21st International Conference on Pattern Recognition (ICPR 2012)
November 11-15, 2012. Tsukuba, Japan

Activity Recognition Based on Semantic Spatial Relation

Lingxun Meng[1,2] Laiyun Qing[1,2] Peng Yang[3] Jun Miao[2] Xilin Chen[2]
Dimitris N. Metaxas[3]
[1]*School of Information Science and Engineering,*
Graduate University of Chinese Academy of Sciences(CAS), Beijing 100049, China
[2]*Key Laboratory of Intelligent Information Processing,*
Institute of Computing Technology, Chinese Academy of Sciences ,Beijing 100190, China
[3]*Computer Science Department, Rutgers University, Piscataway NJ 08854, USA*
{*lxmeng,lyqing,jmiao,xlchen*}*@jdl.ac.cn* {*peyang,dnm*}*@cs.rutgers.edu*

Abstract

We propose an approach to recognize group activities which involve several persons based on modeling the interactions between human bodies. Benefitted from the recent progress in pose estimation [1], we model the activities as the interactions between the parts belong to the same person (intra-person) and those between the parts of different persons (inter-person). Then a unified, discriminative model which integrates both types of interactions is developed. The experiments on the UT-Interaction Dataset [2] show the promising results and demonstrate the power of the interacting models.

1. Introduction

Human activity recognition has great scientific importance and many practical applications, such as surveillance, human-computer interaction, image and video search.

Many progresses have been gained in activity recognition with one actor, such as running, working and waving hands [3, 4, 5]. Some systems considered the interactions between the actor and the object. For example, Yao et al [6] modeled an activity as the pose of the player by human parts detection and the interaction between the key object (e.g. basketball) and the human parts. Few works dealt with activities involving more than one person. Lan et al [7] proposed an algorithm to capture contextual information including person-group interactions and person-person interactions to recognize group activities.

Many types of features have been explored in activity recognition. The appearance features such as the Bag-of-Words (BoWs) method have been successfully applied on the activity recognition [3, 8] by using discriminative space-time features [9]. However, the spatial information is missed in such a representation.

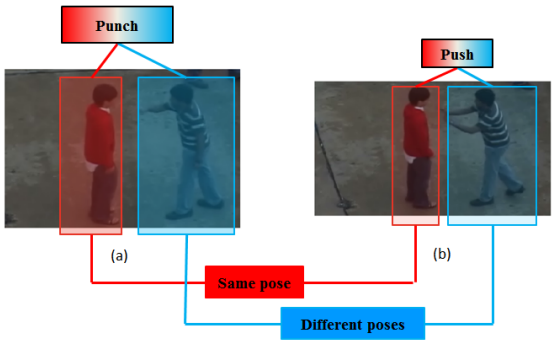


Fig.1: Interactive poses. The red bounding box shows the person with same pose in different activities, while the blue one shows the person with different poses in different activities.

Then the spatial relations are introduced. The representations of spatial relations in activity recognition can be categorized into three levels: points of interest level, which captures the spatial relations between the points of interest [10]; parts level (pose), which uses the relative locations of body parts implicitly [5, 4] or explicitly [6] to recognize activities; persons level, which might be useful for recognizing group activity [7].

Here we focus on the activities involving several persons. It's well known that pose is critical for activity recognition. Different from [4, 5] using human poses to recognize actions implicitly, we use the locations of the human body parts explicitly and therefore the mod-

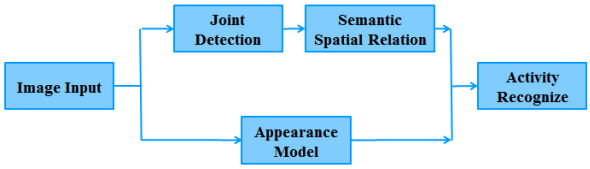


Fig.2: Our framework of activity recognition.

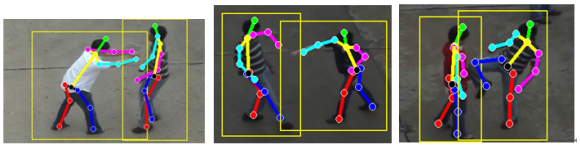


Fig.3: Some examples of the detected joints.

els are more explainable. Fig.3 shows some detected human parts in the different activities. Further, we argue that the activities involves the interactions among the actors, i.e., the activity in Fig.1(a) is not a Punch activity until the punching actor is touching the other one by a punching pose.

In this paper, we propose a unified representation for both the pose of a person and the interactions among different actors. The framework of the proposed method is shown in Fig.2. The joints detection is applied on the input image, and the semantic spatial information within a person and among the actors are extracted based on the locations of the joints. The appearance features are further integrated into action recognition as the complementary.

2. Interaction Representation

Our method models an activity as the interactions between the parts of human bodies. We assume that an image consists of the descriptive local human poses which we refer as *Intra-Person* interactions and the interactions between the parts of different persons which we refer as *Inter-Person* interactions.

2.1. Joints Detection

Recent works on pose estimation have made some big progresses. In this paper we use the state-of-the-art detector [1] to locate the joints' locations.

Yang et al [1] modeled human key joints into a tree structure and constructed a score function to dynamically search human and the poses efficiently in images. The score function combined appearance confidence with limbs' deformation cost, part attributes confidence and relative matching. Their system was tested on several real-life datasets and got a promising improvement.

In order to boost the performance of joints locating, we train multiple pose estimators on the dataset [2] to handle huge variance among activities. While training each model, we choose the training samples of a specific activity as positive samples and the others as negative samples according to the test protocol of the dataset [2].

For a testing image, we try each model to detect joints and choose the one with maximum score. The experimental results show that this strategy can effectively improve locating joints. Some examples of the detected joints are shown in Fig.3.

2.2. Semantic Spatial Relation

We can see from Fig.1 that people's poses discriminate different activities effectively. However, the poses of individual persons are not good enough for the task because of the sharing poses between different interactive actions. For example, the first person is standing still in the activity of Punch (Fig.1(a)) and Push (Fig.1(b)) and the second person with different poses is the main actor. Furthermore, the activity is not a Push until the pushing actor is touching the other one, i.e., the distance between the two actors is also one of the key clues. Therefore both the poses of the individual persons (referred as *Intra-Person*) and the relationships between them (referred as *Inter-Person*) are modeled in our algorithm.

Intra-Person: The pose of an individual person is represented as the pairwise joints' relative locations. As illustrated in Fig.4(a), the *Intra-Person* features of the person p are represented as

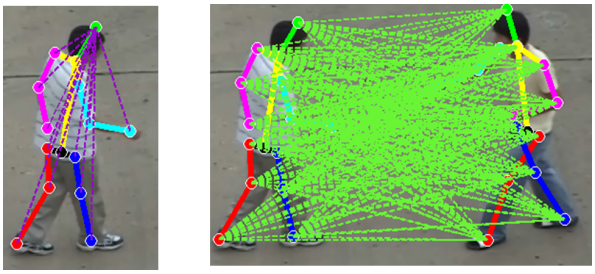
$$\mathbf{s}_{ij}^{p} = [|dx| \quad |dy| \quad d\theta] \tag{1}$$

where $dx = x_i - x_j$, $dy = y_i - y_j$, $d\theta = \theta i - \theta j$, (x_i, y_i),and (x_j, y_j) are the coordinates of joint l_i and l_j respectively, θ_i and θ_j are the orientations of joint l_i and l_j respectively, normalized by the height of person p. $i, j \in \{1, 2, ..., N\}$, and N is the number of joints of one person. In the experiments, the joint orientations are quantized into eighteen discrete values.

Inter-Person: The interactive pose, which is defined as the pairwise relative locations between the joints from different subjects. As shown in Fig.4(b), we represent it as

$$\mathbf{s}_{i,j}^{p_m,p_n} = [|dx| \quad |dy| \quad d\theta] \tag{2}$$

where $s_{i,j}^{p_m,p_n}$ represents interactive pose between person p_m and $p_n, m, n \in \{1, 2, ..., K\}$, K is the number



(a) (b)

Fig.4: Semantic Spatial Relations. (a)Intra-Person spatial relations drawn by dash line in purple; (b) Inter-Person spatial relations drawn by dash line in green.

Fig. 5: UT-Interaction Dataset examples in still images. (a) hand-shake; (b) kick; (c) hug; (d) point; (e) punch; (f) push.

of interactive people. The definitions of dx, dy and $d\theta$ are the same with *Intra-Person*, whereas dx, dy are normalized by the distance between person p_m and p_n.

3. Interaction Recognition

We prefer our task to a discriminative function $F : I \times \mathbf{y} \rightarrow R$, over an image I and its class labels Y where F is parameterized by Θ. During testing, we predict the class label of Y^* of an input image I as:

$$Y^* = \arg\max_{Y \subset \mathbf{y}} F(I, Y \mid \Theta) \tag{3}$$

We can write

$$F(I, Y \mid \Theta) = \sum_{p_m} \alpha_m^T \cdot Intra(p_m, Y) + \sum_{(p_m, p_n)} \beta_{mn}^T \cdot Inter(p_m, p_n, Y) + \gamma^T \cdot App(I, Y) \tag{4}$$

where $Intra(p_m, Y)$ represents the *Intra-Person* spatial relations of person p_m, $Inter(p_m, p_n, Y)$ stands for spatial relations of the *Inter-Person* between person p_m and p_n, and $App(I, Y)$ is the appearance model. In the experiments, the system uses the output of SVMs trained on *Intra-Person* spatial relation features instead of tuning the weights of individual features. The same operations are done on *Inter-Person* and appearance model parts. We choose HoG [11] as the appearance features.

4. Experiment

We test our method on the public UT-Interaction Dataset [2], which has two sets of video data. The video sequences in Set1 are taken in a parking lot with slightly different zoom rate and the background is relatively static, whereas videos in Set2 contains more jitters which results in more phantoms in still images. There are six different interactions: Hand-shake, Hug, Kick, Point, Punch and Push. Fig.5 shows snapshots of these activities.

We extract key frames from the original videos to construct our testing dataset. The key frames defined here are extracted from the original videos' middle one third frames with five frames as the sampling interval. Set1 contains 497 samples and Set2 contains 464 samples. We follow the original video-test protocol,i.e., the 10-fold leave-one-out cross validation per set. For training the joint detection models, fourteen joints on human bodies are manually annotated, namely $N = 14$. The contributes of the three parts in the system (4) are tested: intra personal features, inter personal features and the appearance model. Table 1 shows the experimental results using the mean classification accuracy and mean average precision. In the current implementation, the weights of the three parts in (4) are learned using SVM. The experimental results show that the final results are not sensitive to the weights. It can be seen from Table 1 that the semantic spatial relation(*Intra-Person* and *Inter-Person*) is better than HoG templates. Besides, our semantic spatial relation is complementary to template matching, and combining these two approaches improves the performance significantly.

We also compare our model with the other two state-of-the-art methods [12] and [10] on our dataset. The first baseline BoF is the bag-of-features classifier [12], aggregating quantized responses of densely sampled SIFT features in spatial pyramid representation, using an intersection kernel. Note that this is a strong baseline, which was shown in [12] to outperform other state-of-the-art methods in single-actor recognition. During

Table 1: Each part contributes to the final result

Each Part	mAcc.	mAP
Intra-Person[a]	77.68	78.35
Inter-Person	**84.27**	84.33
Intra+Inter[b]	83.11	**86.24**
HoG Template[c]	83.59	85.96
Final	**87.42**	**91.81**

[a] *Intra-Person* 1 + *Intra-Person* 2
[b] *Intra-Person**0.4 + *Inter-Person**0.6
[c] HoG-Person 1 + HoG-Person 2

testing, we give the interactive bounding box for better results of [12] and set the vocabularies(K=1024). The second baseline, pairwise spatial relations(PSR), is also based on a bag-of-features classifier [10] combining the spatial relationship of pairwise words. we use SIFT instead of STIP-HoG as our local features since our target are interactions in still images. we set vocabularies K=200. During the experiment, the bigger K doesn't bring better performance. We only give the result of pairwise spatial relation of BoWs instead of the combination of PSR and histograms of words. Table 2 shows the results.

Table 2: Methods test on UT-Interaction Dataset

Set	Set1		Set2	
Methods	mAcc.	mAP	mAcc.	mAP
BoF [12]	77.12	79.95	70.06	73.52
PSR [10]	49.09	45.90	45.09	47.26
Intra+Inter	83.11	86.24	77.20	78.26
Ours	**87.42**	**91.81**	**82.19**	**83.60**

Frames-Voting for Video : For proper comparison with previous methods testing on UT-Interaction Video Dataset, the labels of key frames extracted from the original videos are used to vote for the corresponding videos' label. Our final result on UT-Interaction Video Contest reaches **0.983** on Set1 and **0.922** on Set2, compared with best result reported in [13] as 0.88 on Set1 and 0.77 on Set2.

5. Conclusion

We propose a novel activity recognition algorithm based on the interactions within each person and the interactions among the persons involved in the activity. Different from previous works avoiding estimating poses, we detect interactive people's joints by using multiple pose estimators and extract semantic spatial relations: intra/inter spatial relations. Integrated with the appearance features, our method gets the significant improvement compared with the state-of-the-arts, which demonstrates the power of the proposed semantic features. The future work includes how to integrate actor pose estimation with activity recognition into an unified work.

Acknowledgement. This research is partially sponsored by National Basic Research Program of China (No.2009CB320902), Beijing Natural Science Foundation (No.4102013), Natural Science Foundation of China (No.61070116, 60970087, 61070149 and 61001108).

References

[1] Y. Yang and D. Ramanan, "Articulated pose estimation with flexible mixtures-of-parts," in *CVPR*, 2011.

[2] M. Ryoo and J. Aggarwal, "Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities," *ICCV*, 2009.

[3] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: A local svm approach," in *ICPR*, 2004.

[4] W. Yang, Y. Wang, and G. Mori, "Recognizing human actions from still images with latent poses," in *CVPR*, 2010.

[5] Y. Xie, H. Chang, Z. Li, L. Liang, X. Chen, and D. Zhao, "A unified framework for locating and recognizing human actions," in *CVPR*, 2011.

[6] B. Yao and L. Fei-Fei, "Modeling mutual context of object and human pose in human-object interaction activities," in *CVPR*, 2010.

[7] T. Lan, Y. Wang, W. Yang, and G. Mori, "Beyond actions: Discriminative models for contextual group activities," *NIPS*, 2010.

[8] J. Niebles, H. Wang, and L. Fei-Fei, "Unsupervised learning of human action categories using spatial-temporal words," *IJCV*, 2008.

[9] I. Laptev, "On space-time interest points," *IJCV*, 2005.

[10] P. Matikainen, M. Hebert, and R. Sukthankar, "Representing pairwise spatial and temporal relations for action recognition," *ECCV*, 2010.

[11] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *CVPR*, 2005.

[12] V. Delaitre, I. Laptev, and J. Sivic, "Recognizing human actions in still images: a study of bag-of-features and part-based representations," *BMVC*, 2010.

[13] M. Ryoo, C. Chen, J. Aggarwal, and A. Roy-Chowdhury, "An overview of contest on semantic description of human activities (sdha) 2010," *ICPR*, 2010.