Event Semantics in Two-person Interactions

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

This paper presents a method to represent two-person interactions at a semantic level with a natural language description. A human interaction is composed of two single-person actions, which in turn are made up of torso and arm/leg motions. We adopt the 'verb argument structure' in linguistics to represent human action in terms of <agent−motion−target> triplets. Various two-person interactions are represented at a detailed level using multiple triplets aligned along a time line according to the spatial/temporal constraints of the interactions. Our method provides a user-friendly natural-language description of various human interactions, and properly describes positive, neutral, and negative interactions occurring between two persons.

1. Introduction

Understanding human behavior in video data is essential in applications such as surveillance, video annotation/retrieval, and human-computer interfaces. One of the goals in video understanding is to describe the actions between persons at a semantic level. The semantic-level description of human actions and interactions leads to a natural language-based interface between a computer system and ordinary users. Designing a natural language-based human-computer interface is highly desired because of the rich structure of syntax and semantics representing domain-specific rules and contexts. Several researchers have pursued the notion of describing interaction between human and inanimate objects. Single-person activities in an office environment were described in [2, 4]. A single hand's manipulation of objects was interpreted in terms of the laws of physics in [5]. Event description of remote scenes in outdoor surveillance was presented in [6, 3]. 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 information about the individual body parts has not been addressed.

This paper presents a new method to describe 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. 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. In our previous work [7], we presented a method to segment and track multiple body parts in two-person interactions. The segmented body parts include head, upper body, arm(s), lower body, and leg(s). In [8] we presented a hierarchical Bayesian network to estimate body poses and gestures by parameterizing the body parts in terms of ellipses and convex hulls. The Bayesian network estimates instantaneous body poses, such as the orientation of the head, the hand position of the upper body, the foot position of the lower body, etc. Then, a dynamic Bayesian network is constructed to estimate temporal evolution of the poses along the sequence for recognizing the dynamic gestures of the body parts. In this paper we represent human behavior as intentional activity aimed at interacting with an environment that may include objects and other persons.

2. Interaction hierarchy

Our representation of human interaction is based on a hierarchy; a two-person *interaction* is a combination of single-person actions, and the single-person *action* is composed 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. 1.)

We represent the human body as both the autonomous *subject* and *object* in the two-person interaction. Each per-

Proceedings of the 17th International Conference on Pattern Recognition (ICPR'04)
1051-4651/04 $ 20.00 IEEE

IEEE COMPUTER SOCIETY

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 1. Human interaction hierarchy

son 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.

3. Building event semantics for two-person interactions

A gap exists between geometric information obtained from images and semantic information contained in natural language [4]. It is necessary to associate visual features with natural language verbs and symbols to build event semantics of two-person interactions.

The individual body part poses (fig. 2) are obtained from our Bayesian network, and the dynamic gestures of the body parts (fig. 3) are obtained from our dynamic Bayesian network [8].

Head_P = {A: *front-*, B: *left-*, C: *right-*, D: *rear-view*}
Torso_P = {A: *front-*, B: *left-*, C: *right-*, D: *rear-view*}
Arm_{Pv} = {A: *high*, B: *mid-high*, C: *mid-low*, D: *low*}
Arm_{Ph} = {A: *withdraw*, B: *intermediate*, C: *stretch*}
Leg_{Pv} = {A: *low*, B: *middle*, C: *high*}
Leg_{Ph} = {A: *withdraw*, B: *intermediate*, C: *stretch*}

Figure 2. Static poses of body parts.

Torso_G = {'move-right', 'move-left', 'stay stationary'}
Arm_{Gv} = {'raise', 'lower'}
Arm_{Gh} = {'stretch','withdraw'}
Leg_{Gv} = {'raise', 'lower'}
Leg_{Gh} = {'stretch','withdraw'}

Figure 3. Dynamic gestures of body parts.

We conceptualize human actions in terms of an *operation triplet* defined as $triplet = <agent-motion-target>$ according to the theory of 'verb argument structure' in linguistics [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_i | s_i = various body parts as agent term}
= {head, torso, arm, leg}
motion set: V = { v_j | v_j = body-part motion, called 'action-atoms'}
= {stay, move left, move right, raise, lower, stretch, withdraw}
target set: 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.

The task of event understanding is equivalent to the task of transforming a video sequence to a verbal description using the various *operation triplets* filled with the appropriate vocabulary terms in fig. 4. The transformation rules are determined by domain-specific knowledge about human interactions.

Multiple triplets may be involved in a specific action depending on the complexity of the corresponding action, because multiple body parts may be involved in the action. We represent the action of the j-th person as Act^j

$$Act^j = \begin{pmatrix} torso\ pose \\ torso\ triplet \\ arm\ triplet \\ leg\ triplet \end{pmatrix} \quad (1)$$

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 $null$.

IEEE COMPUTER SOCIETY

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 $Interact^{ij}_{\Delta_{tk}}$

$$Interact^{ij}_{\Delta_{tk}} = \begin{pmatrix} Act^i \\ Act^j \end{pmatrix}$$

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 is essential because it constitutes the *pushing* gesture per se. In contrast, the target person's arm motion may not occur if the pushing action is not strong. The target person's arm motion is incidental in constituting the *pushing* interaction. The definitions of the *interaction* classes determine which components are essential and which are incidental, and the decision requires either user discretion or training from data. Based on the dictionary definitions of the interactions, we manually construct the classification rules for the human interactions using the *essential* operational triplets, and convert the rules to a decision-tree structure.

4 Relative Constraints

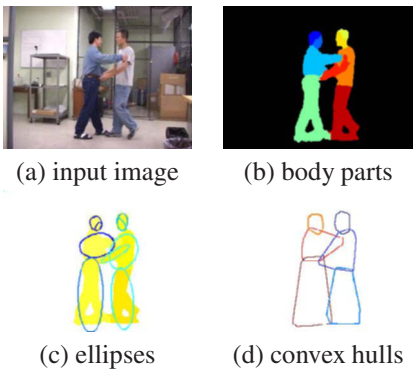
Knowing the individual body poses/gestures is not enough for the recognition of *interaction* between persons. The linkage between the agent and the target requires information about the *relative* positions of the two persons' individual body parts. Information about the relative positions of the body 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.

The spatial constraints of two-person interactions are defined in terms of the torso poses and the distance between two torsos. For example, 'standing hand-in-hand' requires that the two persons' torsos are side by side and facing in the same direction, whereas 'pointing at the opposite person' requires that the two torsos face one another.

The temporal constraints of two-person interactions are defined by causal and coincident relations of multiple actions that represent the two persons' body-part gestures. We adopt Allen's interval temporal logic [1] to represent the causal and coincident relations of two action events in the temporal domain. For example, a 'pushing' interaction involves the i-th person's causal action toward the j-th person $\epsilon_1 = <\text{arm}_i - \text{stretch} - \text{torso}_j>$ followed by the j-th person's effective action $\epsilon_2 = <\text{torso}_j - \text{move-backward} - \text{null}>$.

5. Mapping 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 ellipse and the convex hull of individual body part [8] as shown in fig. 5.



(a) input image (b) body parts

(c) ellipses (d) convex hulls

Figure 5. An example of feature extraction from the 'hugging' sequence. Ellipses and convex hulls are used for pose estimation.

Our method is based on hierarchical processing at multiple levels. At pixel level, individual pixels are classified into homogeneous blobs according to color. 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 multiple body part regions are tracked along the image sequence [7]. The information about body poses and gestures is estimated by a hierarchical Bayesian network (BN) using ellipses and convex hulls as image features. The basic vocabulary sets in figs. 2 and 3 are obtained by the hierarchical BN [8].

The linkage between the agent and the target is achieved by estimating the proximity of the two persons' individual body parts. The proper 'agent' and 'target' terms are chosen from the vocabulary sets in fig. 4 by combining the salient moving body part of a person and the proximal body part of the other person. Knowing the relative positions of the body

IEEE COMPUTER SOCIETY

parts leads to composite verb concepts such as 'approach to the other person', 'touch the chest', etc.

Person-to-person interaction may involve either multiple simultaneous actions or multiple sequential actions. We distinguish 'co-occurrence' and 'sequential occurrence' of actions. For example, "hugging" may involve two *simultaneous* actions of $<\text{arm}_1-\text{stretch}-\text{torso}_2>$ and $<\text{arm}_2-\text{stretch}-\text{torso}_1>$, whereas "punching" may involve two *sequential* actions of $<\text{arm}_2-\text{stretch}-\text{head}_1>$ and then $<\text{torso}_1-\text{move-backward}-\text{null}>$. (See fig. 6.)

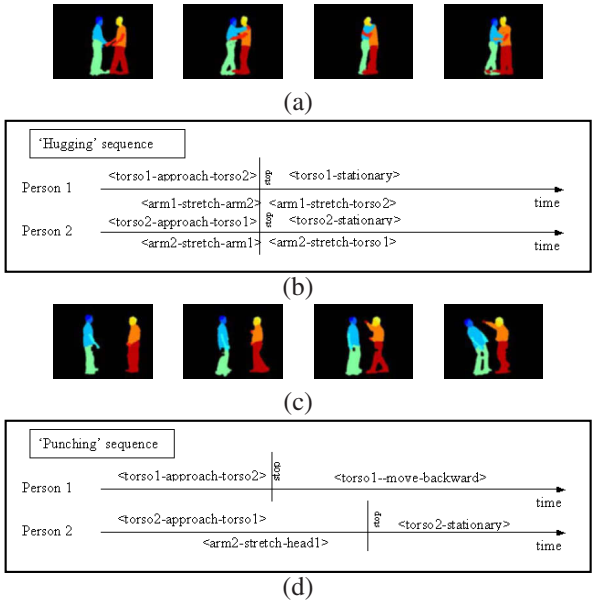


Figure 6. Examples of interaction sequences: (a) 'hugging' and (b) its semantic interpretation, (c) 'punching' and (d) its semantic interpretation.

6. Results and Conclusion

We have tested our methodology for the following human interaction types on real data: (1) *shaking hands*, (2) *standing hand-in-hand*, (3) *hugging*, (4) *approaching*, (5) *departing*, (6) *pointing*, (7) *punching*, (8) *kicking*, and (9) *pushing*. Interactions (1)−(3), (4)−(6), and (7)−(9) correspond to positive, neutral, and negative interactions, respectively. 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 × 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 interaction types (1) - (9) were 100, 83, 50, 100, 100, 67, 67, 83, and 50 percent, respectively. The overall accuracy is 78 percent. Examples of the time line of the interaction behaviors represented in terms of actions and gestures are shown in fig. 6.

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 verb argument structure in linguistics to represent human action in terms of $<\text{agent}-\text{motion}-\text{target}>$ 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 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] Y. Ivanov and A. Bobick. Recognition of visual activities and interactions by stochastic parsing. *IEEE transactions on Pattern Analysis and Machine Intelligence*, 22(8):852–872, 2000.

[4] 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.

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

[6] 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.

[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.

IEEE COMPUTER SOCIETY