RECOGNIZING INTERACTIONS BETWEEN HUMAN AND OBJECT BASED ON SOCIAL NETWORK ANALYSIS

Guang Yang,Yafeng Yin, Jin Xu, Hong Man

ECE Department, Stevens Institute of Technology, Hoboken, NJ, 07030

ABSTRACT

Recognizing human-object interactions in videos is a very challenging problem in computer vision research. There are two major difficulties lying in this task: (1) The detection of human body parts and objects is usually affected by the quality of the videos, for instance, low resolutions of the videos, camera motions, and blurring frames caused by fast motions, as well as the self-occlusions during human-object interactions. (2) The spatial and temporal dynamics of human-object interaction are hard to model. In order to overcome those natural obstacles, we propose a new method using social network analysis (SNA) based features to describe the distributions and relationships of low level objects for human-object interaction recognition. In this approach, the detected human body parts and objects are treated as nodes in social network graphs, and a set of SNA features including *closeness*, *centrality* and *centrality with relative velocity* are extracted for action recognition. A major advantage of SNA based feature set is its robustness to varying node numbers and erroneous node detections, which are very common in human-object interactions. An SNA feature vector will be extracted for each frame and different human-object interactions are classified based on these features. Two classification methods, including Support Vector Machine (SVM) and Hidden Markov Model (HMM), have been used to evaluate the proposed feature set on four different human-object interactions from HMDB dataset [1]. The experimental results demonstrated that the proposed framework can effectively capture the dynamical characteristics of human-object interaction and outperforms the state of art methods in human-object interaction recognition.

1. INTRODUCTION

Human action understanding is a challenge topic and has been widely studied in applications such as surveillance and video retrieval. Many methods [2, 3, 4] have achieved high performance on recognizing single human with periodical actions in clear background scenarios, such as Weizmann human action dataset [5] and KTH human action dataset [6]. With increasing demands on video content analysis, studies have been more focused on complicated scenarios. A recent work by Yin et al. [7] studied the the interactions among people based on BEHAVE dataset [8], which is a recorded data set with interactions within or between small groups, such as fighting, chasing, walking together and etc. These sequences are very close to real surveillance video. However there are more challenges lying in realistic videos, mostly sports and movie clips, which involve the interactions between human and objects.

In the study of recognizing human-object interactions, many researchers started from still images [9, 10, 11]. These existing methods on learning the interactions from static images are mostly using contextual information to build the relations between the object and human poses. Desia et al. [9] provided a unified model based on detecting spatial contextual relations of multiple objects. Yao and Fei-Fei [10] presented a mutual context model to jointly model the human poses with objects in still images by two contextual information, which are the co-occurrence statistics and the spatial context between objects and body part. And Prest et al. [11] introduced a weakly supervised algorithm to learn the object relevant for the action and its spatial relation to the human.

Some recent attempts have been made on recognize interactions between human and object in videos. Gupta et al. [12] added the psychological analyses of human perception to a Bayesian model to recognize objects and actions in videos in a fully supervised manner. Prest et al. [13] further developed their method on realistic videos based on [11], by including spatio-temporal annotations about object's locations and human actions. Another work by Si et al. [14] provided an AND/OR grammar based algorithm to semantically understand certain human daily activities in office.

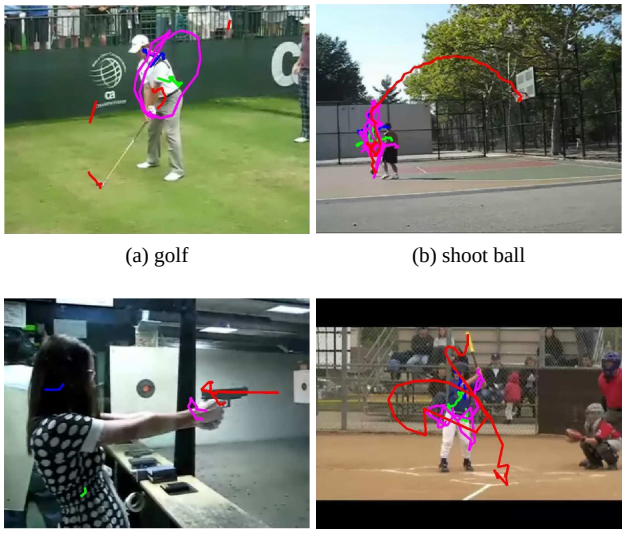
There are many challenges lying in the task of precisely identify the interactions between human and object in realistic videos. First, most existing methods require robust detection or tracking on human and objects, since the inconsistent information on human/body parts causes poor estimations on human poses and object positions. However, these tasks are very difficult in realistic videos. For one thing, it is common to see self-occlusions of the human body parts, or occlusions of objects by human or other less relevant background like branches of the trees. Another potential concern is that the quality of the video may vary significantly. The moving trajectories of objects may temporarily be lost because of the rel-

(a) Frame 1 (b) Frame 23 (c) Frame 48

Fig. 1: Challenges on object detections in realistic videos. In a sequence of human playing golf, (a) two hands are overlapped all the time. In (b), the golf club is invisible temporarily because of the fast motion speed and relatively poor quality of the video. In (c), the club is out of the scene. Besides, the camera itself is not fix and the background is not still.

atively poor quality of the video. The other reason of losing the trajectory of the object or human parts is when those parts reaching out of the camera field of view during the activity. These natural difficulties are illustrated in figure 1. Second, different from surveillance video which have a fixed camera scene, camera motions in realistic video must be taken into account as it affects the human/objects locations and the motion trajectory patterns. In this paper, we propose a novel framework of recognizing human-object interactions by considering the body parts and objects as nodes of social network graphs in the spatial dimension, and analyzing the features of the social network overtime to understand the video sequences. This framework consists of three stages. First is tracking the body parts and object, which provides the spatial information by a tracking algorithm of [15]. Second stage is constructing the social network graphs and extracting the SNA features to describe the temporal dynamic of an interaction in each sequence. This is inspired by Yin et al. [7], in which individual humans were modeled as nodes in social networks and hence the SNA feature set were used to describe small human group activities. At the last stage, two classifiers are applied to the feature vectors, namely, a K-means cluster followed by SVM and a Hidden Markov model classification. Each method reduces the length of feature vectors to a lower dimension. Experiments were conducted on typical sports activities from HMDB dataset [1].

The contribution of our work is threefold. First, this social network based framework characterizes the distribution of the activity globally as well as the distribution of each node in the social network. Second, the social network analysis based feature set dynamically organizes the body parts and object as nodes in a graph. It is able to handle various number of nodes as well as length of the sequence. Last but not least, this framework is able to tolerant missing information during the sequence. Therefore, by using the social network structured feature sets, it does not required strictly precise detections in the earlier stage, which is a major difficulty in realistic videos and many other scenarios.



(a) golf (b) shoot ball

(c) shoot gun (d) swing baseball

Fig. 2: Examples of activity trajectories of the body parts and the objects. Blue and green lines are the trajectories of the head and upper-body center. Magenta represents hands trajectory and red color is for the object.

2. HUMAN AND OBJECT TRACKING

In our approach, the human object interaction is considered as a serial activities happening among the key body parts and the object, which we consider as nodes in a social network graph. It is a challenging task to have perfect detectors or trackers to obtain the precise locations of specific body parts and objects under realistic image quality conditions. In this framework, a reliable tracking algorithm is applies to obtain the locations for these node. We adopt a state-of-the-art tracking algorithm in [15] to have the motion trajectories. In human object interactions, we consider only a few crucial parts providing meaningful information and forming the social network as nodes. The body parts include head and upper-body centers, which represent the human positions in the frame, and hand positions, which are important to reveal the physical contact between human and object. Figure 2 shows some examples of the activity trajectories. The trajectories of head, upper-body center and hands are colored in blue, green and magenta respectively. The red color represents the object motion path. It may discontinue in some places due to the occlusions or the limitation of the video data. However, the proposed social network analysis based framework is robust enough to handle such missing information.

3. SOCIAL NETWORK ANALYSIS BASED FEATURE

Features describing human action can usually be categorized into low level features and middle level features [7]. Low level features such as STIP or SIFT, are computed on the complete image region and more suitable for individual human actions. As discussed in the previous section, there are

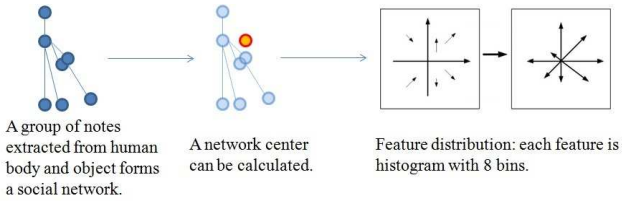


Fig. 3: The social network center is first calculated and other features are histograms distributed in 8 bins.

many difficulties in obtaining accurate low level features for a variety of human appearances, poses and camera motions. However, for the recognition of human object interactions, we need a structural information of the activities which can represent the interactions between body and object in a higher level. The social network analysis based feature set represents a middle level feature set to characterize such complex interactions.

A social network graph model a structure of social relationships (*ties*) among a set of individuals known as *actors* or *nodes*. Social network analysis was originally proposed in [16], which was designed to model the social structure in real world human societies. Inspired by the theoretic analysis of the social network [16] and its extensions on group activity recognitions [7], we introduce a new set of features to describe the dynamic properties of the human object interactions. Figure 3 shows the overview of this approach. To our best knowledge, this is the first time of using social network analysis based features to model human-object interactions.
Network center: Suppose there are n nodes in a network, the center $m_c = \left(\frac{1}{n}\sum_{i=0}^{n} x_i, \frac{1}{n}\sum_{i=0}^{n} y_i\right)$ is defined as the mass center of the network. The network center is calculated first, and other features are related to it.
Centrality: In general, centrality measures how the central node related to all other nodes in a social network. In our framework, centrality is used as a distance measurement between each node and the mess center of the network. Each node has a position $m_i = (x_i, y_i), (i = 1, \ldots, n)$ in the network and the relative position to the network center is a directional vector $ce_i = \overrightarrow{m_i m_c}$. The centrality vector is designed as an 8-bin histogram of directions accumulating the magnitude of the distance and it is normalized. The centrality vector is written as $\mathbf{Ce}_t = \{ce_i\}_t, (i = 1, \ldots, n; t = 8)$.
Closeness: Closeness describes how close an individual is to all the rest nodes in a network. In our framework, the directional distance between each node to every other node in the network is calculated. Therefore, the distance of every pair of nodes $cl_{i,j} = \overrightarrow{m_i m_j}$ are accumulated in the closeness vector which is also a histogram with 8 bins of directions. It is denoted as $\mathbf{Cl}_t = \{cl_{i,j}\}_t, (i, j = 1, \ldots, n; i \neq j; t = 8)$. Following these definitions, a set of social network analysis based features extracted at each frame will form an SNA feature vector with 26 dimensions, including network center, centrality, closeness and centrality with relative velocity. A feature vector is calculated at each desired frame and as one

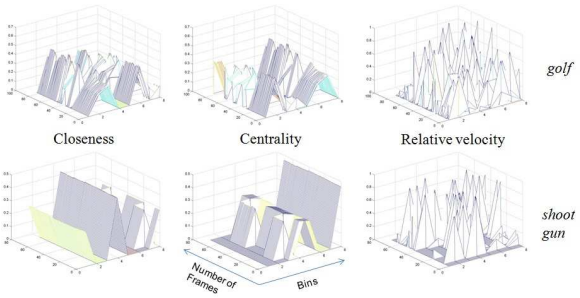


Fig. 4: Examples of social network analysis based features on interactions.

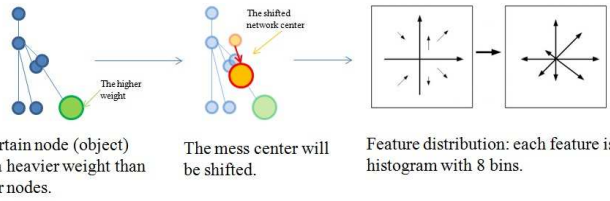


Fig. 5: In a weighted social network, the network center shifted due to the unequal weighted nodes.

entry in the feature matrix. A sequence with N frames will produce a SNA feature set in the dimension of $26 \times N$. Figure 4 shows examples of social network features from two interaction sequences, i.e. golf and shoot gun, respectively.

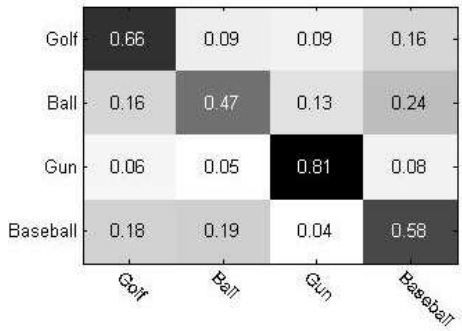
3.1. Weighted Social network

Each node has its contribution in terms of forming a dynamic social network, and some may play more important roles than others. Therefor, the centrality weight is introduced to measure the influence of a node in the network. **Centrality weights:** It assigns relative scores to all nodes in the network based on the concept that connections to high-scoring nodes contribute more to the score of the node than connections to low-scoring nodes. In the social network that describes human and object interaction, there are certain rules should be taken into consideration while assigning the weights of the nodes.

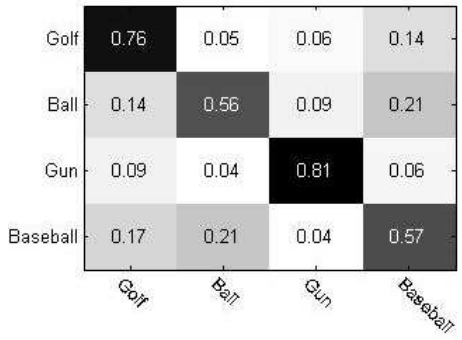
- The total weight of the network is normalized as one. $W_{hum} + W_{obj} = 1$, where $W_{hum} = \sum_{i}^{N_{hum}} w_i$ and $W_{obj} = \sum_{j}^{N_{obj}} w_j$.

- As human has more complicated structures and poses, there are more nodes on describing human than what on objects. $N_{hum} \geq N_{obj}$ and $W_{hum} \geq W_{obj}$.

- The objects have more important roles in understanding the interactions with human. Therefore each node on object has higher score than each node on human. $W_{hum} \geq W_{obj}$ and $w_i^{hum} \leq w_i^{obj}$.

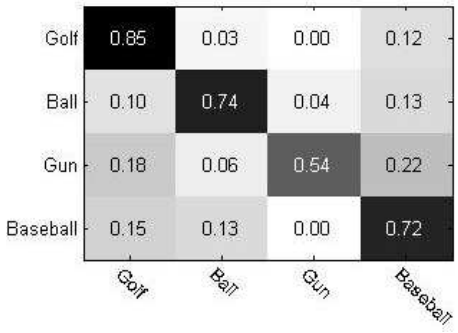


(a) SNA

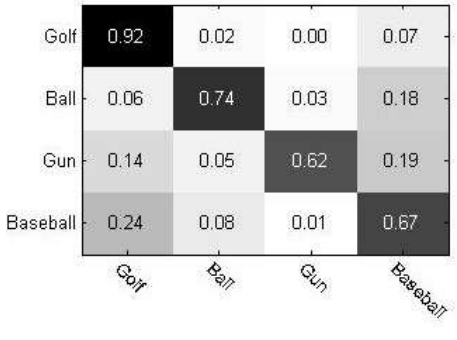


(b) weighted SNA

Fig. 6: The confusion matrix of SVM classification results on SNA and weighted SNA features.



(a) SNA



(b) weighted SNA

Fig. 7: The confusion matrix of HMM classification results on SNA and weighted SNA features.

4. EXPERIMENTAL RESULTS

We validate our method on HMDB dataset [1], which has 51 actions in five general types, and human motion with object interactions is one of them. Videos in this dataset are collected from various source of real world sources, like movies or YouTube. The video quality varies significantly, which makes the recognition task difficult.

In our experiments, we choose four classes of interactions: swing golf club, shoot basketball, shoot gun, and swing baseball bat. Each class has 100 clips. We apply body parts and object detectors on every five frames in each sequence, and then extract the social network features from the detection results. In each activity class, there are four nodes representing human bodies, which are head, upper-body center and both hands, and one more node as the object.

In the classification stage, we apply two classifiers, SVM and HMM. Data clips contain different number of frames, and each frame is represented in a feature vector of 26 dimensions. In the SVM approach, social network analysis based features from all frames are clustered and normalized before applying SVM. In our experiment, SVM with linear kernel is adopted and the training and testing data is divided into 50/50 with five-fold cross-validation. The classification results by SVM are shown in the confusion matrix in figure 6. In the HMM approach, we project the social network features into hidden Markov models with two hidden states and each state with two mixtures of Gaussian. The likelihood is computed between the test data and each trained HMM model, and the classification decisions are made according to the maximum likelihood. This experiment is also cross-validated for five times, and each time training and testing data is randomly divided into half and half. The average classification accuracy is 63% and 67% by SVM classifier on SNA features and weighted SNA features respectively, and 71% and 74% by HMM. Some classes even have over 80% correct recognitions. From the results, we can observed that weighted SNA features outpreform the un-weighted SNA features. The overall performance of our social network analysis based features is much higher than the benchmark [1] result by using the STIP features [4], which has accuracy around 20%.

5. CONCLUSIONS

In this paper, we proposed a new method for recognizing human object interactions. In this framework, key human body parts and the object are considered as nodes in a social network graph. And a set of social network analysis based features is introduced to capture the distributions of motion patterns among all the nodes overtime. It provides a global view of the activity while preserving the individuality of each node. Because of these, our method can tolerate missing information of the low-level detections on human body parts and the small object. We have shown that this method can achieve good performance in very challenging scenarios. In future work, we will extend this framework to model interactions involving more individuals and multiple objects.

6. REFERENCES

[1] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "HMDB: a large video database for human motion recognition," in *ICCV*, 2011.

[2] Juan Carlos Niebles, Hongcheng Wang, and Li Fei-fei, "Unsupervised learning of human action categories using spatial-temporal words," in *BMVC*, 2006.

[3] Shu-Fai Wong and Roberto Cipolla, "Extracting spatiotemporal interest points using global information," in *ICCV*, 2007, pp. 1 –8.

[4] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *CVPR*, 2008, pp. 1 –8.

[5] Moshe Blank, Lena Gorelick, Eli Shechtman, Michal Irani, and Ronen Basri, "Weizmann action database," `http://www.wisdom.weizmann.ac.il/~vision/SpaceTimeActions.html`.

[6] C. Schuldt, I. Laptev, and B. Caputo, "Kth action database," `http://www.nada.kth.se/cvap/actions`.

[7] Y. Yin, G. Yang, J. Xu, and H. Man, "Small group human activity recognition," in *ICIP*, sept. 2012.

[8] S. Blunsden and R. B. Fisher, "The behave video dataset: ground truthed video for multi-person," *BMVA*, vol. 4, pp. 1–12, 2010.

[9] C. Desai, D. Ramanan, and C. Fowlkes, "Discriminative models for static human-object interactions," in *CVPRW*, June 2010, pp. 9 –16.

[10] Bangpeng Yao and Li Fei-Fei, "Recognizing human-object interactions in still images by modeling the mutual context of objects and human poses," *PAMI*, vol. 34, no. 9, pp. 1691 –1703, sept. 2012.

[11] A. Prest, C. Schmid, and V. Ferrari, "Weakly supervised learning of interactions between humans and objects," *PAMI*, vol. 34, no. 3, pp. 601 –614, march 2012.

[12] A. Gupta, A. Kembhavi, and L.S. Davis, "Observing human-object interactions: Using spatial and functional compatibility for recognition," *PAMI*, vol. 31, no. 10, pp. 1775 –1789, oct. 2009.

[13] A. Prest, V. Ferrari, and C. Schmid, "Explicit modeling of human-object interactions in realistic videos," *PAMI*, vol. PP, no. 99, pp. 1, 2012.

[14] Zhangzhang Si, Mingtao Pei, B. Yao, and Song-Chun Zhu, "Unsupervised learning of event and-or grammar and semantics from video," in *ICCV*, nov. 2011, pp. 41 –48.

[15] Kaihua Zhang, Lei Zhang, and Ming-Hsuan Yang, "Real-time compressive tracking," in *ECCV*, 2012, pp. 864–877.

[16] Duncan J. Watts and Duncan H. Strogatz, "Collective dynamics of 'small-world' networks," *Nature*, vol. 393, no. 6684, pp. 440–442, June 1998.