Learning Human Interaction by Interactive Phrases

Yu Kong[1,3], Yunde Jia[1], and Yun Fu[2]

[1] Beijing Laboratory of Intelligent Information Technology
School of Computer Science, Beijing Institute of Technology
Beijing 100081, P.R. China
[2] Department of ECE and College of CIS, Northeastern University, Boston, MA
[3] Department of CSE, State University of New York, Buffalo, NY
{kongyu,jiayunde}@bit.edu.cn,y.fu@neu.edu

Abstract. In this paper, we present a novel approach for human interaction recognition from videos. We introduce high-level descriptions called *interactive phrases* to express binary semantic motion relationships between interacting people. Interactive phrases naturally exploit human knowledge to describe interactions and allow us to construct a more descriptive model for recognizing human interactions. We propose a novel hierarchical model to encode interactive phrases based on the latent SVM framework where interactive phrases are treated as latent variables. The interdependencies between interactive phrases are explicitly captured in the model to deal with motion ambiguity and partial occlusion in interactions. We evaluate our method on a newly collected BIT-Interaction dataset and UT-Interaction dataset. Promising results demonstrate the effectiveness of the proposed method.

1 Introduction

In recent years, interaction recognition has received much attention in computer vision community with applications in areas such as video analysis and surveillance [1–3]. A popular idea for this task in previous approaches is to utilize contextual information of action classes, e.g. human poses, object classes or object locations [2–6], to capture co-occurrence relationships of entities in interactions (human actions or objects). However, misclassifications remain in some challenging situations. This would be probably due to the co-occurrence relationships are not expressive enough to deal with interactions with large variations. For example, in "boxing" interaction, the defender can perform diverse semantic actions to protect himself, e.g. step back, crouch, or even hit back. This requires us to define all possible action co-occurrence relationships and provide sufficient training data for each co-occurrence case, which are infeasible.

We present *interactive phrases*, binary motion relationship descriptions, for recognizing complex human interactions. Essentially, these phrases are descriptive primitives shared in all interaction classes and characterize an interaction from different angles, e.g. motion relationships between arms, legs, and torsos,

A. Fitzgibbon et al. (Eds.): ECCV 2012, Part I, LNCS 7572, pp. 300–313, 2012.

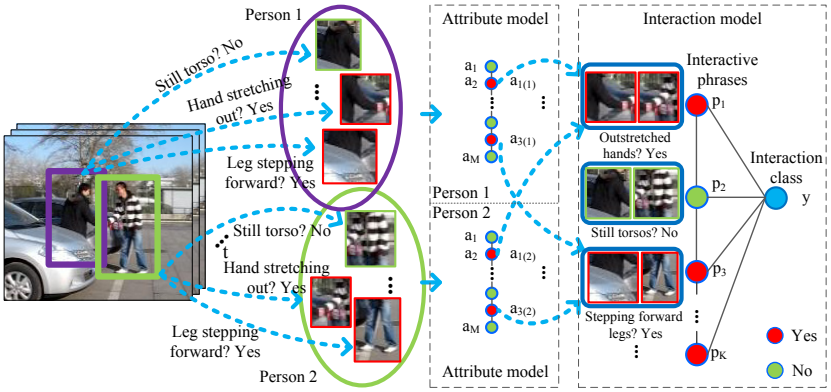


Fig. 1. Framework of our interactive phrase method

etc. Consequently, we can simply use compositions of binary phrases to describe interactions with variations rather than considering all possible action co-occurrences in an interaction class. Moreover, interactive phrases provide a novel type of contextual information, i.e. phrase context, for human interaction recognition. Since phrases describe all the details of an interaction, they provide a strong context for each other and are more expressive than the action context used in previous work [3]. The use of interactive phrases allows us to build a more descriptive model, which can be used to recognize human interactions with large variations (e.g. interactions with partial occlusion).

The significance of interactive phrases is that they incorporate rich human knowledge about motion relationships and thus allow us to better represent complex human interactions. Moreover, they bridge the gap between low-level features and high-level interaction classes and thus improve recognition accuracy. Compared with attributes of objects [7–9], which focus on the intrinsic properties of an object (e.g. "furry", "metal"), interactive phrases provide an effective way to describe motion relationships between interacting people. In other words, attributes represent *unary* relationships of an object while interactive phrases describe high-order relationships between people. In this work, we focus on *binary* relationships in human interactions.

The goal of this paper is to model interactive phrases of human interactions so that complex interactions can be better represented. The flowchart of our method is shown in Fig.1. Given training videos, our method learns motion attributes for each interacting person. These human understandable attributes characterize individual actions and serve as the input of our interaction model. Each interactive phrase is associated with one attribute of people in interactions to express motion relationships between them. To deal with the inherent intra-class variability of each class, we treat interactive phrases as latent variables and formulate the interaction recognition problem based on the latent SVM framework [10, 11]. We explicitly model the co-occurrence relationships between interactive phrase

pairs to address the problems of motion ambiguity and partially occlusion in interactions. Using such co-occurrence relationships will provide a strong context for phrases and make them fit in the context.

2 Related Work

In recent years, human-human and human-object interaction recognition have received increasing attention in computer vision community. Great progress has been made by capturing co-occurrence contextual information. Lan et al.[3] captured action context to aid interaction recognition. Perez et al.[1] employed a structured learning technique to capture spatial relationships between interacting individuals. Choi et al.[2] utilized spatial-temporal crowd context to recognize human interactions. Vahdat et al.[12] represented each individual by a set of key poses and formulated spatial and temporal relationships among key poses in their model. Gupta et al.[4] fused context from object reaction, object class, and manipulation motion into a single framework for analyzing human-object interactions in videos. Yao and Fei-Fei [5] explored mutual context of objects and human poses in human-object interaction recognition. Desai et al.[6] proposed a contextual model utilizing relative locations of objects and human poses. Approaches proposed in [13, 14] treat interacting people as a group and recognize interactions by computing group motion similarities in videos.

To our best knowledge, few attempts have been made to utilize high-level descriptions for human interaction recognition. A related work to ours is Ryoo and Aggarwal [15] in which the context-free grammar is employed to describe spatial and temporal relationships between people. The key difference between our work and theirs is that our method integrates high-level descriptions and interaction classes into a unified probabilistic model. In addition, these descriptions (interactive phrases) are treated as latent variables to deal with the intra-class variability. Our work is also different from [3]. Our model depends on high-level descriptions, i.e. interactive phrases, while their method relies on action co-occurrence. Our method decomposes action co-occurrence into phrase co-occurrence, which provide a more effective way to represent complex interactions.

High-level description-based methods have shown their power in object recognition [8, 9] and action recognition [16]. In these methods, attributes are utilized to describe intrinsic properties of an object (e.g. color, shape) or spatial-temporal visual characteristics of an actor (e.g. single leg motion, torso up-down motion). Our interactive phrases are different from attributes in objects [9] and actions [16]. In their work, attributes represent unary relationships (intrinsic properties of an object or an action), which are directly inferred from low-level features. By contrast, interactive phrases describe binary motion relationships and are built based on semantic motion attributes of each interacting person.

Our work is partially inspired by [17] which used language constructs such as "prepositions" (e.g. above, below) and "comparative adjectives" (e.g. brighter, smaller) to express relationships between objects. The difference is that our interactive phrases describe motion relationships of people rather than spatial

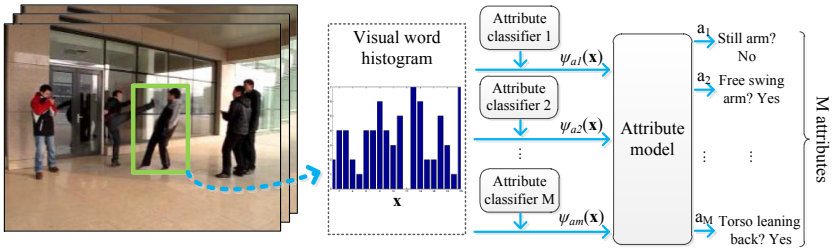


Fig. 2. Framework of detecting motion attributes from videos

relationships of static objects. Moreover, interactive phrases are built upon semantic motion attributes rather than inferred from object classes.

3 Our Method

Our method consists of two main components, the attribute model and the interaction model. The attribute model is utilized to jointly detect all attributes for each person, and the interaction model is applied to recognize an interaction. In this work, we mainly focus on recognizing interactions between two people.

3.1 Attribute Model

We utilize motion attributes to describe individual actions [16], e.g. "arm raising up motion", "leg stepping backward motion", etc. In interactions, both of the two interacting people have the same attribute vocabulary but with different values. Those motion attributes can be inferred from low-level motion features (Fig.2), for example, spatiotemporal interest points [18]. We assume there are certain interdependencies between attribute pairs (a_j, a_k). For instance, attributes "arm stretching out motion" and "leg stepping forward motion" tend to appear together in "handshake". The interdependencies are greatly helpful in dealing with incorrect attributes caused by motion ambiguity and partial occlusion.

We adopt a tree-structured undirected graph [19] $\mathcal{G}_a = (\mathcal{V}_a, \mathcal{E}_a)$ to represent the configurations of attributes. A vertex $a_m \in \mathcal{V}_a$ $(m = 1, \cdots, M)$ corresponds to the m-th attribute and an edge $(a_j, a_k) \in \mathcal{E}_a$ corresponds to the dependency between the two attributes. We use a discriminative function $g_\lambda : \mathbb{X} \times \mathbb{A} \to \mathbb{R}$ to score each training example $(\mathbf{x}, \mathbf{a})$: $g_\lambda(\mathbf{x}, \mathbf{a}) = \lambda^{\mathrm{T}}\Phi(\mathbf{x}, \mathbf{a})$, where $\mathbf{x}$ denotes the feature of a person in an interaction and $\mathbf{a} = (a_1, \cdots, a_M)$ is a binary attribute vector. $a_m = 0$ means the m-th attribute is absent and $a_m = 1$ denotes the attribute is present. We define $\lambda^{\mathrm{T}}\Phi(\mathbf{x}, \mathbf{a})$ as a summation of potential functions:

$$\lambda^{\mathrm{T}}\Phi(\mathbf{x}, \mathbf{a}) = \sum_{a_j \in \mathcal{V}_a} \lambda_{a_j}^{\mathrm{T}} \phi_1(\mathbf{x}, a_j) + \sum_{(a_j, a_k) \in \mathcal{E}_a} \lambda_{a_j a_k}^{\mathrm{T}} \phi_2(a_j, a_k), \tag{1}$$

where $\lambda = \{\lambda_{a_j}, \lambda_{a_j a_k}\}$ is model parameter. In our work, graph structure $\mathcal{E}_a$ is learned by the Chow-Liu algorithm [20]. The potential functions in Eq.(1) are summarized as follows.

Unary potential $\lambda_{a_j}^{\mathrm{T}} \phi_1(\mathbf{x}, a_j)$ provides the score for an attribute a_j and is used to indicate the presence of a_j given the motion feature $\mathbf{x}$. Parameter λ_{a_j} is a template for an attribute a_j. The feature function $\phi_1(\mathbf{x}, a_j)$ models the agreement between observation $\mathbf{x}$ and motion attribute a_j, and is given by

$$\phi_1(\mathbf{x}, a_j) = \delta(a_j = u) \cdot \psi_{a_j}(\mathbf{x}). \tag{2}$$

Here, $\delta(\cdot)$ denotes an indicator function, $u \in \mathcal{A}$ denotes a state of the attribute a_j, where $\mathcal{A}$ is the attribute space. Instead of keeping $\psi_{a_j}(\mathbf{x})$ as a high-dimensional feature vector, we represent it as the score output of a linear SVM trained with attribute a_j. Similar tricks have been used in [9, 21].

Pairwise potential $\lambda_{a_j a_k}^{\mathrm{T}} \phi_2(a_j, a_k)$ captures the co-occurrence of a pair of attributes a_j and a_k, for example, the co-occurrence relationships between attributes "torso bending motion" and "still leg" in "bow". Parameter $\lambda_{a_j a_k}$ is a 4-dimensional vector representing the weights for all configurations of a pair of attributes. The feature function $\phi_2(a_j, a_k)$ models the co-occurrence relationships of two attributes. We define $\phi_2(a_j, a_k)$ for a co-occurrence (u, v) as

$$\phi_2(a_j, a_k) = \delta(a_j = u) \cdot \delta(a_k = v). \tag{3}$$

3.2 Interaction Model

Interactive phrases encode human knowledge about motion relationships between people. The phrases are built on attributes of two interacting people and utilized to describe their co-occurrence relationships. Let p_j be the j-th phrase associated with two people's attributes $a_{j(1)}$ and $a_{j(2)}$[1]. In the interaction model, we use $a_{j(i)}$ to denotes the attribute of the i-th person that links to the j-th phrase. For example, phrase p_j "cooperative interaction" is associated with two people's attributes $a_{j(1)}$ and $a_{j(2)}$ "friendly motion". Note that $a_{j(i)}$ and $a_{k(i)}$ could be the same attribute but link to different phrases. We also assume that there are certain interdependencies between some phrase pairs (p_j, p_k). For example, phrases "interaction between stretching out hands" and "interaction between stepping forward legs" are highly correlated in "handshake".

We employ an undirected graph $\mathcal{G}_p = (\mathcal{V}_p, \mathcal{E}_p)$ to encode the configurations of phrases. A vertex $p_j \in \mathcal{V}_p$ $(j = 1, \cdots, K)$ corresponds to the j-th phrase and an edge $(p_j, p_k) \in \mathcal{E}_p$ corresponds to the dependency between the two phrases. Note that intra-class variability leads to different interactive phrase values in certain interaction classes. For instance, in "handshake", some examples have interactive phrase p_j "interaction between stepping forward legs" but some do not. In addition, labeling attributes is a subjective process and thus would influence the values of interactive phrases. We deal with this problem by treating phrases as

[1] Please refer to the supplemental material to see details about the connectivity patterns of interactive phrases and attributes.

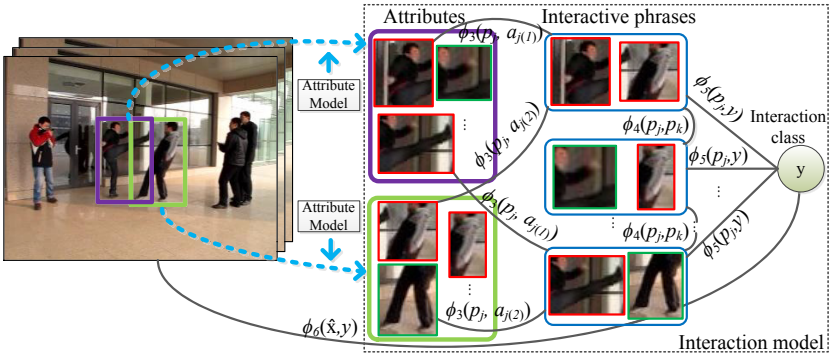


Fig. 3. The unary, pairwise and global interaction potentials in the interaction model

latent variables and formulating the classification problem based on the latent SVM framework [10, 11].

Given training examples $\{\hat{\mathbf{x}}^{(n)}, y^{(n)}\}_{n=1}^{N}$, we are interested in learning a discriminative function $f_{\mathbf{w}}(\hat{\mathbf{x}}, \hat{\mathbf{a}}, y) = \max_{\mathbf{p}} \mathbf{w}^{\mathrm{T}}\Phi(\hat{\mathbf{x}}, \hat{\mathbf{a}}, \mathbf{p}, y)$. Here $\hat{\mathbf{x}} = (\mathbf{x}_1, \mathbf{x}_2)$ is raw features of two interacting people, $\hat{\mathbf{a}} = (\mathbf{a}_1, \mathbf{a}_2)$ denotes two people's attributes, $\mathbf{p} = (p_1, \cdots, p_K)$ is a binary vector of phrases, and y is an interaction class, where $p_k \in \mathcal{P}$ and $y \in \mathcal{Y}$. To obtain $\mathbf{a}_1$ and $\mathbf{a}_2$, we run the attribute model twice with corresponding features. We define $\mathbf{w}^{\mathrm{T}}\Phi(\hat{\mathbf{x}}, \hat{\mathbf{a}}, \mathbf{p}, y)$ as a summation of potential functions:

$$\begin{aligned}\mathbf{w}^{\mathrm{T}}\Phi(\hat{\mathbf{x}}, \hat{\mathbf{a}}, \mathbf{p}, y) = &\sum_{p_j \in \mathcal{V}_p} \sum_{i=1}^{2} \mathbf{w}_{p_j a_{j(i)}}^{\mathrm{T}} \phi_3(p_j, a_{j(i)}) + \sum_{p_j \in \mathcal{V}_p} \mathbf{w}_{p_j y}^{\mathrm{T}} \phi_4(p_j, y) \\ &+ \sum_{(p_j, p_k) \in \mathcal{E}_p} \mathbf{w}_{p_j p_k}^{\mathrm{T}} \phi_5(p_j, p_k) + \mathbf{w}_{\hat{\mathbf{x}} y}^{\mathrm{T}} \phi_6(\hat{\mathbf{x}}, y), \end{aligned} \tag{4}$$

where $\mathbf{w} = \{\mathbf{w}_{p_j a_{j(i)}}, \mathbf{w}_{p_j p_k}, \mathbf{w}_{p_j y}, \mathbf{w}_{\hat{\mathbf{x}}}\}$ is model parameter. Similar to the attribute model, we use the Chow-Liu algorithm [20] to learn graph structure $\mathcal{E}_p$ in the interaction model. The potential functions are enumerated as follows.

Unary potential $\mathbf{w}_{p_j a_{j(i)}}^{\mathrm{T}} \phi_3(p_j, a_{j(i)})$ models the semantic relationships between an interactive phrase p_j and its associated attribute $a_{j(i)}$. Each interactive phrase in this paper is associated with one attribute of each interacting person. Parameter $\mathbf{w}_{p_j a_{j(i)}}$ is a 4-dimensional vector encoding the weights for all configurations between a phrase and an attribute, and feature function $\phi_3(p_j, a_{j(i)})$ models the agreement between them. The feature function $\phi_3(p_j, a_{j(i)})$ for a configuration (h, u), where $h \in \mathcal{P}$ and $u \in \mathcal{A}$, is given by

$$\phi_3(p_j, a_{j(i)}) = \delta(p_j = h) \cdot \delta(a_{j(i)} = u). \tag{5}$$

Unary potential $\mathbf{w}_{p_j y}^{\mathrm{T}} \phi_4(p_j, y)$ indicates that how likely the interaction class is y and the j-th interactive phrase is p_j. Feature function $\phi_4(p_j, y)$ is used to

encode the semantic relationships between an interaction class y and a phrase p_j. We define the feature function for a relationship (h, b), where $b \in \mathcal{Y}$, as

$$\phi_4(p_j, y) = \delta(p_j = h) \cdot \delta(y = b). \tag{6}$$

Parameter $\mathbf{w}_{p_j y}$ indicates the weight encoding valid relationships between a phrase p_j and an interaction class y.

Pairwise potential $\mathbf{w}_{p_j p_k}^{\mathrm{T}} \phi_5(p_j, p_k)$ captures the co-occurrence of a pair of interactive phrases (p_j, p_k). Parameter $\mathbf{w}_{p_j p_k}$ is a 4-dimensional vector denoting the weights of all possible configurations of a pair of phrases. Feature function $\phi_5(p_j, p_k)$ in the pairwise potential captures the co-occurrence relationships between two phrases. We define $\phi_5(p_j, p_k)$ for a relationship (h_1, h_2) as

$$\phi_5(p_j, p_k) = \delta(p_j = h_1) \cdot \delta(p_k = h_2). \tag{7}$$

Global interaction potential $\mathbf{w}_{\hat{\mathbf{x}}y}^{\mathrm{T}} \phi_6(\hat{\mathbf{x}}, y)$ provides the score measuring how well the raw feature $\hat{\mathbf{x}}$ matches the interaction class template $\mathbf{w}_{\hat{\mathbf{x}}y}$. The feature function $\phi_6(\hat{\mathbf{x}}, y)$ models the dependence between an interaction class with its corresponding video evidence. The feature function for interaction class $y = b$ is defined as

$$\phi_6(\hat{\mathbf{x}}, y) = \delta(y = b) \cdot \hat{\mathbf{x}}. \tag{8}$$

3.3 Learning and Inference

Parameter learning in our work consists of two steps: learning parameters of the attribute model and learning parameters of the interaction model.

The max-margin conditional random field formulation [22] is adopted to train the attribute model given training examples $\mathcal{D}_a = \{\mathbf{x}^{(n)}, \mathbf{a}^{(n)}\}_{n=1}^{N_a}$:

$$\begin{aligned} &\min_{\lambda, \xi} \ \frac{1}{2}\|\lambda\|^2 + C \sum_n \xi_n \\ &\text{s.t. } \lambda^{\mathrm{T}} \Phi(\mathbf{x}^{(n)}, \mathbf{a}^{(n)}) - \lambda^{\mathrm{T}} \Phi(\mathbf{x}^{(n)}, \mathbf{a}) \geqslant \Delta(\mathbf{a}, \mathbf{a}^{(n)}) - \xi_n, \forall n, \forall \mathbf{a}, \end{aligned} \tag{9}$$

where C is the trade-off parameter similar to that in SVMs, ξ_n is the slack variable for the n-th training example to handle the case of soft margin, and $\Delta(\mathbf{a}, \mathbf{a}^{(n)})$ is the 0-1 loss function.

Next, the latent SVM formulation [10, 11] is employed to train the parameter $\mathbf{w}$ of the interaction model given training examples $\mathcal{D} = \{\hat{\mathbf{x}}^{(n)}, \hat{\mathbf{a}}^{(n)}, y^{(n)}\}_{n=1}^{N}$, where $\hat{\mathbf{a}}^{(n)} = (\mathbf{a}_1^{(n)}, \mathbf{a}_2^{(n)})$ is the attributes of interacting people inferred by the trained attribute model:

$$\begin{aligned} &\min_{\mathbf{w}, \xi} \ \frac{1}{2}\|\mathbf{w}\|^2 + C \sum_n \xi_n \\ &\text{s.t. } \max_{\mathbf{p}} \mathbf{w}^{\mathrm{T}} \Phi(\hat{\mathbf{x}}^{(n)}, \hat{\mathbf{a}}^{(n)}, \mathbf{p}, y^{(n)}) \\ &\quad - \max_{\mathbf{p}} \mathbf{w}^{\mathrm{T}} \Phi(\hat{\mathbf{x}}^{(n)}, \hat{\mathbf{a}}^{(n)}, \mathbf{p}, y) \geqslant \Delta(y, y^{(n)}) - \xi_n, \forall n, \forall y. \end{aligned} \tag{10}$$

This optimization problem can be solved by the coordinate descent [10]. We first randomly initialize the model parameter $\mathbf{w}$ and then learn the parameter $\mathbf{w}$ by iterating the following two steps:

1. Holding $\mathbf{w}$ fixed, find the best interactive phrase configuration $\mathbf{p}'$ such that $\mathbf{w}^{\mathrm{T}}\Phi(\hat{\mathbf{x}}^{(n)}, \mathbf{a}^{(n)}, \mathbf{p}, y^{(n)})$ is maximized.
2. Holding $\mathbf{p}$ fixed, optimize parameter $\mathbf{w}$ by solving the problem Eq.(10).

In testing, our aim is to infer the interaction class of an unknown example: $y^* = \arg\max_{y \in \mathcal{Y}} f_{\mathbf{w}}(\hat{\mathbf{x}}, \hat{\mathbf{a}}, y)$. However, the attributes $\hat{\mathbf{a}}$ of two interacting people are unknown during testing. We solve this problem by finding the best attribute configuration $\mathbf{a}_i$ for the i-th person by running Belief Propagation (BP) in the attribute model: $\mathbf{a}_i = \arg\max_{\mathbf{a}_i} \lambda^{\mathrm{T}}\Phi(\mathbf{x}_i, \mathbf{a}_i)$. Then attributes $\hat{\mathbf{a}} = (\mathbf{a}_1, \mathbf{a}_2)$ is derived and utilized as the input for inferring the interaction class y. BP is also applied to find the best interactive phrase configuration $\hat{\mathbf{p}}$ in the interaction model: $f_{\mathbf{w}}(\hat{\mathbf{x}}, \hat{\mathbf{a}}, y) = \max_{\mathbf{p}} \mathbf{w}^{\mathrm{T}}\Phi(\hat{\mathbf{x}}, \hat{\mathbf{a}}, \mathbf{p}, y)$.

4 Experiments

4.1 Spatial-temporal Features

The spatial-temporal interest points [18] are detected from videos of human interaction. The spatial-temporal volumes around the detected points are extracted and represented by gradient descriptors. The dimensionality of gradient descriptors is reduced by PCA. All descriptors are quantized to 1000 visual-words using the k-means algorithm. Then videos are represented by histograms of visual-words.

4.2 Datasets

We compile a new dataset, BIT-Interaction dataset, to evaluate our method (see Fig.4) and add a list of 23 interactive phrases based on 17 attributes for all the videos (Please refer to the supplemental material for details.). Videos are captured in realistic scenes with cluttered background and bounding boxes of interacting people are annotated. People in each interaction class behave totally different and thus have diverse motion attributes (e.g. in some "boxing" videos, people step forward but in some videos they do not). This dataset consists of 8 classes of human interactions (bow, boxing, handshake, high-five, hug, kick, pat, and push), with 50 videos per class. The dataset contains a varied set of challenges including variations in subject appearance, scale, illumination condition and viewpoint. In addition, in most of videos, actors are partially occluded by body parts of the other person, poles, bridges, pedestrians, etc. Moreover, in some videos, interacting people have overlapping motion patterns with some irrelevant moving objects in the background (e.g. cars, pedestrians). We randomly choose 272 videos to train the interaction model and use the remaining 128 videos for testing. 144 videos in the training data are utilized to train the attribute model.

Fig. 4. Example frames of BIT-Interaction dataset. This dataset consists of 8 classes of human interactions: bow, boxing, handshake, high-five, hug, kick, pat, and push.

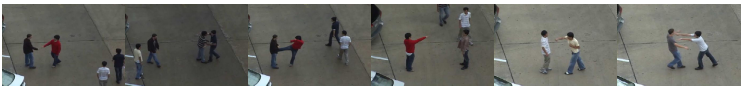


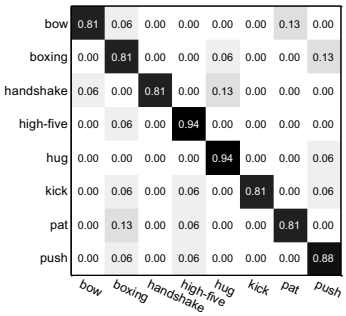
Fig. 5. Example frames of the UT-Interaction dataset. This dataset consists of 6 classes of human interactions: handshake, hug, kick, point, punch and push.

We also test our method on the UT-Interaction dataset [23]. We add 23 interactive phrases to this dataset based on 16 manually defined attributes (please refer to the supplemental material for details). This dataset consists of 6 types of human interactions: handshake, hug, kick, point, punch and push. Each type of interactions contains 10 videos, to provide 60 videos in total. Videos are captured in different scales and illumination conditions. Moreover, some irrelevant pedestrians are present in the videos. Example frames are displayed in Fig.5. We adopt the leave-one-out cross validation training strategy on this dataset.

4.3 Results

We conduct three experiments to evaluate our method. First, we test the proposed method on the BIT-Interaction dataset and compare our method with action context based method [3]. Next, we evaluate the effectiveness of components in the proposed method. At last, we compare our method with previous work [13, 14, 24] on the UT-Interaction dataset.

Evaluation on BIT-Interaction Dataset. In the first experiment, we test the proposed method on BIT-Interaction dataset. The confusion matrix is shown in Fig.6(a). Our method achieves 85.16% accuracy in classifying human interactions. Some of classification examples are displayed in Fig.6(b). Our method can recognize human interactions in some challenging situations, e.g. partially occlusion and background clutter. This is mainly due to the effect of interdependencies between interactive phrases. In such challenging scenarios, the interdependencies provide a strong context for the incorrectly inferred phrases and thus make them better fit in the context. As a result, human interaction in challenging situations can be correctly recognized. As we show in the last row in Fig.6(b), most of misclassifications are due to visually similar movements in different interaction classes (e.g. “boxing” and “pat”) and significant occlusion.



(a) Confusion matrix.

(b) Classification examples of our method

Fig. 6. Results of our method on BIT-Interaction dataset. In (b), correctly recognized examples are in the first two rows and misclassifications are in the last row.

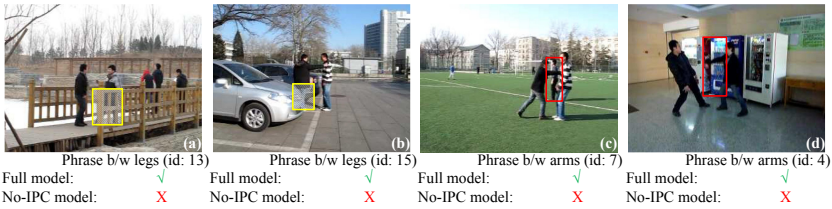
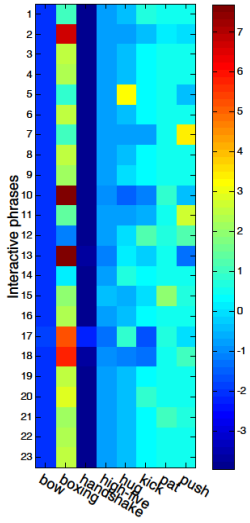


Fig. 7. Classification examples in BIT-Interaction dataset with occlusion and background noise. Yellow boxes denote occlusion and red boxes represent background noise. Please refer to supplemental material for the meaning of phrases according to their id.

To further investigate the effect of the interdependencies between interactive phrases, we remove the interdependencies $\phi_4(p_j, p_k)$ from the full model and compare the no-IPC model (the full model without $\phi_4(p_j, p_k)$) with the full model. Results in Fig.7 demonstrate that, without the interdependencies, the no-IPC model cannot accurately infer phrases from noisy motion attributes by the feature function $\phi_3(p_j, a_{j(i)})$. For example, in Fig.7(a) and (b), the phrases of occluded legs cannot be detected. However, the phrases of legs play key roles in recognizing "boxing" and "pat" (see Fig.8(b)). Without the key phrases, the interactions cannot be recognized. By comparison, the full model can use the feature function $\phi_4(p_j, p_k)$ to learn the interdependencies of a pair of interactive phrases from training data. Once some phrases cannot be inferred from the corresponding attributes, the interdependencies will play a strong prior on the phrases and thus facilitate the recognition task.

Interactive phrases have different importance to an interaction class. We illustrate the learned importance of interactive phrases to 8 interaction classes in Fig.8 (left). This figure demonstrates that our model learns some key interactive phrases to an interaction class (e.g. "interaction between embracing arms" in "hug" interaction). As long as these key interactive phrases are correctly detected, an interaction can be easily recognized. The learned top 3 key interactive phrases in all interaction classes are displayed in Fig.8 (right).



Interactions	Interactive phrases p_j	id
bow	b/w a bending torso and a still torso	17
	b/w still arms	1
	b/w a chest-level moving arm and a free swinging arm	2
boxing	b/w a chest-level moving arm and a leaning backward torso	10
	b/w a stepping forward leg and a stepping backward leg	13
	b/w a chest-level moving arm and a free swinging arm	2
handshake	b/w a bending torso and a still torso	17
	b/w outstretched hands	3
	b/w a leaning forward torso and a leaning backward torso	18
high-five	b/w a stepping forward leg and a still leg	15
	b/w raising up arms	4
	b/w outstretched hands	3
hug	b/w embracing arms	5
	b/w a bending torso and a still torso	17
	b/w stepping forward legs	14
kick	b/w still legs	12
	b/w leaning backward torsos	20
	b/w a kicking leg and a stepping backward leg	16
pat	b/w a stepping forward leg and a still leg	15
	b/w a leaning forward torso and a still torso	21
	b/w still legs	12
push	b/w two chest-level moving arms and a free swinging arm	7
	b/w two chest-level moving arms and a leaning backward torso	11
	b/w a leaning forward torso and a leaning backward torso	18

Fig. 8. [Best viewed in color] (Left) The learned importance of different interactive phrases in 8 interaction classes. (Right) The learned top 3 important interactive phrases for 8 interaction classes, where phrases of significant importance (their weights are at least 10 times greater than the others) are in blue and phrases never showed in the training data of an interaction class are in red. "b/w" is short for the word "between".

Table 1. Accuracies of our method and action co-occurrence based method [3]

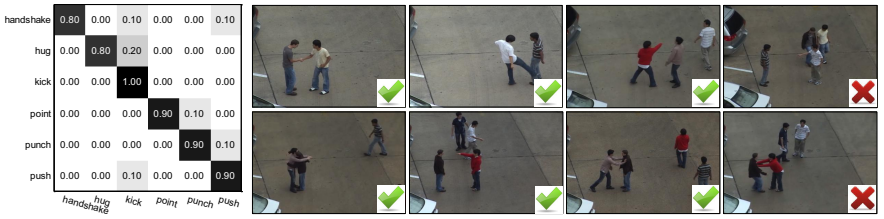
Methods	Lan et al.[3]	**Our method**
Accuracy	82.21%	**85.16%**

We also compare our description-based method with action co-occurrence based method [3] for human interaction recognition. To conduct a fair comparison, we use the same bag-of-words motion representation for the two methods. Results in Table 1 indicate that our method outperforms the action co-occurrence based method. The underlying reason is that the our method decomposes action co-occurrence relationships into a set of phrase co-occurrence relationships. The compositions of binary phrase variables allow us to represent interaction classes with large variations and thus make our method more expressive than [3].

Contributions of Components. In this experiment, we evaluate the contributions of components in the proposed method, including the interdependencies in the attribute model and the interaction model, respectively, and the interactive phrases. We remove these components from our method respectively, and obtain three different methods: the method without connections between attributes (no-AC method), the method without connections between interactive phrases (no-IPC method), and the interaction model without phrases (no-phrase

Table 2. Comparison results of accuracy (%) on the BIT-Interaction dataset

Methods	Overall	bow	boxing	handshake	high-five	hug	kick	pat	push
bag-of-words	70.31	81.25	75	50	75	81.25	68.75	62.5	68.75
no-phrase method	73.43	81.25	68.75	68.75	81.25	68.75	81.25	62.5	75
no-IPC method	80.47	81.25	68.75	81.25	87.5	81.25	81.25	75	87.5
no-AC method	81.25	81.25	62.5	81.25	87.5	93.75	81.25	81.25	81.25
Our method	**85.16**	81.25	81.25	81.25	93.75	93.75	81.25	81.25	87.5



(a) Confusion matrix. (b) Classification examples of our method

Fig. 9. Results of our method on UT-Interaction dataset. In (b), correctly recognized examples are in the first three columns and misclassifications are in the last column.

method). Our method is compared with these three methods as well as the baseline bag-of-words representation with a linear SVM classifier.

Table 2 indicates that our method outperforms all the baseline methods. Compared with the baseline bag-of-words method, the performance gain achieved by our method is significant due to the use of high-level knowledge of human interaction. Our method significantly outperforms the no-phrase method, which demonstrates the effectiveness of the proposed interactive phrases. Our method uses interactive phrases to better represent complex human interactions and thus achieves superior results. As expected, the results of the proposed method are higher than the no-IPC method and the no-AC method, which emphasize the importance of the interdependencies between interactive phrases and attributes, respectively. With the interdependencies, the proposed method can capture the co-occurrences of interactive phrases and thus reduces the number of incorrect interactive phrases. The interdependencies between individual attributes enable to capture the important relationships between individual attributes and reduce inaccurate attribute labels caused by noisy features and subjective attribute labeling. With the interdependencies in both attribute pairs and interactive phrase pairs, our method can recognize some challenging interaction videos and thus achieves higher results.

Results on UT-Interaction Dataset. We test our method on UT-Interaction dataset and show the confusion matrix in Fig.9(a). Our method achieves 88.33% recognition accuracy. Most of confusions are due to visually similar movements in two classes and the influence of moving objects in the background.

Table 3. Recognition accuracy (%) of methods on the UT-Interaction dataset

Methods	Overall	handshake	hug	kick	point	punch	push
bag-of-words	68.33	50	70	80	95	50	70
no-phrase method	70	60	60	70	80	90	60
no-AC method	80	60	80	80	90	90	80
no-IPC method	81.67	80	80	80	90	90	70
Ryoo & Aggarwal [13]	70.8	75	87.5	75	62.5	50	75
Yu et al.[14]	83.33	100	65	75	100	85	75
Ryoo [24]	85	–	–	–	–	–	–
Our method	**88.33**	80	80	100	90	90	90

We compare our full model with previous methods [13, 14, 24], the no-phrase method, the no-AC method and the no-IPC method, and adopt a bag-of-words representation with a linear SVM classifier as the baseline. Results in Table 3 show that our method outperforms all the methods in comparison. The value of our interactive phrases can be clearly seen from the performance differences between our method and the bag-of-words method as well as methods in [13, 14, 24]. Our method exploits rich human knowledge while the these methods only use low-level features. Our method significantly outperforms the no-phrase method, which demonstrates that the phrases provide an effective way to better represent complex interactions. Our full model achieves higher accuracies than the no-AC method and the no-IPC method, which shows the effectiveness of interdependencies in the attribute model and interaction model, respectively.

5 Conclusion

We have proposed interactive phrases, semantic descriptions of motion relationships between people, for human interaction recognition. Interactive phrases incorporate rich human knowledge and thus provide an effective way to represent complex interactions. We have presented a novel method to encode interactive phrases, which is composed of the attribute model and the interaction model. Extensive experiments have been conducted and showed the effectiveness of the proposed method.

The attributes and phrases rely on expert knowledge and are dataset specific. Scaling up attributes and phrases to general datasets remains an open problem. Possible solutions are: 1) cross-dataset techniques and 2) data-driven attributes. Structure learning techniques can also be adopted to adaptively determine the optimal connectivity pattern between attributes and phrases in new datasets. We plan to explore these in future work.

Acknowledgments. This research is supported in part by the Natural Science Foundation of China (NSFC) under grant No. 60905006, and the NSF CNS 1135660.

References

1. Patron-Perez, A., Marszalek, M., Zissermann, A., Reid, I.: High five: Recognising human interactions in tv shows. In: BMVC (2010)
2. Choi, W., Shahid, K., Savarese, S.: Learning context for collective activity recognition. In: CVPR (2011)
3. Lan, T., Wang, Y., Yang, W., Mori, G.: Beyond actions: Discriminative models for contextual group activities. In: NIPS (2010)
4. Gupta, A., Kembhavi, A., Davis, L.: Observing human-object interactions: Using spatial and functional compatibility for recognition. PAMI 31, 1775–1789 (2009)
5. Yao, B., Fei-Fei, L.: Modeling mutual context of object and human pose in human-object interaction activities. In: CVPR, pp. 17–24 (2010)
6. Desai, C., Ramanan, D., Fowlkes, C.: Discriminative models for static human-object interactions. In: CVPR Workshop on Structued Models in Computer Vision (2010)
7. Ferrari, V., Zisserman, A.: Learning visual attributes. In: NIPS (2007)
8. Lampert, C.H., Nickisch, H., Harmeling, S.: Learning to detect unseen object classes by between-class attribute transfer. In: CVPR (2009)
9. Wang, Y., Mori, G.: A Discriminative Latent Model of Object Classes and Attributes. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part V. LNCS, vol. 6315, pp. 155–168. Springer, Heidelberg (2010)
10. Felzenszwalb, P., McAllester, D., Ramanan, D.: A discriminatively trained, multiscale, deformable part model. In: CVPR (2008)
11. Wang, Y., Mori, G.: Max-margin hidden conditional random fields for human action recognition. In: CVPR, pp. 872–879 (2009)
12. Vahdat, A., Gao, B., Ranjbar, M., Mori, G.: A discriminative key pose sequence model for recognizing human interactions. In: ICCV Workshops, pp. 1729–1736 (2011)
13. Ryoo, M., Aggarwal, J.: Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities. In: ICCV, pp. 1593–1600 (2009)
14. Yu, T.H., Kim, T.K., Cipolla, R.: Real-time action recognition by spatiotemporal semantic and structural forests. In: BMVC (2010)
15. Ryoo, M., Aggarwal, J.: Stochastic representation and recognition of high-level group activities. IJCV 93, 183–200 (2011)
16. Liu, J., Kuipers, B., Savarese, S.: Recognizing human actions by attributes. In: CVPR (2011)
17. Gupta, A., Davis, L.S.: Beyond Nouns: Exploiting Prepositions and Comparative Adjectives for Learning Visual Classifiers. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part I. LNCS, vol. 5302, pp. 16–29. Springer, Heidelberg (2008)
18. Dollar, P., Rabaud, V., Cottrell, G., Belongie, S.: Behavior recognition via sparse spatio-temporal features. In: VS-PETS (2005)
19. Lafferty, J., McCallum, A., Pereira, F.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: ICML (2001)
20. Chow, C., Liu, C.: Approximating discrete probability distributions with dependence tree. IEEE Transactions on Information Theory 14, 462–467 (1968)
21. Desai, C., Ramanan, D., Fowlkes, C.: Discriminative models for multi-class object layout. In: ICCV (2009)
22. Taskar, B., Guestrin, C., Koller, D.: Max-margin markov networks. In: NIPS (2003)
23. Ryoo, M.S., Aggarwal, J.K.: UT-Interaction Dataset. In: ICPR Contest on Semantic Description of Human Activities, SDHA (2010), `http://cvrc.ece.utexas.edu/SDHA2010/Human_Interaction.html`
24. Ryoo, M.S.: Human activity prediction: Early recognition of ongoing activities from streaming videos. In: ICCV (2011)