J. Vis. Commun. Image R. 32 (2015) 63–73

Contents lists available at ScienceDirect

J. Vis. Commun. Image R.

journal homepage: www.elsevier.com/locate/jvci

Two-person interaction recognition via spatial multiple instance embedding ☆

Fadime Sener [a], Nazli Ikizler-Cinbis [b,*]

[a] Department of Computer Engineering, Bilkent University, 06800 Ankara, Turkey
[b] Department of Computer Engineering, Hacettepe University, 06800 Ankara, Turkey

ARTICLE INFO

Article history:
Received 6 April 2015
Accepted 30 July 2015
Available online 6 August 2015

Keywords:
Human interaction recognition
Activity recognition
Multiple instance learning
Video retrieval
Video analysis
Human actions
Human interactions
Spatial embedding

ABSTRACT

In this work, we look into the problem of recognizing two-person interactions in videos. Our method integrates multiple visual features in a weakly supervised manner by utilizing an embedding-based multiple instance learning framework. In our proposed method, first, several visual features that capture the shape and motion of the interacting people are extracted from each detected person region in a video. Then, *two-person* visual descriptors are formed. Since the relative spatial locations of interacting people are likely to complement the visual descriptors, we propose to use *spatial multiple instance embedding*, which implicitly incorporates the distances between people into the multiple instance learning process. Experimental results on two benchmark datasets validate that using two-person visual descriptors together with spatial multiple instance learning offers an effective way for inferring the type of the interaction.

1. Introduction

Human activity and interaction recognition remain as an open challenge for computer vision research. Recent years have witnessed quite a number of studies and progression made in this area, especially for the problem of human action/activity recognition. Recent reviews on this topic include [35,50,1]. However, there is still a large room for improvement, especially for the recognition of activities and interactions in unconstrained videos.

In this work, we look into the problem of recognizing interactions that take place between two people. We believe that a model developed for two-person interaction recognition can serve as a primitive for more complex recognition systems that involve multiple people and/or collective interactions. It also has the potential to be deployed in complex systems ranging from surveillance applications to human–computer interfaces for content-based video retrieval.

There are subtle differences between human interactions and singleton activities. For the recognition of singleton activities, the focus is on the body parts of a single person and the related spatio-temporal patterns in general. On the contrary, human–human interactions involve detailed analysis of two people: the proximity, respective positions and poses of interacting people all matter in distinguishing the underlying interaction patterns. This paper looks into this area, and investigates the use of a number of cues to capture the characteristics of two-person interactions.

In this work, we cast the problem of human–human interaction recognition in a weakly supervised setting. The main reason for this choice of formulation is that designing a fully-supervised system is a very cumbersome task which requires annotating every frame of interaction on a large number of videos. We assume that for each video sequence, the only available supervision is the interaction class label. We do not have the information where in the sequence the interaction takes place, *i.e.* the start and the end of the interactions are not marked. In addition, there may be multiple people in a video, where some of them are not involved in any interaction. Such presence of unrelated frames and uninvolved people add a remarkable amount of noise to the problem. Our goal is to be able to distinguish ongoing interactions in the videos in spite of such noise.

In order to deal with such presence of noise, we propose to jointly leverage visual and spatial characteristics of human interactions within a multiple-instance learning (MIL) framework. An outline of the proposed approach is illustrated in Fig. 1. In our proposed framework, first, the bounding boxes and the tracks of the people within a video are extracted using off-the-shelf person detectors and tracking methods. Then, in each frame, two-person pairs are formed by pairing each person region with another person region. We extract multiple *two-person* shape and motion descriptors from these pairs. Later on, these *two-person* descriptors

☆ This paper has been recommended for acceptance by M.T. Sun.

* Corresponding author.
E-mail addresses: fadime.sener@cs.bilkent.edu.tr (F. Sener), nazli@cs.hacettepe.edu.tr (N. Ikizler-Cinbis).

http://dx.doi.org/10.1016/j.jvcir.2015.07.016
1047-3203/

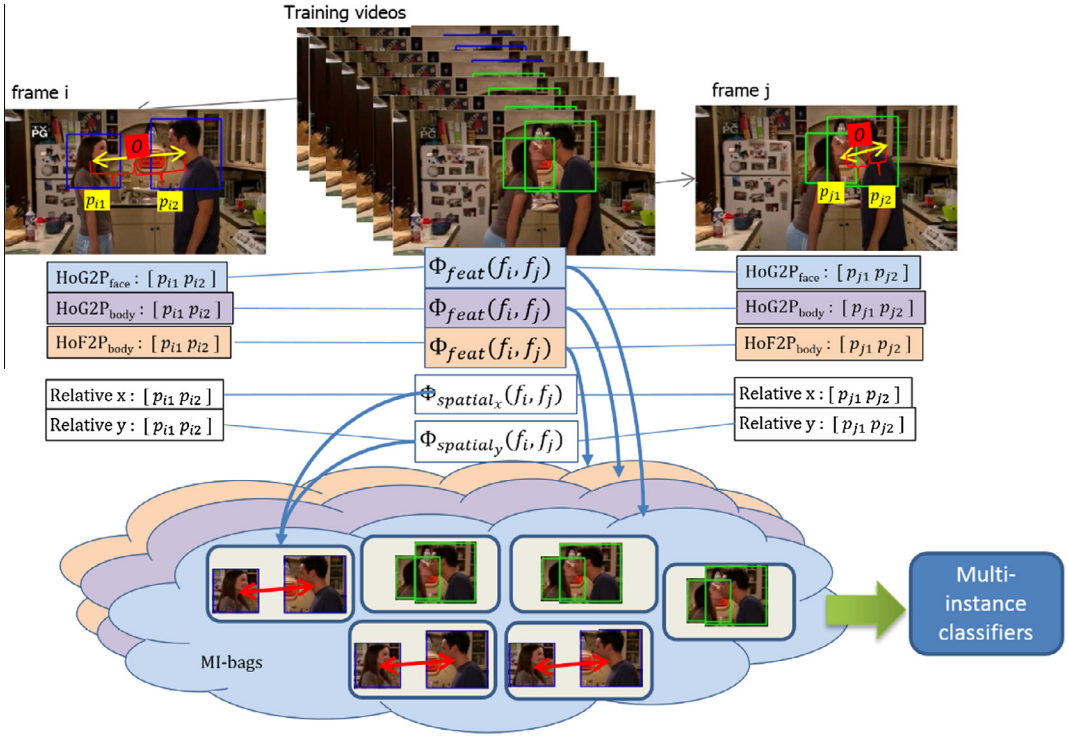


Fig. 1. Our proposed framework for human–human interaction recognition.

become the candidate instances within the MIL bags that are processed in the learning phase. We incorporate the spatial distances between people into the MIL framework by modifying Multiple-Instance Learning via Embedded Instance Selection (MILES) [7] to include two multiplicative spatial kernels. We demonstrate that using two spatial kernels is more suited to the problem, giving the flexibility of modeling variances of spatial distances.

Our contributions in this paper are twofold:

- Instead of using regular single person features, we propose to use two-person descriptors. We show that for recognizing two-person interactions, these descriptors are more effective than their singleton counterparts.
- We propose a embedding based representation that jointly incorporates the appearances and relative spatial positions of the visual interaction elements. We show that this proposed MIL embedding captures the nature of the two-person interactions more accurately.

We evaluate our algorithm on two benchmark datasets for two-person interactions: UT-Interactions [40] and TV Interactions [34] datasets. Our experimental results confirm that the proposed MIL framework obtains state-of-the-art recognition performance, and at the same time, qualitative evaluations show that it offers a simple and interpretable model.

The rest of the paper is organized as follows. In Section 2, we review the existing literature on human interaction recognition. In Section 3, first, we present the visual feature extraction step and then give the details of the proposed spatial multi-instance embedding for human interaction recognition. Section 4 includes the quantitative and qualitative experimental results, and Section 5 presents brief discussions together with potential future directions.

2. Related work

While there is a large body of literature on human action/activity recognition, such as [25,41,48,44], the problem of recognizing human interactions is a relatively less studied topic in computer vision. Related work on human interaction recognition typically addresses one of the following two interaction types: (i) human–object interactions, and (ii) human–human interactions. Prior work on human–object interaction include simultaneous object and action recognition using probabilistic models [18], extraction of distinctive feature groups [52], bag-of-features and part-based representations [11], weakly supervised learning [37]. In this work, we basically focus on the problem of recognizing human–human interactions, specifically two-person interactions.

Two-person interaction recognition: In one of the earliest studies on two-person interaction recognition, Datta et al. [10] focus on the problem of person-to-person violence recognition and uses motion trajectory information. Park and Aggarwal [32] propose to simultaneously segment and track multiple body parts of interacting humans in videos. Ryoo and Aggarwal [39] looks at the matching of local spatio-temporal features which are known to have good performance on atomic action recognition.

Initial attempts [10,32,39] heavily depend on the successes of low level processes such as background subtraction. Such low level processes are likely to fail in the complex settings of the unstructured real world video footage coming from TV shows and YouTube. In this respect, the study of Patron-Perez et al. [34,33] is different. They target at the recognition of two-person interactions such as hand shake, high five that are extracted from TV shows with cluttered backgrounds and introduce a person-centered descriptor which exploits head orientations for the recognition of two-person interactions. Patron-Perez et al. [34] claim that face orientations contain important cues for inferring the type of the action since two people face to each other when they are in interaction. Fathi et al. [13] also consider faces and their locations for recognizing social interactions in egocentric videos. Marin-Jimenez et al.[29] determine whether people are looking at each other by considering eyeline match between people. In our work, we also make use of features extracted around face regions and upper body for aiding two-person interaction recognition, and

we show that our spatial embedding based representation is more suited for this problem.

In their recent work Zhang et al. [56] propose bag-of-phrases approach for activity recognition, in which the visual phrases are constructed via the identification of co-occurring space–time points, which in turn can be used for interaction recognition. On the other hand, Kong et al. [23] proposes to use higher level features such as attributes and build interactive phrase models. Vahdat et al. [45] utilizes a graphical model of key pose sequences for interaction matching. Gaidon et al. [16] propose to use cluster-trees of tracklets. Marin-Jimenez et al. [27] proposes to use audio features, as well as visual features for interaction recognition. While audio features can be useful, in this work, we focus solely on visual features, and show that without the need for complex models, our simple framework of two-person based visual features coupled with spatial multiple instance embedding proves to be an effective way for two-person interaction recognition.

Yang et al. [51] have focused on how people interact in still images. Their method is closely related to *visual phrase* approach [42]. Their claim is that complex interactions can be modeled as a single representation and a joint model of body poses is proposed by focusing personal space in between.

Recently, Hoai and Zisserman [20] have demonstrated the effect of accurate upper body detection and the use of human-focused dense trajectories for interaction recognition. In this paper, our experiments also show that accurate upper body detection can be a helpful cue for recognizing ongoing interactions, even using with simple visual features.

Recognition of people interactions on videos paired with depth information is also been recently explored. van Gemeren et al. [17] introduce a new dataset that involve only two interactions, taken in controlled settings with Kinect assistance. In their work, standard HOG and poselet representations are used for interaction recognition, whilst utilizing the joint locations acquired from depth information in training. Yun et al. [54] also devise the underlying features based on the joint locations estimated from the depth data. On the contrary, our proposed framework is targeted at working without any depth information, in uncontrolled video settings.

Recognition of group activities has also emerged as another line of work. Lan et al. [24] focus on employing a structured SVM framework to capture the structure of group activities, while Choi and Savarese [8] propose a graphical model based framework to jointly track and recognize collective activities. Turn-taking activities has also been studied [36] by means of learning the structure of the causal graphs.

Multiple instance learning: Multiple instance learning (MIL) has been a topic of interest in machine learning community, due to its desirable properties of weak supervision. Earliest attempts at this problem propose probabilistic approaches and define a Diverse Density framework [30]. The k-NN classifier has been adapted for MIL by defining the distance between bags [49]. Later on, kernel methods have also been adapted to work with MI data such as [4,19]; a complete review on such approaches is available in [12]. More recently, algorithms that involve boosting [55], embedding the data into a different feature space [7], or treating the data in bags as graphs [57] have been proposed. A broad review on multiple instance classification can be found in the recent survey of Amores [3].

MIL paradigm is attractive for computer vision research due to the difficulties in obtaining fully supervised systems. Besides other domains such as scene, object recognition and tracking [26,5], MIL has been used in the categorization of singleton human actions in [2,22,43]. Prabhakar and Rehg [36] use multiple instance learning to infer the labels of causal sets which temporally co-occur in turn-taking interactions. Yun et al. [54] focus on interaction recognition on depth and motion capture data and propose to use high-dimensional body-pose features with MILBoost [55] algorithm.

In the context of object recognition, spatial embedding of local features has been exploited. In [21], the Euclidean distance between x–y coordinates of the SIFT points extracted around object regions has been used with a single spatial kernel. On the contrary, we encode the relative face and body positions of interacting people, rather than embedding the absolute spatial positions of local low-level features, and show with experiments that this representation is quite effective for human–human interaction recognition.

3. Proposed approach

Our proposed approach is a simple, interpretable and effective method which is formulated in a weakly supervised setting and thus is able to work in the presence of noise. In this section, we first describe our representation of visual features, which are extracted over pairs of people. Then, we give the details of our proposed spatial instance embedding MIL formulation.

3.1. Two-person features

Facial features can be important cues for recognizing human–human interactions since people typically look at each other while interacting. Similarly, body poses and relative positions of the people can carry strong cues as well. Based on these observations, we extract multiple visual features from the face and body regions of people. These multiple features are selected such that they are likely to be complementary to each other for recognition. Moreover, these features are mostly selected because they are standard, non-complex and easy to extract. We omit the calculation of more complex features (such as dense trajectories [47]) in order to demonstrate the effectiveness of the proposed framework. In the following paragraphs, we give the details of the visual features that we use in our learning framework.

Histogram of oriented gradients: Histogram of Oriented Gradients (HOG) [9] descriptor has been successfully used in person detection and action recognition tasks, see *e.g.* [9,14,25]. A building block of HOG descriptor is orientation histograms extracted in local spatial regions called *HOG cells*. The HOG descriptor of a region is obtained essentially by concatenating local groups of HOG cell descriptors into *HOG blocks* and concatenating the normalized HOG block descriptors. In our approach, we use HOG features in order to encode both the facial features and body poses. More precisely, we extract facial descriptors ($\texttt{HOG}_{\texttt{face}}$) by resizing each face region into 96×96 pixels and extract body-pose descriptors ($\texttt{HOG}_{\texttt{body}}$) by resizing person detection region into 128×128 pixels. In both cases, we use HOG cells of size 8×8 pixels and 2×2 HOG blocks. In order to obtain two-person interaction descriptors, we concatenate HOG descriptors of each person region. We refer to the resulting face and body descriptors as $\texttt{HOG2P}_{\texttt{face}}$ and $\texttt{HOG2P}_{\texttt{body}}$, respectively.

Histogram of optical flow: We expect motion features to be complementary to the shape features for interaction recognition. In order to account for motion information, we extract *Histogram of Optical Flow* (HOF) [25] features from person regions in each frame. Optical flow of each frame is extracted using a simple block matching algorithm and HOFs are formed using four major orientations located in 3×3 spatial grid over each ROI. Similar to HOGs, a two-person HOF descriptor ($\texttt{HOF2P}_{\texttt{body}}$) is obtained by concatenating individual HOF features. Note that, we extract HOF descriptors on the body regions only since typically there is no relationship between face motions and human interactions or the relationship is too subtle to exploit.

Relative distance: People interact in many ways and these interactions show a large amount of variability. While interacting, people keep a certain distance to each other based on the interaction type. In order to capture this information and include it in our framework, we encode spatial relations of people for body ($\texttt{rel}_{\texttt{body}}$) and face ($\texttt{rel}_{\texttt{face}}$) regions in each frame. First, we calculate the Euclidean distance between the individuals based on the x and y coordinates of the body (and face) regions. In order to obtain scale invariance, we normalize these distances with respect to the heights of the person (and face) regions. The relative distance features (denoted as $\texttt{rel2P}_{\texttt{face}}$ and $\texttt{rel2P}_{\texttt{body}}$) are then the concatenation of these relative distances.

Finally, a practical problem is that person detector sometimes fails to localize a second person in a frame. In these cases, we represent the missing detection by averaging the descriptors and spatial coordinates of the corresponding person over all frames. To be more specific, if only one person is detected in a frame, we always assume that it is the first person. To calculate features of the second person, if it is a training video, we take the average over all second-person features in all videos of the training set. During testing, if a person detection is missing, we only consider the video being processed and simply take the average of all person features in that particular test video.

3.2. Multiple instance learning for interaction recognition

In the traditional fully supervised learning, the learning procedure works over instances x_i and their individual corresponding labels y_i. In this setting, the label of each instance should be available in the training phase. In our problem, we do not have the explicit information about on which frames of the video the interaction occurs. Interactions can occur somewhere in the video sequence; and in some videos, there may be other irrelevant actions besides the labeled interaction. Therefore, each video is weakly labeled in the sense that interaction class is the only label provided for the whole sequence.

This case is particularly suitable for multiple instance learning (MIL). MIL operates over bags of instances, as opposed to working on single instances, where each bag B_i is composed of multiple instances x_{ij}. A bag B_i is labeled as positive, if at least one of the instances x_{ij} within the bag is known to be positive, whereas it is labeled as negative, if all the instances are known to be negative. This form of learning is referred as *weakly supervised*, since the labels for the individual instances (in our case, individual frames) are not available, and only the labels of the bags are given.

Bag construction: During training, each video is considered as a bag of instances and associated with an interaction class label. Each instance is represented by a two-person feature vector. Fig. 2 illustrates our bag formation scheme. For each person pair in a frame (as shown in Fig. 2(a)), a two-person descriptor is extracted and added as an instance to the corresponding MIL bag. We assume that at least one pair of people within a video involves the target interaction, so that the MI positivity requirement is obtained.

This MIL formulation implicitly covers the case when there are more than two people per frame. In our formulation, each pair of people is treated as an instance in the MIL bag. More precisely, taking each person as a reference, we match the reference person with each one of the remaining person regions to its right in order to extract our two-person based feature vectors. We repeat this procedure until all left-to-right ordered pairs are included in the bag. See Fig. 2(b) for an illustration.

For each positive training video, the corresponding MIL bag is formed using the aforementioned procedure. Negative bags are formed in a similar fashion, using a uniformly sampled portion from the videos of the remaining interactions, and/or videos that do not contain any interaction.

3.2.1. Spatial multiple instance embedding for interaction modeling

In order to model the spatial relationships between interacting people more efficiently, we propose to use a variant of the Multiple Instance Learning with Embedded Instance Selection [7] (MILES) algorithm. Our proposed framework includes an extension of the original MILES algorithm to include relative spatial distances in the embedding step. By infusing the relative distances within the embedding itself via multiplicative kernels, we can easily and naturally represent the spatial relationships between interacting people, and this information proves to be very useful for recognition of the interactions.

More specifically, we first embed the original feature space x to the instance domain. In this embedding, each multi-instance bag B is represented by its similarity to each concept instance c_k in the training bags. The set of concept instances is denoted by $C = \{c_k : k = 1, \ldots, N\}$. Each concept instance c_k, which can also be considered as a reference point for a target concept, corresponds to a MIL embedding dimension. Therefore, the cardinality of the set C defines the dimensionality of the embedding vectors.

The set of concept instances, C can be obtained in a number of ways. In practice, the most prominent two approaches are (i) aggregation of the complete set of instances in the dataset, or (ii) utilization of the output of an intermediate clustering step. In our case, we use all the instances extracted from the training videos as the set of concept instances for embedding.

The original formulation of MILES [7] depends only on the visual feature similarity, where the similarity $s(\cdot)$ between a bag B_i and a concept instance $c_k \in C$ is given by

$$s(c_k, B_i) = \max_j \left(\phi_{feat}(x_{ij}, c_k)\right). \tag{1}$$

Here, $\phi_{feat}(x_{ij}, c_k)$ is the similarity between feature vectors, defined as

$$\phi_{feat}(x_{ij}, c_k) = \exp\left(-\frac{D(x_{ij}, c_k)}{\sigma}\right), \tag{2}$$

where $D(\cdot)$ measures the similarity between a concept instance c_k and a bag instance x_{ij}. In our experiments, we use simple Euclidean distance as $D(\cdot)$.

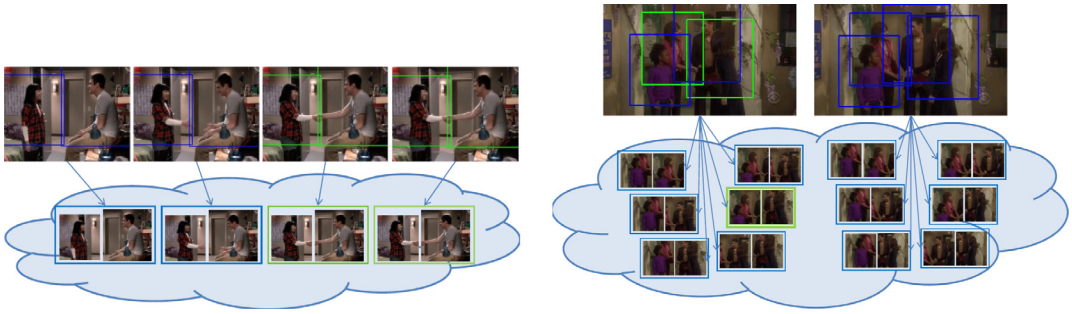
As discussed in Section 3.1, spatial relations between people can provide important additional information about human interactions. In order to incorporate such relationships into the learning framework, we modify this formulation and add two multiplicative spatial kernels. More precisely, the similarity between an instance and a bag is modified to

$$s(c_k, B_i) = \max_j \left(\phi_{feat}(x_{ij}, c_k)\phi_{sp_x}(x_{ij}, c_k)\phi_{sp_y}(x_{ij}, c_k)\right), \tag{3}$$

where $\phi_{sp_x}(x_{ij}, c_k)$ is the spatial closeness between a concept instance c_k and a bag instance x_{ij} over the x coordinate and $\phi_{sp_y}(x_{ij}, c_k)$ is the corresponding spatial closeness over the y coordinate. Replacing θ for x_{ij} and β for c_k for shorthand, spatial kernel $\phi_{sp_x}(\theta, \beta)$ is defined as follows:

$$\phi_{sp_x}(\theta, \beta) = \exp\left(-\frac{\left|d_x(p_1, o_\theta) - d_x(q_1, o_\beta)\right|\left|d_x(p_2, o_\theta) - d_x(q_2, o_\beta)\right|}{\sigma_x}\right), \tag{4}$$

where $\theta = \{p_1, p_2\}, \beta = \{q_1, q_2\}, p_1$ and p_2 are the first and second person in the bag instance x_{ij}, q_1 and q_2 are the first and second person in the concept instance c_k. o_θ represents the middle point of two people in x_{ij} and o_β represents the middle point of two people in c_k respectively. $d_x(\cdot)$ measures the distance in x dimension.



(a) two-person (b) multi-person

Fig. 2. Example multiple instance bag creation for videos with (a) two-person and (b) multi-person. This figure is best viewed in color. Color blue shows the frames with no-interaction, and green shows the presence of the interaction. In two-person case, features extracted from each person region are concatenated and added to the MIL bag as an instance. When multiple people are present in the scene, person regions are paired with eachother and each pair is accounted as a candidate MIL instance. Note that, the presence of multiple people is likely to cause many negative instances in the MIL bags. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

By replacing $d_x(\cdot)$ in Eq. (4) by $d_y(\cdot)$, which corresponds to the distance in y dimension, we get the second spatial kernel $\phi_{sp_y}(x_{ij}, c_k)$ in Eq. (3). Both $d_x(\cdot)$ and $d_y(\cdot)$ are normalized with respect to the related person bounding box size. σ_x and σ_y are the bandwidth parameters that adjust the sensitivity of the measure to the spatial differences and in the experiments, these parameters are selected using cross-validation over the training set.

Eq. (3) allows us to consider the similarity between feature vectors of two-person regions and relative distances of two interacting people in both x and y dimensions together. In this formulation, we consider x and y coordinates separately, rather than having a single distance measure. This nuance is crucial, the relative vertical and horizontal distances may contain important characteristics for each interaction. Therefore, having distinct spatial kernels is necessary to capture such distinguishing properties. Our preliminary experiments also validate this observation; having two separate kernels produce more accurate results.

In the end, each bag can be represented in terms of its similarities to each of the target concepts. The corresponding mapped representation $\mathbf{m}(B_i^f)$ becomes

$$\mathbf{m}(B_i) = [s(c_1, B_i), s(c_2, B_i), \ldots, s(c_N, B_i)]^T \quad (5)$$

and the final classification is performed over this embedded space.

3.2.2. Classification

After the embedding step, an L2-regularized SVM [6] with RBF kernel is trained over the mapped representations $\mathbf{m}(B)$ for building multiple instance classifiers. Separate classifiers are learnt for each of the two-person descriptors $\texttt{HOG2P}_{\texttt{face}}$, $\texttt{HOG2P}_{\texttt{body}}$ and $\texttt{HOF2P}_{\texttt{body}}$, respectively.

The final classification is achieved via a second linear SVM layer learned over the response vector of the individual feature classifiers. This final layer of linear SVM provides the late fusion of the different features and helps to compensate the differences in model biases.

4. Experiments

4.1. Datasets and experimental setup

In order to evaluate the performance of our method, we use two benchmark datasets available for human interaction recognition: These are UT-Interactions [40] dataset and TV Interactions [34] dataset.

UT-Interactions [40] dataset consists of 20 videos, where each video contains six different interactions between two people. These interactions are hand shaking, hugging, kicking, pointing, punching and pushing, and are performed by 10 different actors. There are two sets of videos, where Set 1 is composed of 10 video sequences taken on a parking lot and Set 2 is composed of 10 video sequence taken on a lawn in a windy day. In this dataset, the videos have relatively stable backgrounds, with a resolution of 720 * 480, at a rate of 30 fps. The height of a person is about 200 pixels. In our experiments, we use the segmented version of Set 1 and Set 2 to compare our method's recognition performance with the existing works. We follow the same testing routine of [40], which involves 10-fold leave-one-out cross-validation. As a preprocessing step, we deploy the Felzenszwalb et al.'s person detector [14] and use meanshift tracking to aid in localizing people in frames with no detection.

The second dataset is the more realistic "TV Interactions" dataset collected by Patron-Perez et al. [34]. This dataset consists of 300 videos extracted from different TV shows. The dataset contains four interactions: hand shake, high five, hug and kiss (50 videos for each class) and negative examples (100 videos) which do not contain any of the four interactions. It is a quite challenging dataset with changing camera viewpoints, varying scales, etc. The lengths of the video clips range between 30 and 600 frames. In this dataset, the upper body bounding boxes of the people and interaction labels are provided for each frame. We follow the same evaluation methodology of [34], applying cross-validation using the two splits

Table 1
Comparison between singleton features and 2P features on TV Interactions dataset. In this table, Average Precision (AP) values are reported. The classifiers are learned using the regular MILES with no spatial embedding. The combination of the individual features are done via a linear SVM. Bold values indicate the highest AP scores for each individual interaction class.

Feature	Handshake	Highfive	Hug	Kiss	Avg
$\texttt{HOG}_{\texttt{body}}$	53.69	43.74	50.67	56.15	51.06
$\texttt{HOG}_{\texttt{face}}$	52.30	57.13	63.05	58.93	57.85
$\texttt{HOF}_{\texttt{body}}$	44.28	44.35	35.48	36.75	40.21
$\texttt{rel}_{\texttt{body}}$	52.40	49.45	48.47	36.81	46.78
$\texttt{rel}_{\texttt{face}}$	52.70	51.42	49.69	46.79	50.15
All	56.57	55.00	64.11	53.29	57.24
$\texttt{HOG2P}_{\texttt{body}}$	**63.08**	52.06	62.24	68.64	64.32
$\texttt{HOG2P}_{\texttt{face}}$	58.11	63.60	**74.13**	75.97	66.15
$\texttt{HOF2P}_{\texttt{body}}$	59.90	**63.99**	49.64	49.25	59.47
$\texttt{rel2P}_{\texttt{body}}$	55.73	53.54	66.53	56.92	61.39
$\texttt{rel2P}_{\texttt{face}}$	54.40	54.01	67.73	61.37	62.50
All2P	61.77	63.97	72.73	**76.36**	**68.71**

Table 2
Average Precision (AP) on TV Interactions dataset with 2P features. Bold values indicate the highest AP scores for each individual interaction class.

Method	Feature	hs	hf	h	k	Avg
		Negatives excluded				
MIL embedding [21]	All2P	61.77	63.97	72.73	76.36	68.71
	All2P	60.50	63.33	81.83	74.93	70.15
Our method	$HOG2P_{body}$	66.90	69.99	77.50	74.48	72.22
	$HOG2P_{face}$	63.57	66.47	**87.06**	**84.18**	75.32
	$HOF2P_{body}$	67.72	69.61	69.62	55.60	65.64
Our method	All2P	**68.57**	**70.03**	83.68	80.13	**75.60**
		Negatives included				
MIL embedding [21]	All2P	47.83	51.48	83.29	77.10	64.93
	All2P	48.74	54.07	83.56	66.28	63.16
Our method	$HOG2P_{body}$	49.81	52.29	83.51	67.30	63.23
	$HOG2P_{face}$	41.87	57.90	86.92	**82.94**	67.41
	$HOF2P_{body}$	**53.61**	48.99	71.57	49.11	55.82
Our method	All2P	50.13	**61.28**	**88.69**	69.70	**67.45**

of the data. There are two evaluation schemes that [34] considered, where the first scheme excludes the negative data, and the second includes negative data in training and testing. We report the results for both of these settings.

During experiments, all the parameters are selected using cross-validation over the training set and the results are reported at the sequence level.

4.2. Performance of individual features

We first evaluate the performance of the individual features on the TV Interactions dataset [34]. For the five types of singleton features, we first employ the standard MIL embedding. The instance embedding representation of MILES considers only the low-level feature similarity, and we train an L2-regularized SVM with RBF kernel for each visual feature. For the combination of these classifiers, we use the same late fusion scheme described above.

Average Precision (AP) values obtained using individual features are shown in Table 1. HOG_{face} has the best singleton performance among others, followed by HOG_{body}. For hand shake class HOG_{body} has the best performance, and for the remaining interactions HOG_{face} feature provides the best performance. These results demonstrate that shape features are very informative for inferring the type of the interaction. The high performance of HOG_{face} is not surprising, since most of the interactions occur closer to the facial area. This also coincides with the claims of [34] on the importance of exploiting the visual features extracted around the head region. We observe that, in this dataset HOF_{face} features are not that reliable, showing promising performance only for high five and handshake actions. This may be due to the existing camera motion in this dataset.

As it can be seen from Table 1, relative distance features have also good performance. This observation suggests that the relative spatial locations of people can provide useful information and encourages to further investigate these features. The sixth row of Table 1 is the performance when these singleton feature classifiers

(a) MILES

(b) Our Method

Fig. 3. Ranking results for the TV-Interactions dataset (negatives not included). (a) The result of the MILES algorithm of using HOG_{body}, HOG_{face} and HOF_{body} features, and (b) is the result of the proposed framework with $HOG2P_{body}$, $HOG2P_{face}$ and $HOF2P_{body}$ features. Left column displays the true positives based on their ranking in the retrieval, and right column displays the false positives with their ranks in the list. Note that the green bounding boxes show the ground truth annotations for this dataset. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 3
Classification accuracies of spatial embedding on UT-Interactions dataset. Bold values indicate the highest AP scores for each individual interaction class. The overall accuracy is 93.3%.

Feature	hs	h	k	po	pun	pus	Avg
			SET1				
HOG2P$_{body}$	100	100	90	100	70	90	91.67
HOG2P$_{face}$	100	100	90	100	80	90	93.33
HOF2P$_{body}$	90	100	90	100	50	90	86.67
All2P	**100**	**100**	**90**	**100**	**80**	**100**	**95.00**
			SET2				
HOG2P$_{body}$	100	100	60	100	80	90	88.33
HOG2P$_{face}$	100	100	50	100	90	90	88.33
HOF2P$_{body}$	90	100	70	100	70	50	80.00
All2P	**100**	**100**	**70**	**100**	**90**	**90**	**91.67**

are combined via the final layer of linear SVM. Surprisingly, combination of the singleton visual features does not offer much of a performance difference in this case, and features extracted around the upper body region seem to be dominant in recognition.

A more interesting observation from Table 1 is that, recognition performance significantly improves if two-person (2P) features are used. HOG2P$_{face}$ achieves the best performance amongst 2*P* features and the combination of all 2P features provides a slight increase over HOG2P$_{face}$. Compared to singleton features, this noticeable increase in performance suggests that using 2P features can be a fruitful direction to explore for human interaction recognition.

4.3. Performance of spatial embedding

Next, we look at the performance of the proposed framework based on spatial embedding. Instead of using relative distance features as a separate feature, these distances are incorporated into the embedding procedure. Table 2 shows the results. We observe that encoding spatial information via the proposed kernels to the instance embedding procedure increases the performance for all three feature types. HOG2P$_{face}$ feature has the best performance among others, and especially for hug and kiss interactions the performance gain is noticeable. For these interactions, spatial kernels are shown to be especially useful. Overall, the best results are achieved using spatial embedding with all 2P features.

In Table 2, we also compare our approach with the spatial encoding approach of [21] that uses a single spatial kernel that computes the direct Euclidean distance between features for embedding. As it can be seen, our method outperforms this naive

Fig. 4. The contributions of individual frames to the classification using our approach. In this figure, from top to bottom, the videos belong to hand shake, high five, hug and kiss classes (two rows for each). Note that the green and blue bounding boxes show the ground truth annotations for this dataset. The contribution scores are displayed on the bottom right corner of each frame and the positively contributing instances with respect to each target interaction are marked with green. While some in-class frames are missed (e.g. frame in second row first column for high five interaction), overall, our algorithm is quite successful in discriminating the frames related to each interaction. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Fig. 5. The contributions of individual frames to the classification using our approach on UT-Interactions dataset [40]. In this figure, from top to bottom, the videos belong to hand shake, hug, kick, punch and push classes, from SET1 and SET2 divisions of the dataset. The yellow bounding boxes show the person tracks that are acquired automatically via person detection and meanshift tracking. The contribution scores are displayed on the bottom right corner of each frame and the instances with positive contribution scores for each target interaction are marked with green. As it can be observed, while there are some confusions, overall, our algorithm is quite successful in discriminating the frames related to each interaction. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

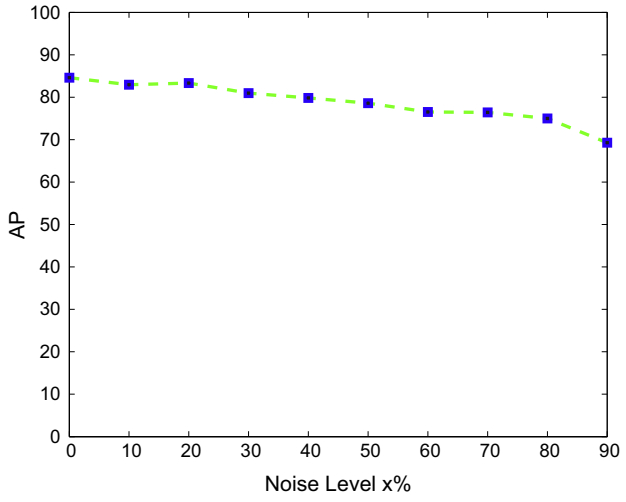


Fig. 6. The effect of the presence of noise. Having a multiple-instance based nature, the proposed framework is not much affected by the amount of noise, aka the people regions that are not involved in any type of interactions, in the video.

spatial encoding, showing that the proposed embedding is more effective for two-person interaction recognition problem.

In Fig. 3, some qualitative examples are given for the rankings obtained using different methods. Fig. 3(a) shows the ranking results for the baseline standard MIL Embedding (MILES) when using with the regular singleton $\texttt{HOG}_{\texttt{body}}$, $\texttt{HOG}_{\texttt{face}}$ and $\texttt{HOF}_{\texttt{body}}$ features. Fig. 3(b) shows results for Spatial Embedding with $\texttt{HOG2P}_{\texttt{body}}$, $\texttt{HOG2P}_{\texttt{face}}$ and $\texttt{HOF2P}_{\texttt{body}}$ features. We observe that the top three high ranking videos are relevant for all interactions regardless of the choice of method. We observe that our proposed framework tend to retrieve more relevant results higher in the list; the false positive with the highest rank for high five action has rank 9 for our method (rank 4 for MILES), and for kiss action the highest ranked false positive has rank 13 (rank 4 for MILES) respectively.

Table 4
The AP comparison of different MI-based methods using `All2P` features on TV Interactions dataset. The highest AP score for each individual interaction class is shown in bold.

Method	hs	hf	h	k	Avg
mi-Graph [57]	60.01	50.73	64.24	68.59	60.89
MILES [7]	61.77	63.97	72.73	76.36	68.71
milBoost [55]	62.32	67.84	76.72	73.97	70.21
Our method	**68.57**	**70.03**	**83.68**	**80.13**	**75.60**

Table 3 shows the performance of the proposed approach on UT-Interactions dataset. Similar to the previous findings, the best performance is achieved with the combination of all features within the spatial embedding framework. Overall, the achieved accuracy on this dataset is 93.3%.

One of the strengths of the proposed approach is its ease for interpretation. Fig. 4 illustrates this property. In this figure, example frames from the test videos of TV Interactions dataset [34] are shown, with their contribution scores overlaid. The positively contributing instances to the classification of the target interaction are framed in green. As it can be seen, our approach successfully discriminates the frames of the target interaction and offers a rough localization of the interaction within the sequence. Similarly, Fig. 5 includes example frames from UT-Interactions dataset [40] with overlaid contribution scores that are output by our method. For both of the datasets, the contribution scores usually increase as the target interaction takes place within the sequence, whereas they are usually lower for frames with no interaction.

Table 5
Comparison to the state-of-the art on "TV Interactions" dataset [34]. Average Precisions are reported for the two separate testing schemes (negatives included and excluded). In this table, l-type represents the localization type of the person regions. For each setting, the best classification performance is shown in bold.

Method	l-type	hs	hf	h	k	Avg
Negatives excluded						
Patron-Perez et al. [34]	Manual	57.83	51.08	71.16	76.54	64.15
Yu et al. [53]	Auto	–	–	–	–	66.16
Our method	Manual	68.57	70.03	83.68	80.13	**75.60**
Negatives included						
Patron-Perez et al. [34]	Auto	35.17	25.39	37.69	32.50	32.76
Patron-Perez et al. [33]	Auto	39.35	45.82	46.99	37.60	42.44
Marin-Jimenez et al. [28]	Auto	–	–	–	–	39.23
Gaidon et al. [15]	Auto	–	–	–	–	55.6
Patron-Perez et al. [33]	Manual	41.32	43.06	66.08	68.57	54.76
Patron-Perez et al. [34]	Manual	45.30	45.07	62.00	70.58	55.74
Yu et al. [53]	Auto	–	–	–	–	55.95
Hoai et al. [20]	Auto	55.8	60.2	60.8	48.2	56.3
Our method	Auto	52.74	44.77	84.33	61.43	60.81
Gaidon et al. [16]	Auto	–	–	–	–	**62.4**
Our method	Manual	50.13	61.28	88.69	69.70	**67.45**

To test our two-person descriptor's robustness to flip, we conducted a small set of experiments. We take the training set of UT-Interactions as it is, whereas we mirror flip the test set and evaluate the trained models on the flipped test set. As expected, we observe no change for the symmetric interactions, i.e. interactions done by the two people simultaneously such as handshaking. For the asymmetric actions, such as kicking and punching, we observed a slight degradation in performance where the average accuracy dropped from 93.3% to 90%. While this is a reasonable amount of performance loss, it suggests a possible bias in the dataset. When we investigate the reason behind this loss, we see that especially in SET 1 of UT-Interactions dataset the asymmetric interactions have an imbalanced distribution of examples, where the main actor of the interaction (e.g. the puncher in the punching action) tend to be on a particular side (e.g. left/right). This bias can be eliminated by mirror-flipping all the dataset. However, since this would double the size of the training data and the existing works that report results on this dataset do not use such flipping, we also omit it for the sake of fair comparison.

A limitation of our method is its dependency on the proper extraction of person regions. While the MIL setting tolerates some amount of noise and the presence of multiple people, it requires at least some representative cases of the interacting people regions to be within the bag of instances. Otherwise, the constructed bags will violate MIL positivity constraint and this is likely to lower the recognition performance. In order to evaluate the performance of the proposed method with respect to the presence of noise, we conduct the following experiment using TV-Interactions dataset: We first run our classification method on the bounding boxes that involve an interaction and we achieve a mAP of 84.61%. This case can be referred as no noise situation, where all the instances belong to one of the existing interactions. Then, we gradually add to the instance space bounding boxes of additional people that are not involved in any interaction. The change of the recognition performance (mAP) with respect to changing amounts of noise is presented in Fig. 6. As it can be seen, the effect of the noise over the recognition performance is quite minor, and the presented method is able to achieve quite competent results even in the presence of 90% noise, achieving 69.3% mAP.

4.4. Comparison to state-of-the-art

We first compare our proposed method to some of the existing Multiple Instance methods. For this purpose, two frequently used Multiple Instance Learning algorithms, [55,57], are applied over the same set of all person features `All2P` on the TV Interactions dataset. The results are shown in Table 4. While MI-graph [57] performs poorly for this problem, the MILBoost algorithm performs comparably with the MILES algorithm. Overall, the best performance is achieved with the proposed spatial embedding framework.

We then compare our method to the state-of-the-art in the literature. The results on TV Interactions dataset are given in Table 5 and results on UT-Interactions dataset are given in Table 6, respectively. For TV Interactions dataset, the reported methods in the literature either use the manually annotated bounding boxes, or they automatically detect and track the person regions within the video frames. We report the performance of our method for both of these cases (denoted by l-type in Table 5). For automatic tracks, we use the automatic track generation method of [34]. For UT-Interactions data, we generated person tracks automatically via utilizing a person detector [14] first, and then using meanshift tracking over the detections to acquire more solid person tracks.

In TV Interactions dataset (Table 5), we observe that our method is able to produce quite successful results both using manual and automatic tracks. It is on par with [16] when provided bounding boxes are not used. It should be noted that the approach of [16] relies on the powerful and computationally expensive feature extraction mechanism of dense trajectories [48], whereas we use simpler features. When such additional computational burden is not a problem, our method can as well benefit from using more advanced features. Our method achieves the state-of-the-art recognition performance when manual person annotations are used. Note that, for the manual localization case, even if the person bounding boxes are provided, there may be irrelevant people in the scene who are not involved in any interactions and this situation still introduces a significant amount of noise.

Table 6
Comparison to the state-of-the art on "UT-Interactions" dataset [40]. Best classification performance in each set of the dataset is shown in bold. Overall performance is 93.3, higher than the best results reported for this dataset so far (92.5 [56]).

Data	Method	Handshake	Hug	Kick	Point	Punch	Push	AVG
SET1	BoW [56]	70	80	90	100	50	70	77
	Waltisberg et al. [46]	50	100	100	100	70	80	83
	Mukherjee et al. [31]	85	85	95	85	75	95	86.7
	Raptis and Sigal [38]	100	100	90	100	80	90	93.3
	Vahdat et al. [45]	90	100	90	90	90	100	93
	Zhang et al. [56]	100	100	100	90	90	90	95
	Our method	100	100	90	100	80	100	**95**
SET2	BoW [56]	70	70	80	80	70	70	73
	Waltisberg et al. [46]	70	90	100	100	80	40	80
	Raptis and Sigal [38]	–	–	–	–	–	–	–
	Vahdat et al. [45]	80	100	100	100	70	90	90
	Zhang et al. [56]	80	100	100	80	90	90	90
	Our method	100	100	70	100	90	90	**91.67**

In UT-Interactions dataset, our method achieves on par or better results compared to the state-of-the-art (Table 6). In the literature, the best reported result on this dataset is 92.5% by Zhang et al. [56] and our method achieves an accuracy of 93.3%.

5. Conclusion

In this study, we propose a multiple instance learning (MIL) based approach for two-person interaction recognition in videos. Our method involves extracting multiple visual features from person regions and leveraging them in a simple form to construct two-person descriptors. Experimental results show that using two-person descriptors yields promising results. In this context, to demonstrate the effectiveness of the proposed MIL framework, we basically rely on simple features (such as HOG and HOF) and even with these simple features, our recognition rates are on par or better than the state-of-the-art on the two well-established human interaction benchmark datasets. Nevertheless, our proposed framework is easily extendible to include more sophisticated features and the recognition rates are likely to benefit from further exploration of such futures.

Another contribution is the introduction of a novel way for incorporating the spatial distances between interacting people to the multiple instance learning. We embed the spatial distances via multiplicative spatial kernels. Our results show that better recognition rates are obtainable by using spatial information in conjunction with the two-person descriptors in the proposed MIL framework.

Future work includes the exploration of different features that may further aid in recognition of everyday interactions. Early fusion techniques, together with Multiple Kernel Learning (MKL) approaches can be explored in the search for better feature combinations. Possible extensions of the proposed method can be developed to handle group interactions or collective actions as well.

Acknowledgment

This work was supported in part by the Scientific and Technological Research Council of Turkey (TUBITAK) Career Development Award numbered 112E149.

References

[1] J. Aggarwal, M. Ryoo, Human activity analysis: a review, ACM Comput. Surv. 43 (3) (2011) 16:1–16:43.
[2] S. Ali, M. Shah, Human action recognition in videos using kinematic features and multiple instance learning, IEEE Trans. Pattern Anal. Mach. Intell. 32 (2) (2010) 288–303.
[3] J. Amores, Multiple instance classification: review, taxonomy and comparative study, Artif. Intell. 201 (2013) 81–105.
[4] S. Andrews, I. Tsochantaridis, T. Hofmann, Support vector machines for multiple-instance learning, in: S. Becker, S. Thrun, K. Obermayer (Eds.), Advances in Neural Information Processing Systems, vol. 15, MIT Press, 2003, pp. 577–584.
[5] B. Babenko, M.H. Yang, S. Belongie, Visual tracking with online multiple instance learning, IEEE Trans. Pattern Anal. Mach. Intell. (PAMI) (2011).
[6] C.C. Chang, C.J. Lin, LIBSVM: a library for support vector machines, ACM Trans. Intell. Syst. Technol. 2 (2011) 27:1–27:27. Software available at <http://www.csie.ntu.edu.tw/cjlin/libsvm>.
[7] Y. Chen, J. Bi, J.Z. Wang, Miles: multiple-instance learning via embedded instance selection, IEEE TPAMI 28 (12) (2006) 1931–1947.
[8] W. Choi, S. Savarese, A unified framework for multi-target tracking and collective activity recognition, in: ECCV, 2012, pp. 215–230.
[9] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: CVPR, 2005, pp. I: 886–893.
[10] A. Datta, M. Shah, N. da Vitoria Lobo, Person-on-person violence detection in video data, in: ICPR, 2002, pp. I: 433–438.
[11] V. Delaitre, I. Laptev, J. Sivic, Recognizing human actions in still images: a study of bag-of-features and part-based representations, in: BMVC, 2010.
[12] G. Doran, S. Ray, A theoretical and empirical analysis of support vector machine methods for multiple-instance classification, Mach. Learn. 97 (1–2) (2014) 79–102.
[13] A. Fathi, J.K. Hodgins, J.M. Rehg, Social interactions: a first-person perspective, in: CVPR, IEEE, 2012, pp. 1226–1233.
[14] P.F. Felzenszwalb, R.B. Girshick, D.A. McAllester, D. Ramanan, Object detection with discriminatively trained part-based models, IEEE Trans. Pattern Anal. Mach. Intell. 32 (9) (2010) 1627–1645.
[15] A. Gaidon, Z. Harchaoui, C. Schmid, Recognizing activities with cluster-trees of tracklets, in: BMVC 2012 – British Machine Vision Conference, 2012, pp. 30.1–30.13.
[16] A. Gaidon, Z. Harchaoui, C. Schmid, Activity representation with motion hierarchies, Int. J. Comput. Vis. 107 (3) (2014) 219–238.
[17] C. van Gemeren, R.T. Tan, R. Poppe, R.C. Veltkamp, Dyadic interaction detection from pose and flow, in: Human Behavior Understanding Workshop, 2014, pp. 101–115.
[18] A. Gupta, A. Kembhavi, L. Davis, Observing human–object interactions: using spatial and functional compatibility for recognition, IEEE Trans. Pattern Anal. Mach. Intell. 31 (10) (2009) 1775–1789.
[19] T. Grtner, P.A. Flach, A. Kowalczyk, A.J. Smola, Multi-instance kernels, in: In Proc. 19th International Conf. on Machine Learning, Morgan Kaufman, 2002, pp. 179–186.
[20] M. Hoai, A. Zisserman, Talking heads: detecting humans and recognizing their interactions, in: IEEE Conference on Computer Vision and Pattern Recognition, 2014.
[21] N. Ikizler-Cinbis, S. Sclaroff, Object recognition and localization via spatial instance embedding, in: International Conference on Pattern Recognition (ICPR), IEEE, 2010, pp. 452–455.
[22] N. Ikizler-Cinbis, S. Sclaroff, Object, scene and actions: combining multiple features for human action recognition, in: ECCV, 2010, pp. 494–507.
[23] Y. Kong, Y. Jia, Y. Fu, Learning human interaction by interactive phrases, in: Proceedings of European Conference on Computer Vision, ECCV, 2012, pp. 300–313.
[24] T. Lan, L. Sigal, G. Mori, Social roles in hierarchical models for human activity recognition, in: CVPR, IEEE, 2012, pp. 1354–1361.
[25] I. Laptev, M. Marszalek, C. Schmid, B. Rozenfeld, Learning realistic human actions from movies, in: CVPR, 2008.
[26] C. Leistner, A. Saffari, H. Bischof, MILForests: multiple-instance learning with randomized trees, in: ECCV, 2010.
[27] M. Marin-Jimenez, R. Muoz-Salinas, E. Yeguas-Bolivar, N. Prez de la Blanca, Human interaction categorization by using audio-visual cues, Mach. Vis. Appl. 25 (1) (2014) 71–84.
[28] M. Marin-Jimenez, E. Yeguas, N.P. de la Blanca, Exploring stip-based models for recognizing human interactions in tv videos, Pattern Recognit. Lett. 34 (15) (2013) 1819–1828.
[29] M. Marin-Jimenez, A. Zisserman, M. Eichner, V. Ferrari, Detecting people looking at each other in videos, Int. J. Comput. Vis. 106 (3) (2013) 282–296.
[30] O. Maron, T. Lozano-Pérez, A framework for multiple-instance learning, in: Proceedings of the 1997 Conference on Advances in Neural Information Processing Systems 10, NIPS '97, 1998, pp. 570–576.
[31] S. Mukherjee, S.K. Biswas, D.P. Mukherjee, Recognizing interactions between human performers by dominating pose doublet, Mach. Vis. Appl. 25 (4) (2014) 1033–1052.
[32] S. Park, J.K. Aggarwal, Simultaneous tracking of multiple body parts of interacting persons, Comput. Vis. Image Underst. 102 (1) (2006) 1–21.
[33] A. Patron-Perez, M. Marszalek, I. Reid, A. Zisserman, Structured learning of human interactions in tv shows, IEEE Trans. Pattern Anal. Mach. Intell. 34 (12) (2012) 2441–2453.
[34] A. Patron-Perez, M. Marszalek, A. Zisserman, I. Reid, High five: recognising human interactions in TV shows, in: British Machine Vision Conference, 2010.
[35] R. Poppe, A survey on vision-based human action recognition, Image Vis. Comput. 28 (6) (2010) 976–990.
[36] K. Prabhakar, J.M. Rehg, Categorizing turn-taking interactions, in: ECCV (5), 2012, pp. 383–396.
[37] A. Prest, C. Schmid, V. Ferrari, Weakly supervised learning of interactions between humans and objects, IEEE TPAMI 34 (2012) 601–614.
[38] M. Raptis, L. Sigal, Poselet key-framing: a model for human activity recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013.
[39] M.S. Ryoo, J. Aggarwal, Spatio-temporal relationship match: video structure comparison for recognition of complex human activities, in: 2009 IEEE 12th International Conference on Computer Vision, 2009, pp. 1593–1600.
[40] M.S. Ryoo, J.K. Aggarwal, UT-Interaction Dataset, ICPR contest on Semantic Description of Human Activities (SDHA), 2010. <http://cvrc.ece.utexas.edu/SDHA2010/Human_Interaction.html>.
[41] S. Sadanand, J.J. Corso, Action bank: a high-level representation of activity in video, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2012.
[42] M.A. Sadeghi, A. Farhadi, Recognition using visual phrases, in: CVPR. IEEE, 2011.
[43] M. Sapienza, F. Cuzzolin, P. Torr, Learning discriminative spacetime action parts from weakly labelled videos, Int. J. Comput. Vis. (2013) 1–18.
[44] B. Solmaz, S.M. Assari, M. Shah, Classifying web videos using a global video descriptor, Mach. Vis. Appl. 24 (7) (2013) 1473–1485.
[45] A. Vahdat, B. Gao, M. Ranjbar, G. Mori, A discriminative key pose sequence model for recognizing human interactions, in: Eleventh IEEE International Workshop on Visual Surveillance, 2011.
[46] D. Waltisberg, A. Yao, J. Gall, L. Van Gool, Variations of a hough-voting action recognition system, in: ICPR, 2010.

[47] H. Wang, A. Kläser, C. Schmid, C.L. Liu, Action recognition by dense trajectories, in: IEEE Conference on Computer Vision & Pattern Recognition, 2011, pp. 3169–3176.
[48] H. Wang, C. Schmid, Action recognition with improved trajectories, in: IEEE International Conference on Computer Vision (ICCV), 2013.
[49] J. Wang, J.D. Zucker, Solving the multiple-instance problem: a lazy learning approach, in: In Proc. 17th International Conf. on Machine Learning, 2000, pp. 1119–1125.
[50] D. Weinland, R. Ronfard, E. Boyer, A survey of vision-based methods for action representation, segmentation and recognition, Comput. Vis. Image Underst. 115 (2) (2011) 224–241.
[51] Y. Yang, S. Baker, A. Kannan, D. Ramanan, Recognizing proxemics in personal photos, in: CVPR, IEEE, 2012, pp. 3522–3529.
[52] B. Yao, L. Fei-Fei, Grouplet: a structured image representation for recognizing human and object interactions, in: The Twenty-Third IEEE Conference on Computer Vision and Pattern Recognition. San Francisco, CA, 2010.
[53] G. Yu, J. Yuan, Z. Liu, Propagative hough voting for human activity recognition, in: Computer Vision ECCV 2012, Springer-Verlag, 2012, pp. 693–706.
[54] K. Yun, J. Honorio, D. Chattopadhyay, T. Berg, D. Samaras, Two-person interaction detection using body-pose features and multiple instance learning, in: Computer Vision and Pattern Recognition Workshops (CVPRW), 2012, pp. 28–35.
[55] C. Zhang, J.C. Platt, P.A. Viola, Multiple instance boosting for object detection, in: Y. Weiss, B. Schölkopf, J. Platt (Eds.), Advances in Neural Information Processing Systems, vol. 18, MIT Press, 2006, pp. 1417–1424.
[56] Y. Zhang, X. Liu, M.C. Chang, W. Ge, T. Chen, Spatio-temporal phrases for activity recognition, in: Proceedings of the 12th European Conference on Computer Vision, ECCV'12, vol. Part III, 2012, pp. 707–721.
[57] Z.H. Zhou, Y.Y. Sun, Y.F. Li, Multi-instance learning by treating instances as non-i.i.d. samples, in: Proceedings of the 26th Annual International Conference on Machine Learning, ICML '09, 2009, pp. 1249–1256.