Human Group Activity Analysis with Fusion of Motion and Appearance Information

Zhongwei Cheng[1], Lei Qin[2], Qingming Huang[1,2], Shuqiang Jiang[2], Shuicheng Yan[3], Qi Tian[4]

[1]Graduate University of Chinese Academy of Sciences
Beijing 100190, China

[2]Key Lab of Intelli. Info. Process., Inst. of Comput. Tech., CAS
Beijing 100190, China

[3]Electrical and Computer Engineering, National University of Singapore
Singapore, 117576

[4]Department of Computer Science, University of Texas at San Antonio
TX 78249, U.S.A.

{zwcheng, lqin, qmhuang, sqjiang}@jdl.ac.cn, eleyans@nus.edu.sg, qitian@cs.utsa.edu

ABSTRACT

Human activity analysis is an important and challenging task in video content analysis and understanding. In this paper, we focus on the activity of small human group, which involves countable persons and complex interactions. To cope with the variant number of participants and inherent interactions within the activity, we propose a hierarchical model with three layers to depict the characteristics at different granularities. In traditional methods, group activity is represented mainly based on motion information, such as human trajectories, but ignoring discriminative appearance information, e.g. the rough sketch of a pose style. In our approach, we take advantage of both the motion and the appearance information in the spatiotemporal activity context under the hierarchical model. These features are inhomogeneous. Therefore, we employ multiple kernel learning methods to fuse the features for group activity recognition. Experiments on a surveillance-like human group activity database demonstrate the validity of our approach and the recognition performance is promising.

Categories and Subject Descriptors

H.3.1 [**INFORMATION STORAGE AND RETRIEVAL**]: Content Analysis and Indexing – *Abstracting methods*;

I.2.10 [**ARTIFICIAL INTELLIGENCE**]: Vision and Scene Understanding – *Video analysis*

General Terms

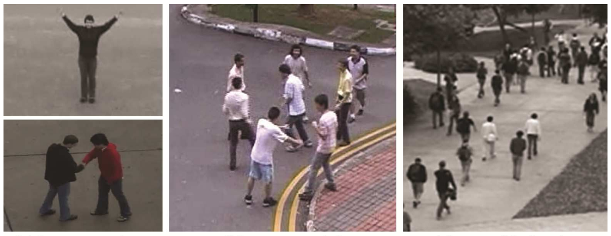
Algorithms, Design, Experimentation

Keywords

Human Group Activity, Activity Analysis, Feature Fusion

1. INTRODUCTION

As the amount of digital media grows rapidly, the demand of analyzing, understanding and utilizing these data is rising. Human activity analysis, as an important and challenging task in video content analysis, has drawn growing attention of worldwide researchers for its great potential and promising applications in industry, entertainment, security and medical domains. In recent years, human activity analysis has made notable progress and the representative state-of-the-art approaches are reviewed in [1].



(a) human action (b) group activity (c) crowd behavior

Figure 1. Human activity categorization

Human activity is a complex concept with diverse semanteme, various expressions and different scales. To present our work clearly, we categorize human activities into three levels: human action, group activity and crowd behavior according to the number of participants and the complexity of the activity. Fig.1 illustrates instances of the three categories. The methods to analyze these activities should be different and be adjusted to their inherent characteristics. As the smallest scale of activity, human action covers single person action and the interaction between a pair of persons. For the movements of human body are significant properties of action, the related methods aim at modeling the action pattern with localized features of motion or appearance information [2, 3]. Group activity is the intermediate level activity with countable people and complex interactions. The “group” here is also mentioned as “small group”, which consists of three or more persons with possible occlusions. To recognize group activity, the analysis of actions of individuals as well as their overall relations becomes essential [1]. For crowd behavior which consists of visually uncountable people, it is impossible to track individuals and recognize their actions to understand the whole crowd. Thus a reasonable choice is to model the entire crowd with motion information [4], such as optical flows or motion trajectories.

In this paper, we focus on the group activity. It is difficult to handle the structured property of human group activity through single models. Therefore, researchers have attempted to take advantage of layered models. Three level localized causalities are introduced to characterize relations within, between and among motion trajectories [5]. Cheng *et al* proposed a three layered model to describe the activity patterns [8]. However, the top level of their models, which is to depict the holistic activity pattern, is not represented properly. Interactions of person-person and person-group are considered in [7], but the complete layered structure is not explicitly defined.

To represent the activity patterns of human group, motion information is widely utilized. Ni *et al*. describes motion information by digital filter responses on participants’ motion trajectories [5], while Gaussian Process regression is applied to represent motion patterns in [8]. Although the motion information is proper to depict the activity pattern, appearance information can

*Area Chair: Kiyoharu Aizawa

MM’11, November 28–December 1, 2011, Scottsdale, Arizona, USA.
Copyright 2011 ACM 978-1-4503-0616-4/11/11…$10.00.

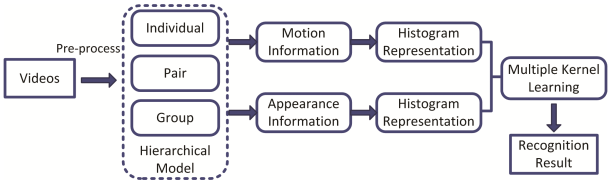


Figure 2. Framework of our approach

be a complementary to draw the discriminative characters for promoting the understanding of group activity. Zhu *et al* utilized localized appearance, SIFT, as a descriptor of activity [6]. But localized features can be unreliable due to appearance variance and noise. Thus the holistic appearance features are more appropriate to represent activity patterns.

In this paper, we propose a new approach to analyze human group activity, with the framework demonstrated in Fig.2. Different to previous works, by considering spatiotemporal context information of the activity, we construct a hierarchical model for human group activity and represent activity pattern with both the motion and the appearance information. The inhomogeneous features are expressed in histogram form respectively and fused with Multiple Kernel Learning methods.

The remainder of this paper is organized as follows. In section 2, we introduce the hierarchical model of group activity. The representations of motion and appearance information are described in section 3. Experimental validation and conclusions come with the last two sections.

2. HIERARCHICAL MODEL FOR HUMAN GROUP ACTIVITY

As introduced in the previous section, group activity has more participants and interactions than human action and more visible individual movements than crowd behavior. To analyze the group activity, we need to recognize the individual actions, the pairwise interactions, and the overall motion pattern of the group. Semantically, they are mutual promotional components for understanding the group activity. However, due to the diversity of participants and their variant numbers, it is hard to cope with the patterns of human group activity. Jointly considering different granularities of activity pattern is a reasonable solution for modeling the activity. In this paper, we introduce a three layered hierarchical model for group activity recognition. Fig.3. illustrates an activity instance represented under our hierarchical model. The bottom row shows all components of activity patterns in three granularities. The corresponding realistic examples of video frames are demonstrated in the top row. The three levels represent the group activity pattern from different aspects, and they are complementary for analyzing the group activity. Details of the model levels are introduced in the rest of this section.

Individual Level: This level focuses on the action pattern of a single participant. By depicting the individual movements of people within a group activity, we can obtain a general knowledge of that activity. It should be noted that appearance information is useful to represent discriminative individual patterns, which is consistent with human cognition. Especially, the general holistic appearances of people’s movements can properly describe the style of key poses in actions. Different with usual human action representation, action patterns of a single participant do not work separately for group activity. As our target is to represent the group activity, knowing a participant doing a specific type of action is not of much significance. While to present the frequency of an action appearing in a category of activities makes more sense.

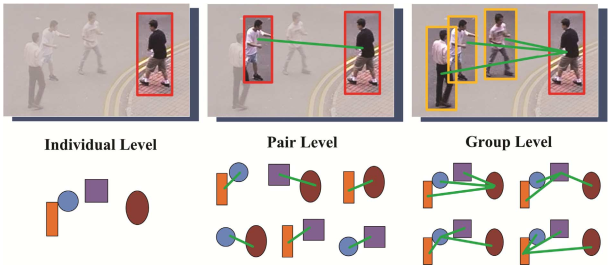


Figure 3. Hierarchical model for group activity

Pair Level: Pairwise representations are designed to handle the internal interactions of group activity. Each pair of participants reveals some kind of interaction within the activity, no matter if it is semantically significant. These pairwise patterns can be regarded as medial level components of group activity pattern. Although some pairs may connect unrelated persons in the activity, it is reasonable that the discriminative pairwise patterns for a specific activity class have statistical significance.

Group Level: The top level of group activity model is to express the pattern of the entire group, with handling the great diversification of group activity instances. We introduce a representation of the holistic pattern, noted as grouptron. The grouptron can be taken as high level component of the activity pattern. To treat a specific person as a reference, the grouptron represents the relativities in action with all other participants. One grouptron indicates a specific view of the group activity. Thus we think that collection of the grouptrons of a group activity can depict the activity pattern properly in group level.

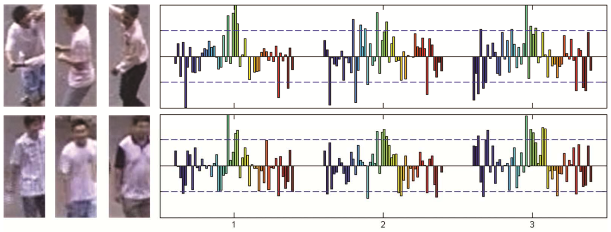
As presented above, our hierarchical model of group activity employs statistical property of component patterns in each level to handle variations and generate discriminative representations. The decomposition of group level pattern into different personal views of the group is different from directly modeling of the entire group, which makes our model more expressive. This model also provides a framework for considering different types of information at multiple granularities.

3. FEATURE REPRESENTATION

We adopt different features for different levels of the hierarchical model. Both motion and appearance information are taken into account. Motion features are based on trajectories of participants, and as object tracking is not the topic of this paper, it is considered as a pre-processing step. To ease the complexity of tracking, we split videos into smaller video segments according to the timelines. Thus motion features are extracted from video segments. Nevertheless, appearance features are extracted on the basis of video frames.

3.1 Motion Representation

Motion information, especially the motion trajectories of people [5, 6, 8], is effective for activity representation. In this paper, we step forward by not just describing the trajectory pattern itself but also representing the context information of the activity as well. For the individual level, motion pattern lies in the time-variant locations of the humans, or specifically their motion trajectories. For the pair level, as pairwise motion patterns reflect the interactions, we represent it through the time-variant distances between the pair of humans. As to the group level, the



(a) Appearance samples (b) 50D PCA of corresponding HoG features

Figure 4. Appearance information in activity

representation of a grouptron needs to cover the characters of all pairwise relationships between the reference person and all the other ones. Thus we present it through the statistics of all the pairwise information (the time-variant distances). As all motion representations are extracted from unified form (time-variant data, which can be treated as some kind of trajectory), we apply the same feature extracting paradigm for all the three levels of the activity model. To simplify the explanation, we take the physical meanings of motion trajectory in the individual level to introduce the motion features extracted.

Given a set of motion trajectories within a video segment, we have $\{T_i(t_j) \mid i=1,..,n \text{ and } j=1,..,m\}$, where n is the number of trajectories and m is the length of trajectories. The motion representation consists of two parts:

Movement Property: This part depicts the characteristic of a trajectory itself. Actually, a trajectory consists of temporal conjunctive positions, which can be regarded as a kind of temporal context of the activity. Inspired by [8], we employ Gaussian Process (GP) to represent the probabilistic variation of the trajectory. With modeling a trajectory as a GP over time and applying the squared exponential covariance function, which is denoted in Eq.1, the hyper-parameters $\theta = [\sigma_f , l , \sigma_n]$ are utilized to represent the character of the trajectory. Please refer to [9] for the details of GP regression to obtain the hyper-parameters θ.

$$T(t) \sim \mathcal{GP}(0,\Sigma),\ \Sigma = {\sigma_f}^2 \exp\left(-\frac{(\mathrm{t} - \mathrm{t}')^2}{2l^2} \right) + \sigma_n^2 \delta_{jj'} \tag{1}$$

Moreover, location change of the person within a trajectory is used to express the movement scale.

$$c_i = \left| T_i(t_1) - T_i(t_m) \right| \tag{2}$$

Velocity is also an important property of movement. We adopt the average velocity ϑ and velocity ratio r to represent the intensity and the complexity of the movement respectively.

$$\vartheta_i = \frac{1}{m-1} \sum_{j=2}^{m} v_j \tag{3}$$

$$r_i = \left| \vartheta_i \right| \Big/ \frac{1}{m-1} \sum_{j=2}^{m} \left| v_j \right| = \left| \sum_{j=2}^{m} v_j \right| \Big/ \sum_{j=2}^{m} \left| v_j \right| \tag{4}$$

where $v_j = T(t_j) - T(t_{j-1})$.

Movement Context: For group activity, individual's movement is influenced by other participants. By considering it in the feature view, we bring the activity context information into the motion representation. This activity context is a kind of spatial context of one's movement in a group activity, which indicates the movements of other persons in the same activity and the influence to him/her. The relative location change *rc* and relative average velocity *rv* are used to depict the context.

$$rc_i = c_i - \frac{1}{n} \sum_{p=1}^{n} c_p \tag{5}$$

$$rv_i = \vartheta_i - \frac{1}{n} \sum_{p=1}^{n} \vartheta_p \tag{6}$$

Therefore, the motion information of a trajectory T_i can be represented as the feature vector of $[\theta_i, c_i, \vartheta_i, r_i, rc_i, rv_i]$.

3.2 Appearance Representation

Besides motion, general appearance, like shape, is also an important clue for activity recognition. It provides additional discriminative information which is complementary to motion information. For example, we can estimate the activity in Fig.1(b) without motion information. We employ the Histograms of Oriented Gradients (HoG) [10] to represent the shape information in this paper. The appearance of activity may be diverse majorly due to the reason that an activity usually consists of several individual actions. To grab the holistic, or as to say the 'style', of the appearance information of activity, we apply Principal Components Analysis (PCA) to the HoG features. And we believe these principal components can reflect the general appearance characters. As illustrated in Fig.4(a), in the top row are three samples of *fight* activity, and in the bottom are of *walk-in-group*. These samples are visually discriminative. And Fig.4(b) demonstrates 50 dimensions (50D) PCA to the HoG features corresponding with samples in Fig.4(a). It can be observed that the appearance representations can supply some discriminative information.

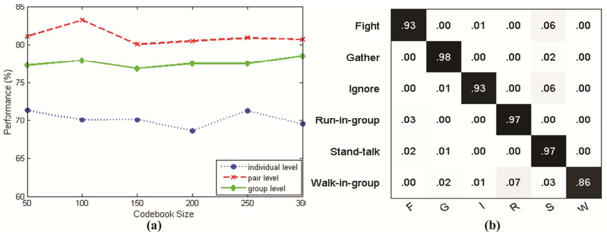
In this paper, we apply appearance representation just to the individual level of our activity model. It may be a possible way to compose pair of samples to a new pairwise sample and then apply the representation on it for the pair level. And other appearance descriptors, like GIST, can be employed to represent the activity appearance "style".

With motion and appearance feature representations on the hierarchical activity model, multiple inhomogeneous features are extracted for the group activity recognition. The way to fuse these features affects the recognition performance. Directly concatenating them into a larger feature is a practicable method, but would not be suitable for the inconsistent dimensions of different features. In this paper, we deal with the fusion of features by Multiple Kernel Learning with different features generating kernel matrix respectively.

4. EXPERIMENTS

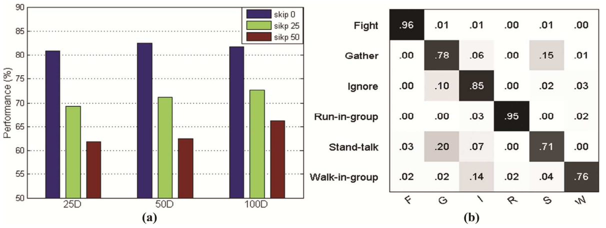
Experiments are performed on Human Group Activity (HGA) dataset in [5] to validate our approach. The HGA dataset consists of 476 videos in total with 6 categories of group activities. Each activity instance contains 4-8 participants. The trajectories of people in videos are generated by blob tracking with manual initializations [5]. According to different collecting conditions, the dataset is organized in 5 sessions. In the experiments, we test the performance with the average classification accuracy of the 6 activity categories through leave-one-session-out strategy. We first test the performance of motion feature, then the appearance feature, and finally we compare our approach with the state-of-the-art methods.

To generate motion features of activities, we extract motion representations from three model levels respectively. Then we apply Bag of Words (BoW) method to obtain statistical histogram features from every activity video [2]. About 50% of data are selected randomly for codebook generation. K-means clustering is applied separately for different model levels. Multi-



(a) (b)

Figure 5. Performance of motion feature



(a) (b)

Figure 6. Performance of appearance feature

class Support Vector Machines with a χ^2 distance based kernel are employed to recognize the group activities. Fig.5(a) shows the performance under different codebook sizes with motion features from our three model levels. The performance of pair and group level features is obviously better than individual level, which implies the interaction and holistic patterns are more discriminative for group activity. We can notice that the performances of three levels reach optimum at different codebook sizes, probably due to diverse sparseness of the histograms caused by the varied feature amount in different model levels. The confusion matrix of recognition results with all motion information is shown in Fig.5(b).

Appearance features are extracted just in individual level for this dataset. We assign a bounding box to a person according to his/her center point from the trajectory to get an appearance sample. With an estimated vertical compensation for the affine transformation, size of bounding box varies to get proper sample size. These samples are resized to the same scale, and then 972D HoG features are extracted. Then the HoG features are projected to lower dimensionality by PCA and clustered to generate codebook. As appearance feature is based on frames, samples from a small number of video data are sufficient. Thus we randomly choose only 10% of data to generate PCA coefficients and the codebook of BoW. The size of codebook is set to 512 for the plentiful appearance samples. Different dimensions and component selection strategies of PCA are tested and results are presented in Fig.6(a). The "skip *M*" stands for dropping the top significant *M* principle components and selecting the following *N* as features, demonstrated with *M*=0, 25, 50 and *N*=25, 50, 100. Performance of "skip 0" outperforms the others, illustrating the top principle components, which present the general and holistic appearance, are more expressive than those in the rear, which present the specific and detailed appearance. Another observation is that the performance of the appearance features is superior to motion feature in individual level, expressing the value of appearance in activity recognition. From the confusion matrix in Fig.6(b), we can discover that appearance information is more effective for intense activities such as *fight* and *run-in-group*.

Feature fusion is achieved by multiplication of kernel matrices of motion and appearance features in different model levels respectively (totally four kernels, one for appearance and three for motion). The performance of ours and other methods are listed in Table.1. It shows that our approach outperforms previous methods. And it can be observed that combining both motion and appearance features is better than just using motion information with over 3% performance gain, which reveals the impact of appearance.

Table 1. Performance comparison with other methods

Methods	*Performance*
Ni. *et al* [5]	73.5%
Cheng. *et al* [8]	91.8%
Zhu. *et al* [6]	87%
motion	93.7%
motion + appearance	**96.8%**

5. CONCLUSION

We propose an approach for the analysis of human group activity in this paper. With a hierarchical model, we represent the group activity pattern from three complementary levels. Other than previous work, we consider both the motion and appearance information with activity context. Experiments on HGA dataset validate the effectiveness of our approach and the recognition performance is notable.

To design informative descriptors for grouptrons is nontrivial and worth of interests in the future work. We will also investigate appearance representations for pair and group levels and explore effective expressions of discriminative components of group activity pattern.

6. ACKNOWLEDGMENTS

This work was supported in part by National Natural Science Foundation of China: 61025011, 61035001 and 61003165, in part by National Basic Research Program of China (973 Program): 2009CB320906, and in part by Beijing Natural Science Foundation: 4111003. This work was supported in part to Dr. Qi Tian by NSF IIS 1052851, Faculty Research Awards by Google, FXPAL and NEC Laboratories of America, respectively.

7. REFERENCES

[1] Aggarwal, J.K. and Ryoo, M.S. 2011. Human activity analysis: A Review. ACM Computing Surveys (CSUR), v.43 n.3, p.1-43.

[2] Niebles, J.C., Wang, H. and Fei-Fei, L. 2008. Unsupervised learning of human action categories using spatial-temporal words. International Journal of Computer Vision (IJCV) 79, 3(Sep).

[3] Ryoo, M.S. and Aggarwal, J.K. 2009. Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities. In ICCV.

[4] Mahadevan, V., Li,W., Bhalodia, V. and Vasconcelos, N. 2010. Anomaly Detection in Crowded Scenes. In IEEE Conference on Computer Vision and Pattern Recognition.

[5] Ni, B., Yan, S. and Kassim, A. 2009. Recognizing Human Group Activities with Localized Causalities. In IEEE Conference on Computer Vision and Pattern Recognition.

[6] Zhu, G., Yan, S., Han, T.X. and Xu, C. 2011. Generative group activity analysis with quaternion descriptor. In International Conference on Advances in Multimedia Modeling.

[7] Lan, T., Wang, Y., Yang, W. and Mori, G. 2010. Beyond Actions: Discriminative Models for Contextual Group Activities. In Annual Conference on Neural Information Processing Systems.

[8] Cheng, Z., Qin, L., Huang, Q., Jiang, S. and Tian, Q. 2010. Group Activity Recognition by Gaussian Process Estimation. In International Conference on Pattern Recognition.

[9] Rasmussen, C. E. and Williams, C. K. I.2006. Gaussian Processes for Machine Learning. The MIT Press.

[10] Dalal, N. and Triggs, B. 2005. Histograms of oriented gradients for human detection. In IEEE Conference on Computer Vision and Pattern Recognition.