Modeling Supporting Regions for Close Human Interaction Recognition

Yu Kong[1] and Yun Fu[1,2]

[1]Department of Electrical and Computer Engineering,
[2]College of Computer and Information Science.
Northeastern University, Boston, MA, USA
{yukong,yunfu}@ece.neu.edu

Abstract. This paper addresses the problem of recognizing human interactions with close physical contact from videos. Different from conventional human interaction recognition, recognizing close interactions faces the problems of ambiguities in feature-to-person assignments and frequent occlusions. Therefore, it is infeasible to accurately extract the interacting people, and the recognition performance of an interaction model is degraded. We propose a patch-aware model to overcome the two problems in close interaction recognition. Our model learns discriminative supporting regions for each interacting individual. The learned supporting regions accurately extract individuals at patch level, and explicitly indicate feature assignments. In addition, our model encodes a set of body part configurations for one interaction class, which provide rich representations for frequent occlusions. Our approach is evaluated on the UT-Interaction dataset and the BIT-Interaction dataset, and achieves promising results.

1 Introduction

Automatic understanding human actions in videos is important to several real-world applications, for example, video retrieval, video annotation, and visual surveillance. These videos often contain close interactions between multiple people with physical contact (e.g., “hug” and “fight”). This raises two major challenges in understanding this type of interaction videos: the occlusion of body parts and the ambiguity in feature assignments (features such as interest points are difficult to be uniquely assigned to a particular person in close interactions).

Unfortunately, the aforementioned problems are not addressed in existing interaction recognition methods [11, 1, 12, 24]. Methods in [11, 1] use trackers/detectors to roughly extract people, and assume interactions do not contain close physical contact (e.g., “walk” and “talk”). Their performance are limited in close interactions since the feature of one single person may contain noises from background or the other interacting people. Feature assignment problem is avoided in [12, 24] by treating the interaction people as a group. However, they do not utilize the intrinsic rich context of the interaction. Interest points have shown that they can be mainly associated with foreground moving human bodies in conventional

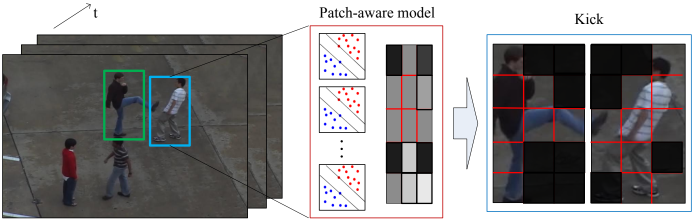


Fig. 1: Example of the inference results of our patch-aware model. Our model recognizes human interaction and discriminatively learns the supporting regions for each interacting people.

single-person action recognition methods [13, 21]. However, since multiple people present in interactions, it is difficult to accurately assign interest points to a single person, especially in close interactions. Therefore, action representations of people are extremely noisy and consequently degrade the recognition performance.

In this paper, we propose a novel patch-aware model for solving the aforementioned problems in close human interaction recognition from videos (Figure 1). Our model learns discriminative supporting regions for each interacting person, which accurately separate the target person from background. The learned supporting regions also indicate the feature-to-person assignments, which consequently help better represent individual actions. In addition, each interaction class associates with a variety of supporting region configurations, thereby providing rich and robust representations for different occlusion cases.

We propose a rich representation for close interaction recognition. Specifically, we introduce a set of binary latent variables for 3D patches indicating which subject the patch is associated with (background, person 1 or person 2), and encourage consistency of the latent variables across all of the training data. The appearance and structural information of patches is jointly captured in our model, which captures the motion and pose variations of interacting people. To address the challenge of an exponentially large label space, we use a structured output framework, employing a latent SVM [6]. During training, the model learns which patterns belong to the foreground and background, allowing for better labeling of body parts and identification of individual people. Results show that the learned supporting patches significantly facilitate the recognition task.

Our work differs from [11, 2, 1, 14] in that they can only deal with interactions that do not contain close physical contact (e.g. “queueing” and “talking”) while our method specifically aims at recognizing close interactions. Different from [24, 19, 17] which treat the interacting people as a group, our model provides fine-grained supporting regions for each interacting person, which allows us to recognize individual action. Although methods in [18, 22] can roughly extract each interacting person using a tracker or detector, they do not model 3D patches and background, and cannot accurately separate people. Our method, in contrast, captures different importance of 3D patches in interaction classes and thus can accurately separate people.

2 Related Work

Multi-person activity recognition has been receiving much attention in computer vision community. Methods in [2, 11] studied the collective activity recognition problem using crowd context. People in a collective activity have no close physical contact with each other and perform similar action, e.g. “crossing the road”, “talking”, or “waiting”. Specifically, Choi *et al.*[2] utilized human pose, velocity and spatiotemporal distribution of individuals to represent the crowd context information. They further developed a system that can simultaneously track multiple people and recognize their interactions [1]. Lan *et al.*[11] represented crowd context by action co-occurrence of interacting people. Odashima *et al.*[14] proposed the Contextual Spatial Pyramid to detect the action of multiple people.

Human interactions, e.g. “hug”, “push”, and “hi-five”, usually involve frequent close physical contact. Perez *et al.*[15] investigated interaction recognition between two people in realistic scenarios. They adopted a human detector to extract individual in videos. However, the ambiguities in feature-to-person assignments during close physical contact remains a problem. Ryoo and Aggarwal [18] utilized body part tracker to extract each individual in videos and then applied context-free grammar to describe spatial and temporal relationships between people. To avoid the extraction of individual people, approaches in [19, 24, 12] treat interacting people as a group and recognize their interactions based on group motion patterns.

Human-object and object-object interaction have also been investigated in recent work. Gupta *et al.*[8] incorporated rich context derived from object class, object reaction, and manipulation motion into Bayesian models for recognizing human-object interaction from videos and static images. Mutual context of objects and human poses was explored by Yao and Fei-Fei [23]. Their work showed that using mutual context, human pose estimation and object detection can greatly benefit each other. Gong and Xiang [7] proposed a dynamically multi-linked Hidden Markov Model for recognizing group actions in-volving multiple objects. Desai *et al.*[3] encoded geometric configurations of objects and human pose in contextual models for recognizing human-object interactions (e.g. tennis-serve and tennis-forehand).

3 Interaction Representation

Our approach takes advantage of 3D space-time local features to jointly recognize interaction and segment people in the interaction. Given a video, we first use a tracker to extract interacting people from each other, and also differentiate them from the background at a patch-level. In each bounding box, spatiotemporal interest points [5] and tracklet [16] are computed within each 3D patch, and described using the bag-of-word model [5, 12, 13] (Figure 2). Spatiotemporal

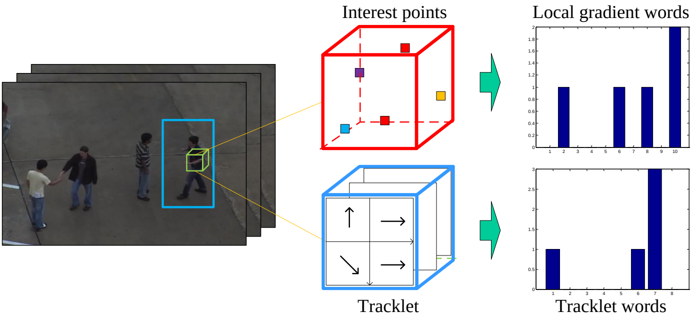


Fig. 2: Illustration of feature representation. We extract both interest points and tracklet from 3D patches.

patches are obtained by decomposing a video of size $R \times C \times T$ into a set of non-overlapping spatiotemporal 3D patches, each of which is of size $r \times c \times t$. Similar to action representation based on histograms of video words [5, 17, 13], we describe each patch by the histogram of video words within the patch.

Noted that the detected interest points and tracklet are mainly associated with salient regions in human body; few of them are associated with background. This results in an inexpressive representation for background. Our aim in this paper is to extract each interacting people from the interactions and thus the background must be described.In this paper, we augment *virtual video words* (VVWs) to describe background.

The idea of VVWs is to build a discriminative feature for background so that background and foreground can be well differentiated. Consider the features of patches as data points in a high-dimensional space. Then patch features associated with foreground are distributed subjecting to an unknown probability. We would like to define some virtual data points for background and make them as far as possible from those foreground data points in order to make these two-class data points well separated. Since we use linear kernel in the model, the best choice for virtual data points is the one that can be linearly separated from foreground data points). In our work, we use origin point for virtual data points, i.e. all the bins in the histogram of a 3D patch which have no video words in it are set to 0.

4 Patch-aware Model

Given the representation of an interaction video, our goal is to determine the interaction class (e.g. "push") as well as infer supporting regions for each interacting person. These 3D regions in this work can be associated with background or one of the interacting people.

Suppose we are given N training samples $\{\mathbf{x}^{(i)}, y^{(i)}\}_{i=1}^{N}$, where $\mathbf{x} \in \mathbb{R}^D$ denotes the video feature and $y \in \mathcal{Y}$ is the interaction class. Our purpose is to learn a discriminative function $f_{\mathbf{w}} : \mathbf{x} \to y$, which infers the interaction class for an unknown interaction video. To model the supporting regions for each interacting person, we introduce a set of auxiliary binary latent variables $\{h_j\}_{j=1}^{M} \in \mathcal{H}$ ($h_j \in \{0, 1\}$), each of which associates with one patch. $h_j = 0$ denotes that the j-th patch is associated with the background and $h_j = 1$ means it is with foreground. Note that intra-class variability leads to different patch configurations in certain interaction classes. For instance, in "handshake", some people would like to pat the other people while shaking hands with the people but some do not like that. We solve this problem by treating regions as latent variables and inferring the most probable states of latent variables in training. An undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is employed to encode the configurations of these patches. A vertex $h_j \in \mathcal{V}$ ($j = 1, \cdots, M$) corresponds to the j-th patch and an edge $(h_j, h_k) \in \mathcal{E}$ corresponds to the dependency between the two patches.

We define the discriminative function as

$$f(\mathbf{x}; \mathbf{w}) = \arg\max_{y} \left[\max_{\mathbf{h}} F(\mathbf{x}, \mathbf{h}, y; \mathbf{w}) \right] \tag{1}$$

where $\mathbf{h}$ is vector of all latent variables. The scoring function $F(\mathbf{x}, \mathbf{h}, y; \mathbf{w})$ is used to measure the compatibility between between the video data $\mathbf{x}$, the interaction class y and the latent patch labels $\mathbf{h}$.

We model the scoring function $F(\cdot)$ as a linear function $F(\mathbf{x}, \mathbf{h}, y; \mathbf{w}) = \langle \mathbf{w}, \Phi(\mathbf{x}, \mathbf{h}, y) \rangle$ with $\mathbf{w}$ being model parameter and $\Phi(\mathbf{x}, \mathbf{h}, y)$ being a feature vector. Specifically, the scoring function $F(\cdot)$ is defined as the summation of four components:

$$\begin{aligned} F(\mathbf{x}, \mathbf{h}, y; \mathbf{w}) = & \sum_{j \in \mathcal{V}} \alpha^{\mathrm{T}} \psi(\mathbf{x}_j, h_j, y) + \sum_{j \in \mathcal{V}} \beta^{\mathrm{T}} \theta(\mathbf{x}_j, h_j) \\ & + \sum_{j \in \mathcal{V}} \gamma_j^{\mathrm{T}} \eta(h_j, y) + \lambda^{\mathrm{T}} \pi(\mathbf{x}, y), \end{aligned} \tag{2}$$

where $\mathbf{w} = \{\alpha, \beta, \gamma, \lambda\}$ is model parameter, $\mathbf{x}_j$ is the feature extracted from the j-th patch.

Class-specific Patch Model $\alpha^{\mathrm{T}} \psi(\mathbf{x}_j, h_j, y)$ models the agreement between the observed patch feature $\mathbf{x}_j$, the patch label h_j and the interaction class y. The definition of the feature vector $\psi(\mathbf{x}_j, h_j, y)$ is given by

$$\psi(\mathbf{x}_j, h_j, y) = \mathbf{1}(y = a) \cdot \mathbf{1}(h_j = 1) \cdot f(\mathbf{x}_j), \tag{3}$$

where $f(\mathbf{x}_j)$ denotes the local feature of the j-th patch and $\mathbf{1}(\cdot)$ is an indicator function. In our work, $f(\mathbf{x}_j)$ encodes both appearance information and structural information of the j-th patch: $f(\mathbf{x}_j) = [f_a(\mathbf{x}_j), f_s(\mathbf{x}_j)]$. The appearance information $f_a(\mathbf{x}_j)$ is the distribution of words in the patch, and the structural information $f_s(\mathbf{x}_j)$ is the location of the patch. To compute the structural feature $f_s(\mathbf{x}_j)$, we discretize the bounding box into M patches and the spatial location feature of a patch $\mathbf{x}_j$ is a vector of all zeros with a single 1 for the bin occupied by $\mathbf{x}_j$. We apply a template α of size $(D + M) \times H \times Y$ on the feature function $\psi(\mathbf{x}_j, h_j, y)$ to weigh the different importance of elements in the feature function, where Y is the number of interaction classes, and H is the number of patches labels. Each entry in α_{yhcm} can be interpreted as, for patch of state h, how much the model prefers to see a discriminative word in the m-th bin when the codeword is c and the interaction label is y. The class-specific patch model $\alpha^{\mathrm{T}}\psi(\mathbf{x}_j, h_j, y)$ can be regarded as a linear classifier and scores the feature vector $\psi(\mathbf{x}_j, h_j, y)$.

The model encodes class-specific discriminative patch information which is of great importance in recognition. Note that the patch label h is unobserved during training and the feature function defined above models the implicit relationship between an interaction class and supporting regions. During training, the model automatically "aware" the supporting regions for an interaction class by maximizing the score $F(\mathbf{x}, \mathbf{h}, y; \mathbf{w})$.

Global Patch Model $\beta^{\mathrm{T}}\theta(\mathbf{x}_j, h_j)$ measures the compatibility between the observed patch feature $\mathbf{x}_j$ and the patch label h_j. We define the feature function $\theta(\mathbf{x}_j, h_j)$ as

$$\theta(\mathbf{x}_j, h_j) = \mathbf{1}(h_j = b) \cdot f(\mathbf{x}_j), \tag{4}$$

where $f(\mathbf{x}_j)$ is the local feature of the j-th patch used in the class-specific patch model. This model encodes shared patch information across interaction classes. It is a standard linear classifier trained to infer the label (0 or 1) of the j-th patch given patch feature $\mathbf{x}_j$. The parameter β is a template, which can be considered as the parameter of a binary linear SVM trained with data $\{\mathbf{x}_j, h_j\}_{j=1}^{M}$.

Essentially, the global patch model encodes the shared patch information across interaction classes. For example, since we use a tracker to obtain a bounding box of an interacting person, this person tends to appear in the middle of the box and thus the patches in the middle of the box are likely to be labeled as foreground. This information is shared across all interaction classes and can be elegantly encoded by our global patch model.

Class-specific Structure Model $\gamma_j^{\mathrm{T}}\eta(h_j, y)$ encodes the structural information of patches in one interaction class. Intuitively, human poses are different in various interaction classes. Although this information are unobserved in training samples, we treat them as latent variables so that they can be automatically discovered during model training. The class-specific structure model is given by

$$\eta(h_j, y) = \mathbf{1}(h_i = b) \cdot \mathbf{1}(y = a). \tag{5}$$

Clearly, the label of a patch is related to its location. Therefore, we use a set of untied weights $\{\gamma\}_{j=1}^{M}$ for the j-th patch, each of which is of size $H \times Y$, where M is the number of patches. The class-specific structure model expresses the prior that, without observing any feature, given an interaction class a, which state of the j-th patch is likely to be.

The class-specific structure model expresses the idea that, without observing any low-level feature, given an interaction class a, which state of the j-th patch is likely to be. The model shows its preference by scoring the feature vector $\eta(h_j, y)$ using a weight vector γ_j. Since the feature vector is a $0-1$ vector, if an entry in $\gamma_j(b,a)$ is positive, the model encourages labeling the j-th patch as b when current interaction class is a.

Global Interaction Model $\lambda^{\mathrm{T}}\pi(\mathbf{x},y)$ is used to differentiate different interaction classes. We define this feature vector as

$$\pi(\mathbf{x_0},y) = \mathbf{1}(y=a)\cdot\mathbf{x_0}, \tag{6}$$

where $\mathbf{x}_0 \in \mathbb{R}^d$ is a feature vector extracted from the whole video. Here we use a statistical bag-of-words style representation for the whole video. This potential function represents a standard linear model for interaction recognition without considering other components. If we ignore other potential functions in Eq.(2) and only consider the global interaction potential function, the parameter λ can be obtained by training a standard multi-class linear SVM.

Discussion. The proposed patch-aware model is specifically designed for interaction recognition with close physical contact. Compared with exiting interaction recognition methods [22, 17, 11, 2, 18, 24, 19, 1, 14], our model accounts for motion at a fine-grain patch level using the three components, the class-specific patch component, the global patch component, and the class-specific structure component. These three components model the appearance and structural information of local 3D patches and allow us to accurately separate interacting people at patch-level. To our best knowledge, our work is the first one that provides supporting patches for close interaction recognition, which can be used to separate interacting people.

5 Model Learning and Testing

Learning. The latent SVM formulation is employed to train our model given the training examples $\mathcal{D} = \{\mathbf{x}^{(n)}, y^{(n)}\}_{n=1}^N$:

$$\min_{\mathbf{w},\xi} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_n(\xi_n+\sigma_n) \tag{7}$$

$$\text{s.t.} \max_{\mathbf{h}} \mathbf{w}^{\mathrm{T}}\Phi(\mathbf{x}^{(n)}, \mathbf{h}_{y^{(n)}}, y^{(n)}) - \max_{\mathbf{h}} \mathbf{w}^{\mathrm{T}}\Phi(\mathbf{x}^{(n)}, \mathbf{h}, y) \tag{8}$$

$$\geqslant \varDelta(y, y^{(n)}) - \xi_n, \forall n, \forall y,$$

$$\pi(\mathbf{h}_{y^{(n)}}, y^{(n)}, \mathbf{h}_y, y) \leqslant \sigma_n, \forall n, \forall y, \tag{9}$$

where $\mathbf{w}$ denotes model parameter, ξ and σ are slack variable that allow for soft margin, and C is the soft-margin parameter. $\varDelta(y, y^{(n)})$ represents the 0-1 loss function. $\pi(\mathbf{h}_{y^{(n)}}, y^{(n)}, \mathbf{h}_y, y)$ in Constraint (9) enforces the similarity over latent regions for training videos. Our assumption is that, for videos in the same category, they are likely to have the same latent variable values. We define $\pi(\mathbf{h}_{y^{(n)}}, y^{(n)}, \mathbf{h}_y, y)$ as

$$\pi(\mathbf{h}_{y^{(n)}}, y^{(n)}, \mathbf{h}_y, y) = \frac{1}{M}d(\mathbf{h}_{y^{(n)}}, \mathbf{h}_y)\cdot\mathbf{1}(y=y^{(n)}), \tag{10}$$

where $d(\cdot,\cdot)$ computes the Hamming distance between the two vectors. The optimization problem (7-9) can be solved using the latent SVM framework [6].

Computing subgradient. The above optimization problem can be efficiently solved by the non-convex cutting plane algorithm [4]. In a nutshell, the algorithm iteratively builds an increasingly accurate piecewise quadratic approximation to the objective function. During each iteration, a new cutting plane is computed by the subgradient of the objective function and added to the quadratic approximation. The key two steps of the algorithm are to compute the empirical loss $R(\mathbf{w}) = \sum_n(\xi_n + \sigma_n)$ and the subgradient $\frac{\partial R}{\partial \mathbf{w}}$.

The computation of a subgradient is relatively straight-forward, assuming the inference over $\mathbf{h}$ can be done. Denote the empirical loss $R(\mathbf{w})$ as $R(\mathbf{w}) = \sum_n R^n(\mathbf{w})$, then the subgradient can be computed by

$$\frac{\partial R}{\partial \mathbf{w}} = \Phi(\mathbf{x}^{(n)}, \mathbf{h}^*, y^*) - \Phi(\mathbf{x}^{(n)}, \mathbf{h}', y^{(n)}), \tag{11}$$

where $(\mathbf{h}^*, y^*)$ and $\mathbf{h}'$ are computed by

$$(\mathbf{h}^*, y^*) = \arg\max_{y,\mathbf{h}} \mathbf{w}^{\mathrm{T}}\Phi(\mathbf{x}^{(n)}, \mathbf{h}, y) + \varDelta(y^{(n)}, y), \tag{12}$$

$$\mathbf{h}' = \arg\max_{\mathbf{h}} \mathbf{w}^{\mathrm{T}}\Phi(\mathbf{x}^{(n)}, \mathbf{h}, y^{(n)}) - \pi(\mathbf{h}_{y^{(n)}}, y^{(n)}, \mathbf{h}, y). \tag{13}$$

Testing. Given an unknown interaction video, we assume that the interaction region in the video is known. Our aim is to infer the optimal interaction label y^* and the optimal configurations of 3D patches $\mathbf{h}^*$:

$$\max_y \max_{\mathbf{h}} \mathbf{w}^{\mathrm{T}}\Phi(\mathbf{x}, \mathbf{h}, y). \tag{14}$$

To solve the above optimization problem, we enumerate all possible interaction classes $y \in \{\mathcal{Y}\}$ and solve the following optimization problem:

$$\mathbf{h}^*_y = \arg\max_{\mathbf{h}} \mathbf{w}^{\mathrm{T}}\Phi(\mathbf{x}, \mathbf{h}, y), \forall y \in \mathcal{Y}. \tag{15}$$

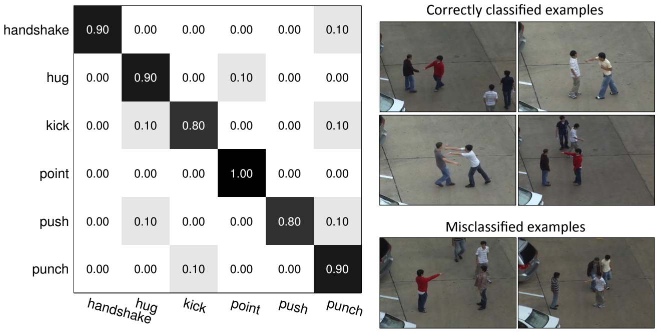
Here, the latent variables $\mathbf{h}$ are connected by a lattice. In this work, we adopt loopy belief propagation to solve the above optimization problem.

Given the latent variable vector $\mathbf{h}^*_y$, we then compute the score $f_{\mathbf{w}}(\mathbf{x}, \mathbf{h}^*_y, y) = \mathbf{w}^{\mathrm{T}}\Phi(\mathbf{x}, \mathbf{h}^*_y, y)$ for all interaction classes $y \in \mathcal{Y}$ and pick up the optimal interaction class y^* which maximizes the score $F(\mathbf{x}, \mathbf{h}^*_y, y; \mathbf{w})$.

6 Experiments

6.1 Datasets

We test our method on the UT-Interaction dataset [20] and the BIT-Interaction dataset [9]. UT dataset consists of 6 classes of human interactions: handshake, hug, kick, point, punch and push. The UT dataset was used for the human activity recognition contest (SDHA 2010) [20], and it has been tested by several state-of-the-art methods [24, 17, 19]. BIT dataset consists of 8 classes of human interactions: bow, boxing, handshake, high-five, hug, kick, pat, and push. Each class contains 50 videos, to provide a total of 400 videos.



(a) Confusion matrix. (b) Predicted examples.

Fig. 3: Confusion matrix and classification examples of our method on UT dataset.

6.2 Experiment Settings

We extract 300 interest points [5] from a video on both datasets. Gradient descriptors are utilized to characterize the motion around interest points. Principal component analysis algorithm is applied to reduce the dimensionality of descriptors to 100 and build a visual word vocabulary of size 1000. We use a visual tracker to obtain a bounding box for each interacting people. Then a 3D volume computed by stacking bounding boxes along temporal axis is split into non-overlapping spatiotemporal cuboids of size 15 × 15 × 15. We use the histogram of the video words in a 3D patch as the patch feature.

We adopt the leave-one-out training strategy on the UT dataset. The split training strategy is applied on BIT dataset to train our model. 272 videos are randomly chosen for training our patch-aware model and the remaining videos are used for testing.

6.3 Experimental Results

Results on UT-Interaction dataset. On UT dataset, we first evaluate the recognition accuracy of our method and report supporting region results. Then we compare with state-of-the-art methods [24, 17, 11, 13, 10].

Recognition Accuracy. We test our method on UT dataset and show the confusion matrix in Figure 3. Our method achieves 88.33% recognition accuracy. Confusions are mainly due to visually similar movements in two classes (e.g. "push" and "punch") and the influence of moving objects in the background. Classification examples are illustrated in Figure 3.

Eq.(5) defines a class-specific structure model for all classes. It would be interesting to investigate the performance of a shared pose prior. We replace the class-specific structure prior in Eq.(5) with a shared one which is defined as $\eta(h_j, y) = 1(h_i = b)$. Results are shown in Table 1. The accuracy difference

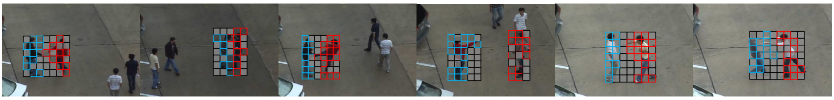

(a) handshake (b) hug (c) kick (d) point (e) punch (f) push

Fig. 4: The learned supporting regions on the UT dataset.

between the two priors is 5%. This is mainly due to that motion variations in individual actions are significant. The model with class-specific prior is able to learn pose under different classes, and benefits the recognition task.

Table 1: Accuracies of different pose prior on UT dataset.

Pose prior	shared	class-specific
Accuracy	83.33%	88.33%

Supporting Regions. The learned supporting regions on the UT dataset are shown in Fig. 4. Our model can accurately discover supporting regions of interacting people. This is achieved by finding the most discriminative regions (e.g. hand and leg) that support an interaction class. Note that some videos in the UT dataset have background motion, e.g.,"point", which introduces noise in the video. However, our model uses the structure prior component in Eq. (5) and the consistency Constraint (9) to enforce a strong structure prior information on the patches, and thus can determine which patches are unlikely to be associated with foreground. This leads to accurate patch labeling results. Some of the patch labels are incorrect mainly due to intra-class variations. People in an interaction class may behave differently according to their personal habits. This increases the difficulty of learning class-specific pose prior.

Comparison Results. We evaluate the value of components in the proposed model, including the global interaction model, the structure prior model, and the patch models. We remove these from our patch-aware model respectively, and obtain three different methods: the no-GI method that removes global interaction potential $\lambda^{\mathrm{T}}\pi(\mathbf{x}, y)$, the no-SP method that removes the structure prior potential $\gamma_j^{\mathrm{T}}\eta(h_j, y)$, and the no-CGP method which removes both class-specific and global patch model $\alpha^{\mathrm{T}}\psi(\mathbf{x}_j, h_j, y)$ and $\beta^{\mathrm{T}}\theta(\mathbf{x}_j, h_j)$ from the full model.

We compare our full model with previous methods $[17, 24, 11, 13]$, the no-GI method, no-SP method and no-CGP method, and adopt a bag-of-words representation with a linear SVM classifier as the baseline. Results in Table 2 show that our method outperforms all the comparison methods. It should be noted that our method learns supporting regions, which can be used to separate people while the methods in $[17, 24, 11, 13]$ cannot achieve this goal.

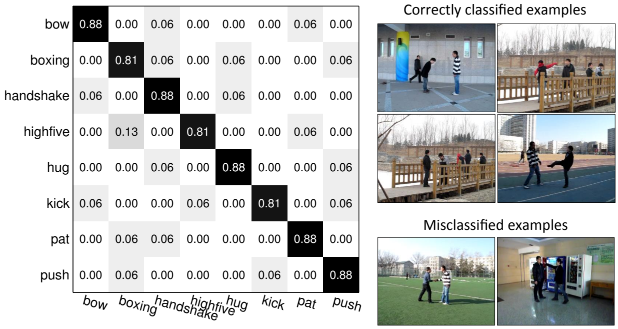
Table 2: Recognition accuracy (%) of methods on the UT dataset.

Methods	Function	handshake	hug	kick	point	punch	push	Overall
bag-of-words	only Rec.	70	70	80	90	70	70	75
no-GI method	Rec. and Seg.	20	30	40	30	10	20	25
no-SP method	Rec. and Seg.	70	80	70	70	80	80	75
no-CGP method	Rec. and Seg.	80	90	70	90	80	80	81.67
Liu *et al.*[13]	only Rec.	60	70	100	80	60	70	73.33
Lan *et al.*[11]	only Rec.	70	80	80	80	90	70	78.33
Yu *et al.*[24]	only Rec.	100	65	75	100	85	75	83.33
Ryoo & Aggarwal [17]	only Rec.	80	90	90	80	90	80	85
Our method	Rec. and Seg.	90	90	80	100	80	90	**88.33**

Results in Table 2 show that our method outperforms [24, 17, 11, 13]. The baseline bag-of-words method simply uses low-level features for recognition. By comparison, our method treats cuboid variables as mid-level features and utilize them to describe local motion information. With rich representation of interaction, our method achieves superior performance. Our method outperforms the method proposed in [17]. Their method uses structural information between interest points to aid recognition. In this work, we adopt a different scheme to encode structure information of interest points. The information is encoded by the location of spatiotemporal cuboids which contains the interest points. Besides, the learned supporting regions in our model can also be used to separate people in interactions while their method cannot. Lan *et al.*[11] utilized action context to recognize interactions. We argue that action context may not able to capture complex action co-occurrence since individual motion could be totally different in an interaction class. Thus modeling the action context may not capture significant motion variations in individual actions. We infer an interaction based on the mid-level patch features. The mid-level features we build can provide detailed regional motion information of interactions and thus improve recognition results. Compared with [24], our method learns supporting regions to separate people while [24] treats interacting people as a group and do not consider separation.

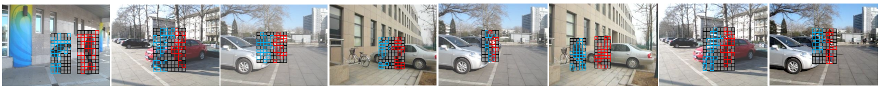
Evaluation on BIT-Interaction Dataset. We conduct two groups of experiments on BIT dataset. First, we test the recognition performance of our method, and show the results on supporting regions and the structure prior. We then test the effectiveness of each component in our patch-aware model.

Recognition Results. In the first experiment, we test the proposed method on BIT dataset. The confusion matrix is shown in Figure 5(a). Our method achieves 85.38% accuracy in classifying human interactions. Results show that the method can recognize interactions in some challenging situations, e.g. partially occlusion and background clutter (Figure 5(b)). This is mainly due to the modeling of the supporting regions. In such challenging scenarios, the supporting



(a) (b)

Fig. 5: (a) Confusion matrix of our method and (b) classification examples on the BIT dataset. Our method achieves 85.38% accuracy on the BIT dataset.



(a) bow (b) boxing (c) handshake (d) high-five (e) hug (f) kick (g) pat (h) push

Fig. 6: The learned supporting regions on the BIT dataset.

regions of each interacting people can be accurately inferred by the patch-aware model according to their appearance and structural information. If the region belongs to the interacting people, the model would assign high weights to the feature extracted from that region and thus more trust the region. If not, the feature extracted from the region will receive low weight and thus play trivial role in interaction recognition.

Most of the misclassifications are due to the visual similarity, e.g. “boxing” and “push”, “pat” and “boxing”. In addition, some temporal segments in the classes of “boxing” and “push” are shared with other classes. For example, in “boxing”, some early segments are visually similar to some “hug” segments. Both of them are “stretching out hand”. Since we adopt voting strategy for classification, these misclassified segments would result in the misclassification of the video. Moreover, some of misclassifications are due to significant occlusion in which the extracted interest points are no discriminative enough for differentiating interactions.

Supporting Regions. Results in Figure 6 show that our model can accurately find supporting regions for interacting people in close interactions. For example, in “hug” interaction (Figure 6(e)), the supporting regions of two close people can be accurately labeled. Our model essentially conducts a refinement in the bounding box. It utilizes both appearance and structure information of patches, and learns latent pose prior for each interaction class. The optimal patch label configuration (supporting regions) that maximizes the score of an interaction class are automatically discovered in the learning procedure. The learned supporting regions overcome the problem of ambiguity in feature assignments and thus facilitate the recognition task. Some of the labels are incorrect. This is mainly due to intra-class variations. People in an interaction class may behave differently according their personal habits. This increases the difficulty of learning class-specific pose priors. We do not fully use temporal information in our model since the inference on a loopy graph is inefficient.

Structure Prior. We encode the shared structure prior potential function into our patch-aware model (refer to as SS model) and compare it with the proposed model defined in Eq.(2) (called full model). Results in Table 3 indicate that the full model outperforms the SS model. The reason can be explained from the view of parameters. For j-th patch, a shared prior for the patch is associated with parameter γ_j where γ_j is a vector of length H. This shared model is too simple to capture pose variations among all the classes. By comparison, a class-specific prior for the cuboid is associated with parameter γ_j where γ_j is a vector of length $Y \times H$. With a more complex structure prior, the full model can easily capture large pose variations and separate background and foreground for each interaction class. Thus the recognition performance is improved.

Table 3: Accuracies of different pose prior on BIT dataset.

Pose prior	shared	class-specific
Accuracy	80.47%	85.38%

Comparison Results. In this experiment, we evaluate the value of components in the proposed method, including the global interaction potential, the structure prior potential, and the potential encoding appearance and structure information of observations. We remove these from our patch-aware model respectively, and obtain three different methods: the no-GI method that removes global interaction potential $\lambda^{\mathrm{T}}\pi(\mathbf{x}, y)$, the no-SP method that removes the structure prior potential $\gamma_j^{\mathrm{T}}\eta(h_j, y)$, and the no-CGP method which removes the appearance and structure information of observations $\alpha^{\mathrm{T}}\psi(\mathbf{x}_j, h_j, y)$ and $\beta^{\mathrm{T}}\theta(\mathbf{x}_j, h_j)$ from the full model. Our patch-aware model is compared with these three methods as well as the baseline bag-of-words representation with a linear SVM classifier.

Table 4 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 since our model is able to automatically infer the supporting regions and treat them as mid-level features. As expected, our method significantly outperforms the no-GI method, which emphasizes the importance of global interaction potential in interaction recognition. The global interaction

Table 4: Recognition accuracy (%) on the BIT dataset. R. and S. are short for recognition and segmentation, respectively.

Methods	Func.	bow	boxing	handshake	high-five	hug	kick	pat	push	Overall
bag-of-words	only R.	81.25	75	50	75	81.25	68.75	62.5	68.75	70.31
no-GI model	R. & S.	20.31	20.31	25	18.75	37.5	18.75	31.25	18.75	23.83
no-SP model	R. & S.	75	68.75	68.75	75	68.75	87.5	81.25	68.75	74.22
no-CGP model	R. & S.	62.5	56.25	62.5	87.5	81.25	87.5	87.5	68.75	75
Lan *et al.*[11]	only R.	81.25	75	81.25	87.5	87.5	81.25	81.25	81.25	82.03
Ours	R. & S.	87.5	81.25	87.5	81.25	87.5	81.25	87.5	87.5	**85.38**

potential function can be considered as a standard linear model for interaction recognition without considering other components. Without this potential, the model mainly focuses on the cuboid features which would be not discriminative enough. The results of the proposed method are higher than the no-SP method, which indicates the effectiveness of the structure prior in recognition. Without the structure prior, the no-SP method is unable to capture mid-level features in cuboids. The information the no-SP model can capture is simply the noisy low-level features rather than meaningful regional information. Since the segmentation and recognition tasks are smoothly connected in our work, the lack of semantic understanding of cuboids would influence the recognition results. As a result, the recognition accuracy of the no-SP method is decreased. The full model outperforms the no-CGP method. The appearance and structure information in the full model serves as local features and complements the global interaction information. The local features are able to describe local motion of interaction and provide detailed information. With appearance and structure information, our method can recognize more challenging interaction videos and thus achieves higher results.

7 Conclusion

We have proposed a novel model for jointly recognizing human interaction and segmenting people in the interaction. Our model is built upon the latent structural support vector machine in which the patches are treated as latent variables. The consistency of latent variables are encouraged across all the training data. The learned patch labels indicate the supporting regions for interacting people, and thus solve the problems of feature assignment and occlusion. Experiments show that our method achieves promising recognition results and can segment people at patch level during an interaction, even in a close interaction.

Acknowledgements. This research is supported in part by the NSF CNS award 1314484, ONR award N00014-12-1-1028, ONR Young Investigator Award N00014-14-1-0484, and U.S. Army Research Office Young Investigator Award W911NF-14-1-0218.

References

1. Choi, W., Savarese, S.: A unified framework for multi-target tracking and collective activity recognition. In: ECCV. pp. 215–230. Springer (2012)
2. Choi, W., Shahid, K., Savarese, S.: Learning context for collective activity recognition. In: CVPR (2011)
3. Desai, C., Ramanan, D., Fowlkes, C.: Discriminative models for static human-object interactions. In: CVPR Workshop on Structued Models in Computer Vision (2010)
4. Do, T.M.T., Artieres, T.: Large margin training for hidden markov models with partially observed states. In: ICML (2009)
5. Dollar, P., Rabaud, V., Cottrell, G., Belongie, S.: Behavior recognition via sparse spatio-temporal features. In: VS-PETS (2005)
6. Felzenszwalb, P., McAllester, D., Ramanan, D.: A discriminatively trained, multi-scale, deformable part model. In: CVPR (2008)
7. Gong, S., Xiang, T.: Recognition of group activities using dynamic probabilistic networks. In: ICCV. vol. 2, pp. 742 –749 (2003)
8. Gupta, A., Kembhavi, A., Davis, L.: Observing human-object interactions: Using spatial and functional compatibility for recognition. PAMI 31(10), 1775–1789 (2009)
9. Kong, Y., Jia, Y., Fu, Y.: Learning human interaction by interactive phrases. In: ECCV (2012)
10. Kong, Y., Jia, Y., Fu, Y.: Interactive phrases: Semantic descriptions for human interaction recognition. In: PAMI (2014)
11. Lan, T., Wang, Y., Yang, W., Robinovitch, S.N., Mori, G.: Discriminative latent models for recognizing contextual group activities. PAMI 34(8), 1549–1562 (2012)
12. Laptev, I., Marszałek, M., Schmid, C., Rozenfeld, B.: Learning realistic human actions from movies. In: CVPR (2008)
13. Liu, J., Kuipers, B., Savarese, S.: Recognizing human actions by attributes. In: CVPR (2011)
14. Odashima, S., Shimosaka, M., Kaneko, T., Fuikui, R., Sato, T.: Collective activity localization with contextual spatial pyramid. In: ECCV (2012)
15. Patron-Perez, A., Marszalek, M., Reid, I., Zissermann, A.: Structured learning of human interaction in tv shows. PAMI 34(12), 2441–2453 (2012)
16. Raptis, M., Soatto, S.: Tracklet descriptors for action modeling and video analysis. In: ECCV (2010)
17. Ryoo, M.S.: Human activity prediction: Early recognition of ongoing activities from streaming videos. In: ICCV (2011)
18. Ryoo, M., Aggarwal, J.: Recognition of composite human activities through context-free grammar based representation. In: CVPR. vol. 2, pp. 1709–1718 (2006)
19. Ryoo, M., Aggarwal, J.: Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities. In: ICCV. pp. 1593–1600 (2009)
20. Ryoo, M., Aggarwal, J.: UT-Interaction Dataset, ICPR contest on Semantic Description of Human Activities (SDHA) (2010)
21. Shi, Q., Cheng, L., Wang, L., Smola, A.: Human action segmentation and recognition using discriminative semi-markov models. IJCV 93, 22–32 (2011)
22. 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)

23. Yao, B., Fei-Fei, L.: Modeling mutual context of object and human pose in human-object interaction activities. In: CVPR. pp. 17 –24 (2010)
24. Yu, T.H., Kim, T.K., Cipolla, R.: Real-time action recognition by spatiotemporal semantic and structural forests. In: BMVC (2010)