terms are Leg instead of Arm.**

4. Learning the relations of multiple action events

Depending on the complexity of an action, multiple triplets may be involved in a specific action. We choose only the triplets that are salient to the specific action. For example, the 'pointing' action involves the 'raising' motion of an arm followed by the 'lowering' motion of the arm regardless of any motion of the legs. In this case, the arm is regarded as the salient agent for the 'pointing' action. Two-person interaction is represented by a pair of single-person actions juxtaposed along a common time line.

The linkage between the agent and the target requires information about the *relative* positions of the two persons' individual body parts. Information about the relative positions of the body parts leads to composite verb concepts such as 'approach', for example. In order to recognize 'approaching', we need to know the relative distance and direction of the torso 'motion' with respect to the other person. This fact leads us to the notion of spatial/temporal constraints between the body parts' states regarding their static poses and dynamic gestures (Fig. 4).

We have generated discrete indices for the poses of individual body parts using our Bayesian network in [8]. The alphabetical indices A,B,C, etc. in Fig. 6 correspond to the individual body part poses.

Torso = {A:'front-', B:'left-', C:'right-', D:'rear-view'}
Head = {A:'front-', B:'left-', C:'right-', D:'rear-view'}
ArmV = {A:'high', B:'mid-high',C:'mid-low', D:'low'}
ArmH = {A:'withdrawn', B:'intermediate', C:'stretching'}
LegV = {A:'high', B:'middle', C:'low'}
LegH = {A:'withdrawn', B:'intermediate', C:'out-reached'}

Figure 6. Static poses of body parts.

The sequences of the discrete indices for the body-part poses are concatenated to form body-part *gestures* as shown in Fig. 7 for 'pointing' interaction.

ϕ_t^j represents the vector of the torso pose, head pose, vertical arm pose, horizontal arm pose, vertical leg pose, and horizontal leg pose for the j-th person at frame t. $\phi_{1:m}^j$ is the concatenation of the instantaneous poses along the frames 1 through m. The bounding boxes in the example sequence show that the arm gesture, *stretch*, is detected for the second person at the initial stage of the input stream, and that the arm gestures, *lower* and *withdraw*, are detected at the final stage of the input stream. The rest of the input streams corresponds to the default gesture, *stay-stationary*.

The temporal constraints in two-person interactions are defined by causal and coincident relations represented in
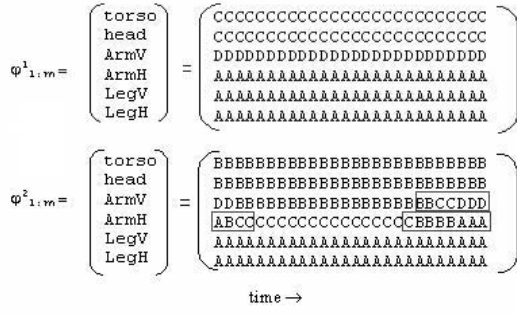
IEEE
COMPUTER
SOCIETY

$$\varphi^1_{1:m} = \begin{pmatrix} torso \\ head \\ ArmV \\ ArmH \\ LegV \\ LegH \end{pmatrix} = \begin{pmatrix} \texttt{CCCCCCCCCCCCCCCCCCCCCCCCCC} \\ \texttt{CCCCCCCCCCCCCCCCCCCCCCCCCC} \\ \texttt{DDDDDDDDDDDDDDDDDDDDDDDDDD} \\ \texttt{AAAAAAAAAAAAAAAAAAAAAAAAAA} \\ \texttt{AAAAAAAAAAAAAAAAAAAAAAAAAA} \\ \texttt{AAAAAAAAAAAAAAAAAAAAAAAAAA} \end{pmatrix}$$

$$\varphi^2_{1:m} = \begin{pmatrix} torso \\ head \\ ArmV \\ ArmH \\ LegV \\ LegH \end{pmatrix} = \begin{pmatrix} \texttt{BBBBBBBBBBBBBBBBBBBBBBBBBB} \\ \texttt{BBBBBBBBBBBBBBBBBBBBBBBBBB} \\ \texttt{DDBBBBBBBBBBBBBBBBBBBCCDDD} \\ \texttt{ABCCCCCCCCCCCCCCCCBBBBAAA} \\ \texttt{AAAAAAAAAAAAAAAAAAAAAAAAAA} \\ \texttt{AAAAAAAAAAAAAAAAAAAAAAAAAA} \end{pmatrix}$$

$time \rightarrow$

Figure 7. An example sequence of discrete states of body poses for 'pointing' interaction. Bounding boxes show salient gestures of body parts.

Figure 8. Examples of different torso poses $[\Theta^1_{TO}, \Theta^2_{TO}]$ **between two persons: (a)** $[\Theta^1_{TO}, \Theta^2_{TO}] = [D, B]$**, (b)** $[\Theta^1_{TO}, \Theta^2_{TO}] = [A, B]$**, (c)** $[\Theta^1_{TO}, \Theta^2_{TO}] = [B, B]$**, (d)** $[\Theta^1_{TO}, \Theta^2_{TO}] = [B, D]$**.**

terms of various operational triplets in Fig. 5. We adopt Allen's interval temporal logic [1] to represent the causal and coincident relations of two action events in the temporal domain (i.e., *before, meet, overlap, start, during,* and *finish* etc.).

The cardinality of possible triplet combinations for representing a single person's composite action involving torso, arm, and leg motions may be computed in a brute-force manner as follows; if we consider a person's torso, arm, and leg as agents moving simultaneously, then the cardinality of a single person action is the multiplication of the triplet cardinalities for the torso, arm and leg.

$$\begin{aligned} |action| &= | < \textbf{torso} - motion - target > | \times \\ &\quad | < \textbf{arm} - motion - target > | \times \\ &\quad | < \textbf{leg} - motion - target > | \\ &\quad |\{\alpha_{1-5}\} \times |\{\alpha_{6-34}\}| \times |\{\alpha_{35-63}\} \\ &= (5) \times (29) \times (29) \\ &= 4205 \end{aligned} \quad (1)$$

The brute-force cardinality in equation (1) shows the number of possible variations of a single person's whole body action, which is a big space! For a two-person interaction, the whole configuration space would be $4205^2 = 17,682,025$. Obviously, it is impractical to process the whole configuration space to recognize human interaction. Fortunately, many of the possible combinations are not likely to occur in a normal human body given its usual kinematic constraints. In our study, we focus on realistic meaningful combinations of torso, arm, and leg actions involved in usual human actions and interactions.

We learn the appropriate spatial/temporal constraints for individual interaction patterns as domain knowledge, and organize the domain knowledge into the rules in a decision-tree structure for classifying the two-person interactions.
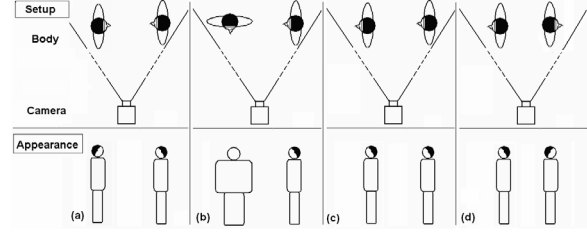
5. Interaction Phrase

The action of a single person may involve the event of the motion of a single body part or the event of the simultaneous motions of multiple body parts (i.e., torso, arm, and/or leg) in a given torso-pose configuration Θ^j_{TO}. Θ^j_{TO} represents the j-th person's static pose of the torso defined in Fig. 6.

Depending on the torso-pose configurations between the two persons, an interaction between the two persons facing each other may have a very different connotation than a similar interaction in which one faces the other's back. Figure 8 shows the environment setup for the image capture and the examples of relative torso poses; the upper panel depicts the geometric camera setup in which individual persons are viewed from distance, and the lower panel shows the corresponding appearances of the persons. The total 16 combinations of the two torso poses are categorized as:

- $S^1 = \{[A, B], [C, B], [D, A], [D, C]\}$

- $S^2 = \{[A, D], [B, A], [B, C], [B, D], [C, D]\}$

- $S^3 = \{[A, A], [C, C]\}$

- $S^4 = \{[B, B], [D, D]\}$

- $S^5 = \{[A, C], [C, A]\}$

- $S^6 = \{[D, B]\}$

We represent the action of the j-th person as Act^j

$$Act^j = \begin{pmatrix} torso\ orientation \\ torso\ triplet \\ arm\ triplet \\ leg\ triplet \end{pmatrix} = \begin{pmatrix} \Theta^j_{TO} \\ \alpha^j_{1:5} \\ \alpha^j_{6:34} \\ \alpha^j_{35:63} \end{pmatrix} \quad (2)$$

where if a single body part, e.g., an arm, is involved in a motion, then the triplets of the other body parts (i.e., torso and leg) in Act^j are assigned with $null$.

The representation of a two-person interaction requires the representation of two Acts at a given time period Δ_{tk}. We represent the interaction of the two persons as $Interact^{ij}_{\Delta_{tk}}$

$$Interact^{ij}_{\Delta_{tk}} = \left(\begin{array}{c} Act^i \\ Act^j \end{array} \right)$$

If the interaction is composite in causal relation, then we need to represent the interaction in terms of the **CAUSE** and the **EFFECT** juxtaposed along a timeline.

$$Interact^{ij}_{\Delta_t} = [\mathbf{CAUSE} \quad , \quad \mathbf{EFFECT}]$$

$$= [Interact^{ij}_{\Delta_{t1}}, \ Interact^{ij}_{\Delta_{t2}}]$$

where the total duration Δ_t spans Δ_{t1} and Δ_{t2}.

We observe that some components of the operational triplets are *essential* for a given interaction according to the constraints of the interpersonal configuration involved in the interaction, while other components of the operational triplets are *incidental*. For example, in the *pushing* interaction, the agent person's (say, the left person's) arm motion α^1_{27} is essential because it constitutes the *pushing* gesture per se. The target person's (say, the right person's) torso motion α^2_4 is also essential because it constitutes the effect of the interaction. In contrast, the target person's arm motion α^2_{19} may not occur if the pushing action is not strong. The target person's arm motion is incidental in constituting the *pushing* interaction. The *interaction* classes determine which components are essential and which are incidental, and the decision requires either user discretion or training from data. We manually construct the classification rules for the human interactions using the *essential* operational triplets, and convert the rules to a decision-tree structure.

6. Operational Definition of Human Interactions

We focus on three categories of human interactions: neutral, positive, and negative. Three interaction types are selected as examples of each of the three categories, as follows: *approaching*, *departing*, and *pointing* as neutral interactions, *hand-shaking*, *standing hand-in-hand*, and *hugging* as positive interactions, and *punching*, *pushing*, and *kicking* as negative interactions. Note that each interaction is mainly referred to in terms of a verb.

Webster's Dictionary defines each of the interaction verbs as shown in figure 9.

Our operational definitions of the human interactions are constructed as follows. We transform the natural language-based, dictionary definitions of human interactions to the operational triplet-based representations of human actions

NEUTRAL INTERACTIONS:
approach: to draw closer to
depart: to go away from
point: to indicate the position or direction of especially by extending a finger

POSITIVE INTERACTIONS:
shake hands: to clasp usually right hands by two people (as in greeting or farewell)
standing hand-in-hand: stand with hands clasped (as in intimacy or affection)
hug: to press tightly especially in the arms

NEGATIVE INTERACTIONS:
punch: to strike with a forward thrust especially of the fist
push: to press against with force in order to drive or impel
kick: to strike out with the foot or feet

Figure 9. Dictionary definitions of human interaction types

in video. We include additional constraints to our operational definitions in order to enhance discrimination between the interactions and to reduce ambiguity.

The operational definitions of the human interactions are shown in figure 10. The operational definitions of the interactions make it possible to represent the interactions in terms of a collection of the operational triplets listed in sections 4 and 5. We construct the classification rules for the human interactions using the operational triplet notation, and convert the rules to a decision-tree structure.

7. Rule-based Classification of Interaction

We learn the appropriate spatial/temporal constraints for individual interaction patterns, and organize the constraints into the rules in a decision-tree structure for classifying the two-person interactions.

The overall procedure for classifying human interaction is described in figure 11. If the two persons' torsos are far from each other (beyond a threshold), then the human interaction is undefined. If the torso distance is near (within the threshold), then the torso orientations are classified into different classes that form subtrees. The human interactions are represented in terms of a series of operational triplets aligned according to the spatio-temporal constraints to form the interaction phrase. Interaction classification is basically a procedure of pruning a decision tree by referring to a lookup table. The satisfaction of the specific spatial/temporal constraints controls the activation of the

NEUTRAL INTERACTIONS:
approach: torso moves forward and the distance between the two
         torsos decreases, and no salient motion of arm/leg is
         involved
depart: torso moves forward and the distance between the two
         torsos increases, and no salient motion of arm/leg is
         involved
point: arm stretches to the upper body or head of the other
         person, and torso(s) stay stationary

POSITIVE INTERACTIONS:
shake hands: two torsos face each other, and the arms of the two
           persons stretch simultaneously and touch each other.
standing hand in hand: two torsos are side-by-side and face
                  in the same direction,
                  and the arms of the two persons touch
                  each other simultaneously
hug: two torsos touch each other, and the arms of the two
         persons touch the other person's upper body simultaneously.

NEGATIVE INTERACTIONS:
punch: arm stretches to the upper body or head of the person,
         and the torso of the other person moves backward
push: arm stretches and contacts the other person's upper body,
         and the torso of the other person moves backward later
kick: leg raises and stretches to the other person
         and the torso of the other person moves backward

Figure 10. Operational definitions of human interaction types

Figure 11. The procedure to recognize human interactions

8. Results and Conclusion

We have tested our methodology for the human interaction types on real data. The images used in this work are 320×240 pixels in size, obtained at a rate of 15 frames/sec. Six pairs of different interacting persons with various clothing were used to obtain the total 54 sequences (9 interactions $\times$ 6 pairs of persons) with 2445 frames total. The 6 sequences were tested for each of the 9 interaction types using the decision-tree, which classifies the pattern of 'co-occurrences' and 'sequential occurrences' of the triplets transformed from the video sequences to recognize the type of interaction.

The accuracies of the sequence classification for various interaction types are shown in table 1.

Example sequences of real persons and their corresponding results are presented in figs. 14 and15.

We have presented a new framework for describing human actions and interactions at a semantic level. Our method is based on the hierarchy of action concepts: static pose, dynamic gesture, single-person action and person-to-person interaction. Our method combines statistical methods for estimating poses/gestures and syntactic methods for verbal description. We adopt the linguistic 'verb argument structure' to represent human action in terms of $<$agent$-$motion$-$target$>$ triplets. Human interaction is represented by multiple triplets aligned according to spatial/temporal constraints between the actions. Various two-person interactions are described at a detailed level in terms

proper rules to generate a verbal description of the interaction. The decision tree is represented in Fig. 12 The satisfaction of the specific spatial/temporal constraints controls the activation of the proper rules to generate a verbal description of the interaction. For example, the decision tree's initial part (i.e., levels 1–3) is shown in figure 13. If the two persons' torsos are far from each other (beyond a threshold), then the human interaction is undefined. If the torso distance is near (within the threshold), then the torso orientations are classified into different classes: $S^1, S^2, \ldots, S^6$ defined in section 5. In each of the subtrees with roots, $\tau_1 - \tau_6$, branching is determined by torso orientations first at level 4, and by the operational triplets involved in the specific interactions at level 5. The individual leaf node in Fig. 12 represents an interaction type and it may be reached with different combinations of the operational triplets.
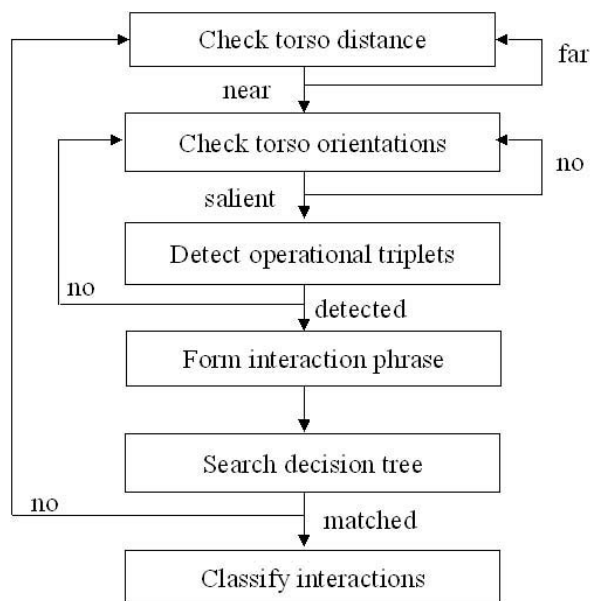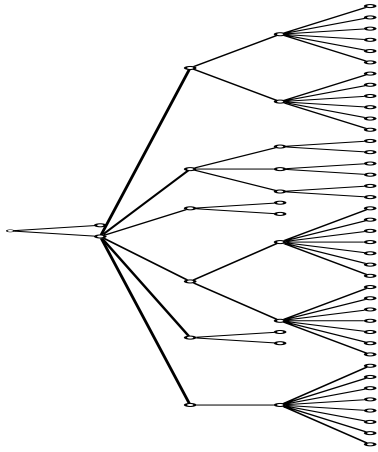
Figure 12. The overall decision tree with 5 layers: the leftmost layer is layer 1, and the rightmost layer is layer 5.
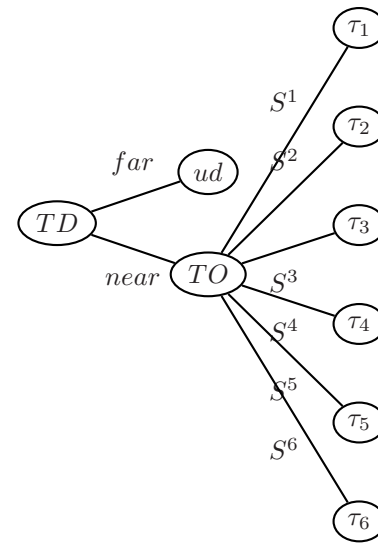

Figure 13. Decision tree's initial part (layer 1 through 3). Branching is determined by torso distance (TD) and torso orientation (TO). 'ud' denotes 'undefined'.

index	interaction	total	correct	accuracy
(a)	approach	6	6	1
(b)	depart	6	6	1
(c)	point	6	4	0.67
(d)	hand-in-hand	6	5	0.83
(e)	shake hands	6	6	1
(f)	hug	6	3	0.5
(g)	punch	6	4	0.67
(h)	kick	6	5	0.83
(i)	push	6	3	0.5
total				0.78

Table 1. Recognition accuracy.

of user-friendly verbal description of single-person actions. Our method properly describes positive, neutral, and negative interactions occurring between two persons.

References

[1] J. F. Allen and G. Ferguson. Actions and events in interval temporal logic. *Journal of Logic and Computation*, 4(5):531–579, 1994.

[2] D. Ayers and M. Shah. Monitoring human behavior from video taken in an office environment. *Image and Vision Computing*, 19(12):833–846, October 2001.

[3] A. Kojima, T. Tamura, and K. Fukunaga. Textual description of human activities by tracking head and hand motions. In *International Conference on Pattern Recognition*, volume 2, pages 1073–1077, 2002.

[4] R. Mann, A. Jepson, and J. Siskind. Computational perception of scene dynamics. *Computer Vision and Image Understanding*, 65 (2):113–128, 1997.

[5] R. Nevatia, T. Zhao, and S. Hongeng. Hierarchical language-based representation of events in video streams. In *IEEE Conf. on Computer Vision and Pattern Recognition*, volume 4, pages 39–46, 2003.

[6] N. M. Oliver, B. Rosario, and A. P. Pentland. A Bayesian Computer Vision System for Modeling Human Interactions. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 22(8):831–843, August 2000.

[7] S. Park and J. K. Aggarwal. Segmentation and tracking of interacting human body parts under occlusion and shadowing. In *IEEE Workshop on Motion and Video Computing*, pages 105–111, Orlando, FL, 2002.

[8] S. Park and J. K. Aggarwal. Recognition of two-person interactions using a hierarchical Bayesian network. In *ACM SIGMM International Workshop on Video Surveillance*, pages 65–76, Berkeley, CA, USA, 2003.

[9] A. Sarkar and W. Tripasai. Learning verb argument structure from minimally annotated corpora. In *Proceedings of COLING 2002*, Taipei, Taiwan, August 2002.

[10] K. Sato and J.K. Aggarwal. Temporal spatio-velocity transform and its application to tracking and interaction. *to appear in Computer Vision and Image Understanding*, 2004.

[11] R. C. Schank and R. P. Abelson. *Scripts, plans, goals, and understanding : an inquiry into human knowledge structures*. John Wiley and Sons Inc., Hillsdale, N.J., 1977.

[12] T. Starner and A. Pentland. Real-time american sign language recognition from video using hidden markov models. In *Proc. of SCV95*, pages 265–270, 1995.
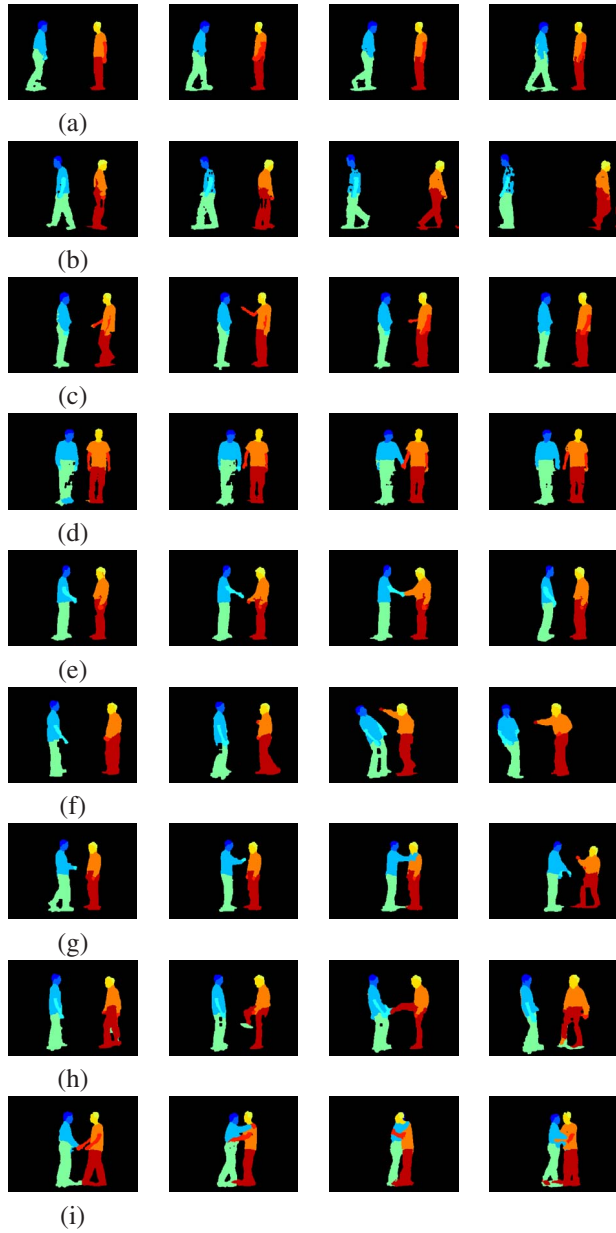
Figure 14. Subsampled frames of various interaction sequences. (a) *approach*, (b) *depart*, (c) *point*, (d) *stand hand-in-hand*, (e) *shake hands*, (f) *punch*, (g) *push*, (h) *kick*, and (i) *hug*.
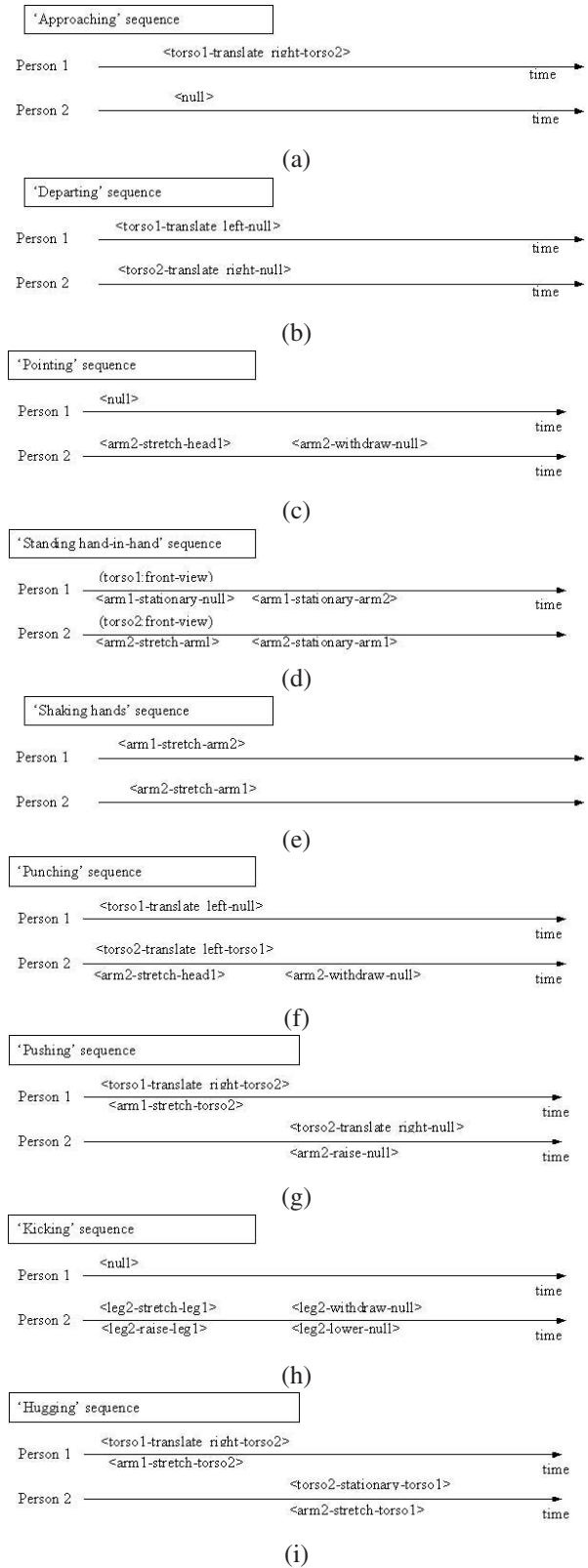

Figure 15. Semantic interpretation of two-person interactions corresponding to Fig. 14: (a) – (i).