Unsupervised Group Activity Detection by Hierarchical Dirichlet Processes

Ali Al-Raziqi, Joachim Denzler

Computer Vision Group, Friedrich-Schiller-Universität Jena, Germany
{ali.al-raziqi,Joachim.Denzler }@uni-jena.de

Abstract. Detecting groups plays an important role for group activity detection. In this paper, we propose an automatic group activity detection by segmenting the video sequences automatically into dynamic clips. As the first step, groups are detected by adopting a bottom-up hierarchical clustering, where the number of groups is not provided beforehand. Then, groups are tracked over time to generate consistent trajectories. Furthermore, the Granger causality is used to compute the mutual effect between objects based on motion and appearances features. Finally, the Hierarchical Dirichlet Process is used to cluster the groups. Our approach not only detects the activity among the objects of a particular group (intra-group) but also extracts the activities among multiple groups (inter-group). The experiments on public datasets demonstrate the effectiveness of the proposed method. Although our approach is completely unsupervised, we achieved results with a clustering accuracy of up to 79.35% and up to 81.94% on the Behave and the NUS-HGA datasets.

1 Introduction

Public spaces are characterized by the existence of several activities. Many researchers have contributed in activity recognition. The approaches can be divided into three categories: (I) Action recognition, which is handled by analyzing the action of a single object through extracting features of the whole object or the segmented body parts [1]. (II) Pair activity, which is interpreted by analyzing the relationships of a pair of objects [8, 11]. (III) Group activity, which is considered as coherent activities performed by multiple objects. In this paper, we focus on the group activity recognition. The analysis of group activity plays an important role in video analysis. Accordingly, localizing and understanding the group activity is an important topic in many applications such as security and surveillance interaction detection. In addition, it can help in detecting suspicious and illegal group behavior.

Most of the group activity recognition methods are supervised [4, 13, 14, 16, 18–20]. In contrast, we focus on detecting the group activity in an unsupervised manner using *Hierarchical Dirichlet Processes* (HDP). Although recent work applied HDP in interaction detection [2], they rely on optical flow features, which is not helpful in case of the fixed objects. The intuition behind their work is to segment activities into spatio-temporal patterns. Also, a video sequence was divided temporally into equally sized clips without overlap. As a consequence, too short clips will split up an activity into sub-activities, and thus too long clips might join non-relevant activities.

We tackle this problem by dividing a video automatically into clips using an unsupervised clustering approach. As a result, the clips might have overlap and have different lengths. To this end, first, relevant groups of objects are detected using a *bottom-up hierarchical clustering*. The groups are tracked over time to form consistent trajectories, then, each group is treated as one clip. Finally, the HDP is used to cluster the clips.

The main contributions of this paper are as follows: (I) We presented a novel approach for detecting meaningful groups without training. This addresses (1) a varying number of involved objects and (2) an unknown number of groups. (II) The Granger causality is used to measure the mutual effect among objects in a particular group and among groups as well based on motion trajectories and appearances features.

The rest of this paper is organized as follows. Sect. 2 provides an overview of the existing literature on group activity recognition. The proposed framework of group activity is described in Sect. 3. The experiments and results conducted on the Behave and NUS-HGA datasets are described in Sect. 4 along with results.

2 Related Work

Ni *et al.* in [14], analyzed the self, pair, and inter-group causalities to detect group activities based on trajectories. They assumed that there is only one group activity in the scene. Hence, they cannot handle more complicated environments of simultaneous activities. In contrast, we handle all the activities in the scene. Zhang *et al.* in [19] tried to detect multiple group activities. This was achieved by clustering the objects into subgroups using K-means. But, providing the number of groups in advance is not a robust solution. An interesting work has been presented in [13], Kim *et al.* tried to overcome

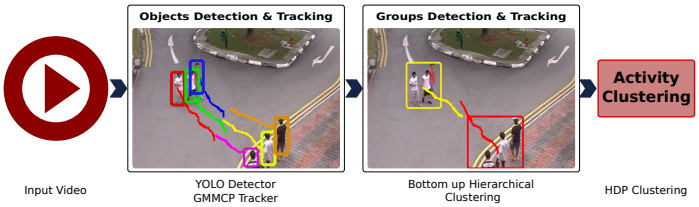

Fig. 1. Proposed framework for automatic group activity detection.

the fixed number of groups. They recognized the groups by modeling proxemics. This was achieved by defining Interaction Potential Zone (IPZ) around each object (bubbles with 58 pixels). However, using the same IPZ value whether objects are far or near from the camera leads to dispersing the relevant objects in irrelevant groups and vice versa. In contrast to them, we cluster the objects using bottom-up hierarchical clustering based on the velocity and motion direction.

Additionally, deep neural networks have been recently applied for group activity detection [6, 7, 10]. Thus, they are more robust and effective, but they used supervised learning methods. In work presented in [2], spatio-temporal patterns are analyzed automatically by using HDP to extract the hidden topics. They divided the video into short and equally sized clips without overlap, where the clip size effects on the performance. Another interesting method which tried to tackle this problem using an extended probabilistic Latent Semantic Analysis (pLSA) [21]. Unlike them, our approach extracts the number of activities automatically. Unlike many of the approaches described above, our approach extracts the group activity by dividing a video automatically into clips without further knowledge.

3 Framework

Our framework for group activity detection has several stages as shown in Figure 1. Given an input video, objects are detected using YOLO [15], which is a unified neural network based approach. All detected objects are tracked by the GMMCP tracker [5]. After generating the groups of objects, each group is tracked over time and treated as one clip. Afterward, these clips are clustered by the HDP as shown in Figure 2. The last two steps will be described in detail as follows.

3.1 Groups Detection and Tracking

In common situations, multiple objects are involved simultaneously in separate activities, and those objects may further interact with each other. To detect all activities in the scene, the main step is the detection of groups. The key assumption of our approach is to cluster objects that are spatially close and moving in the same direction with the same speed. Given objects trajectories, each object represented by 3 tuples $(\mathcal{P}, \mathcal{V}, \theta)$, where $\mathcal{P}$ is the center of mass coordinate (x, y), $\mathcal{V}$ represents the velocity and θ is the motion direction. However, the pairwise distance $d^t(i, j)$ is computed for the trajectories i, j as

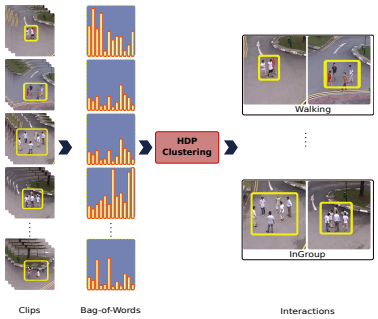

Fig. 2. Process of extracting BoWs and clips clustering.

$$\begin{aligned} d^t(i,j) &= \xi^t_{i,j} \cdot \vartheta_t(i,j) \\ \xi^t_{i,j} &= |\mathcal{P}^t_i - \mathcal{P}^t_j| \end{aligned} \tag{1}$$

$$\vartheta_t(i,j) = \begin{cases} 1 & \text{if} \quad |\mathcal{V}^t_i - \mathcal{V}^t_j| < \mathcal{T}_v \quad \wedge \quad |\theta^t_i - \theta^t_j| \quad < \mathcal{T}_\theta \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

Consequently, we compute d^t by multiplying ξ^t and ϑ^t and normalized, which ensures that the objects are spatially close and moving in the same direction with the same speed. Where $\mathcal{T}_v$ and $\mathcal{T}_\theta$ are predefined thresholds. Then, the adjacency matrix $\mathcal{A}$ is built as a result of $1-d^t$ for each pair of detections. Equation 3 shows an example of 4 objects. In matrix $\mathcal{A}$, a large value means that the two objects are most close and they are moving with the same speed and in the same direction.

$$\mathcal{A} = \begin{pmatrix} \overline{1-d^t(1,2)} & 1-d^t(1,3) & 1-d^t(1,4) \\ & \overline{1-d^t(2,3)} & 1-d^t(2,4) \\ & & \overline{1-d^t(3,4)} \end{pmatrix} \rightarrow \begin{pmatrix} \overline{0.5} & 0.2 & 0 \\ & \overline{0.3} & 0.4 \\ & & \overline{0.1} \end{pmatrix} \tag{3}$$

After that, the groups are detected using bottom-up hierarchical clustering. In the first step, by taking the matrix $\mathcal{A}$ as input, each object is assigned to a separate cluster (4 clusters in this case) and merged with the most similar clusters in the next iterations. For instance, objects $1,2$ will be assigned to one group because they have the maximum value (0.5). In the next step, 1,2 and 3 will be in one group, etc. In our case, the number of groups is not required comparing to traditional methods (e.g., K-means or spectral).

Once the groups are generated, dense SIFT features are extracted for each group. All features are clustered into k clusters using K-means. Then the Bag of Words (BoWs) histograms are extracted. Hence, each group is described by BoWs. The matching between the groups' BoWs is computed by the distance of the histogram intersection, which is bounded by $[0,1]$. Finally, each tracked group is treated as one clip.

3.2 Activity Clustering

When the clips have been generated, our approach does not only detect the activity among the objects of a particular group (intra-group) but also extracts the activ-

ities among multiple groups (inter-group). Suppose that G_j is a group of size n objects, the group center is determined by the average position of all objects $G_j(c^t, y^t) = \frac{1}{n}\left(\sum_{i=1}^{|n|} x_i, \sum_{i=1}^{|n|} y_i\right)$. In order to describe the activities of intra-group and inter-group, trajectories-based features are extracted for every time window $\mathcal{K}$. Some important features are as follows.

Causality The temporal causality is an usual way to recognize the group activity. In this paper, Granger causality (GC) [9] is used to measure the causal relationships between objects. Generally speaking, given two time series A and B, A is said to be Granger-cause B ($A \rightarrow_G B$) if the past values of A with the past values of B provide significant information of B. Many approaches [13, 14, 19, 21] have focused on measuring GC between objects trajectories in terms of center mass coordinates (x, y). In our approach, we compute the causality of both, the objects coordinates and the appearance features. The appearance SIFT features are extracted for each object, the dictionary is built using K-means, then each object represented by concatenating BoWs over the window $\mathcal{K}$.

To infer the GC, the null hypothesis $A \nrightarrow_G B$ has to be tested first, by evaluating the autoregression as

$$\begin{aligned} B^{(i)} &= \beta_0 + \beta_1 B^{(i-1)} + \cdots + \beta_l B^{(i-l)} \\ \bar{B}^{(i)} &= B^{(i)} + \vartheta_0 + \vartheta_1 A^{(i-1)} + \cdots + \vartheta_l A^{(i-l)}, \end{aligned} \tag{4}$$

where β_l and ϑ_l are the model parameters. Therefore, the residual sum of square errors $RSS_B, RSS_{\bar{B}}$ are used for the evaluation. Finally, the causality calculated by

$$F_{A \rightarrow B} = \frac{(RSS_{\bar{B}} - RSS_B)/l}{RSS_{\bar{B}}/(K - 2l - 1)} \tag{5}$$

where K is the number of samples considered for the analysis and l represents the lag.

Shape Similarity Suppose we have two trajectories $\mathcal{A}$ and $\mathcal{B}$, Dynamic Time Warping (DTW) is used to map one trajectory to another by minimizing the distance between the two. In particular, the sum of the Euclidean distances is used as feature.

Velocity and Distance Features: We extracted some other features like velocity $\mathcal{V}$, absolute change $\hat{\mathcal{V}}$ of the same object and the absolute difference $\mathcal{V}_{[i,j]}$ in velocity of a pair of objects i and j. In addition to that, vorticity $\mathcal{V}$ can be measured as a deviation of the center mass of an object from a line. The line is calculated by fitting a line to the positions of the trajectory in window $\mathcal{K}$. Moreover, computing the distance $d_{[i,j]}$ and the difference in distance $\hat{d}_{[i,j]}$ are useful to distinguish the interaction among objects. Since they are sensitive to tracking errors, the distance $\tilde{d}_{[i,j]}$ is calculated for every point in window $\mathcal{K}$. More information can be found in [3]. Since the number of features varies, encoding those normalized features using K-means is required. The BoWs histograms are extracted, that each activity is represented by BoWs.

Hierarchical Dirichlet Processes Once the BoWs histograms are computed, the HDP is used to extract the activities as shown in Figure 2. HDP is a generative clustering technique used to cluster words in documents into K latent *topics* [17]. In HDP, the number of topics is inferred automatically from the data and the hyper-parameters.

4 Experimental Results

We validate our proposed approach on two benchmark datasets, the Behave and NUS-HGA [3, 14]. Since the NUS-HGA dataset does not contain the tracking ground truth, the objects are detected and tracked using YOLO detector and GMMCP tracker. For evaluation purposes, the extracted HDP topics are mapped to the ground truth labels by voting among the topics. From the perspective of the HDP theory, each document is a distribution over all extracted topics, so we restricted one topic for one class. Then the evaluation is done as a classification problem. Due to the randomness in the Bayesian inference, each experiment runs ten times, and we report the average performance.

Concerning the parameters analyzing in Sect. 4.2, the parameters of the experiments are chosen for both datasets as follows. For feature extraction, the dictionary size of the SIFT features for computing the Granger causality are 20 and 90 for Behave and NUS- HGA. For the group activity recognition experiments, the dictionary sizes for K-means of the whole features are 50 and 150 for Behave and NUS- HGA. HDP hyper-parameters α and η are set as 0.8 and 0.1 for both dataset.

4.1 Behave and NUS-HGA Datasets

In the Behave dataset, multiple objects ranging from two to five are involved in each activity. We achieved clustering accuracy of up to 79.35($\pm$5%). As can be seen in Figure 3(a), when the dictionary size is increased, the performance is further decreased. The Behave dataset is represented significantly by dictionary size 50. We compare our approach with [2,3,12–14,16,18], according to Table 1, we outperformed all unsupervised approaches and the one that presented in [13] on the Behave dataset.

Table 1. Comparison with other works on the Behave and NUS-HGA datasets

	Method	Accuracy %	
		Behave Dataset	NUS-HGA Dataset
Supervised	[4]	42.50	93.50
	[13]	**93.74**	96.02
	[14]	-	74.16
	[16]	-	**98.00**
	[18]	93.65	-
	[3]	93.67	-
Unsupervised	[2]	65.95	-
	[12]	66.25	-
	Our Approach	**79.35**	81.94

NUS-HGA dataset has different group activities, *WalkInGroup, Gather, RunInGroup, Fight, StandTalk, and Ignore*. We achieved clustering accuracy of up to 81.94 ($\pm$3.07%). As can be seen in Figure 3(a), as the dictionary size is increased, the performance is further decreased. The performance is compared with [4, 13, 14, 16] as shown in Table 1. Despite the fact that our approach is unsupervised, we outperformed the supervised work that is presented in [14].

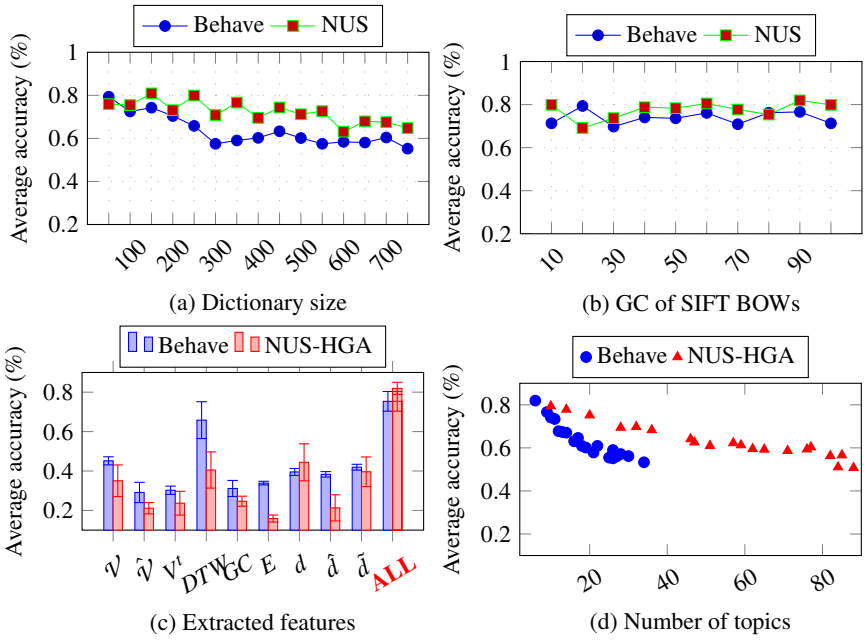

Fig. 3. Illustration of impact of the dictionary size, the GC of SIFT BOWs, the extracted features and HDP η hyperparameter on the Behave and NUS-HGA datasets.

4.2 Influence of Parameters

We studied the influence of the GC, the extracted features and HDP parameters.
Causality: As can be seen in Figure 3(b), we achieved the highest accuracy with the Granger causality of SIFT BoW as 20 and 90 combined with the other features on both datasets respectively. This shows that the Granger causality of SIFT BoWs is robust even with high dimensional data.
Features: As can be seen from Figure 3(c), the most represented feature is the shape similarity DTW of the trajectories for both datasets. The combination of all features improves the performance significantly.
HDP Hyper-Parameters: In these experiments, η ranges from 0.1 to 2. With increasing the η, the number of extracted topics increases linearly, as can be seen in Figure 3(d). From the perspective of the HDP theory, an infinite number of topics are extracted. Therefore, a too small number of topics would under-represent the group activities, which causes joining of similar activities into one. On the other hand, a too large number of topics would lead to over-fitting.

5 Conclusion

The aim of this paper was to address the problem of group activity detection in an unsupervised manner. We introduced a new approach to segment video sequences automatically into clips based on the occurring activities. The main step was the detection of

groups using the bottom-up hierarchical clustering. Furthermore, the Granger causality is used to measure the mutual effect among objects in a particular group and among groups as well based on motion trajectories and appearances features. Finally, the activities are extracted by using HDP. We achieved results with a clustering accuracy of up to 79.35% on the Behave dataset and up to 81.94% on the NUS-HGA dataset.

Acknowledgements The authors are thankful to Mahesh Krishna and Manuel Amthor for useful discussions and suggestions.

References

1. Jake K Aggarwal and Lu Xia. Human activity recognition from 3d data: A review. *Pattern Recognition Letters*, 48:70–80, 2014.
2. Ali Al-Raziqi and Joachim Denzler. Unsupervised framework for interactions modeling between multiple objects. In *VISAPP*, volume 4, pages 509–516, 2016.
3. Scott Blunsden and RB Fisher. The behave video dataset: ground truthed video for multi-person behavior classification. *BMVA*, 4:1–12, 2010.
4. Zhongwei Cheng, Lei Qin, Shuicheng Huang, Yan, and Tian. Recognizing human group action by layered model with multiple cues. *Neurocomputing*, 136:124–135, 2014.
5. A. Dehghan, S. Assari, and M. Shah. Gmmcp tracker: Globally optimal generalized maximum multi clique problem for multiple object tracking. In *CVPR*, pages 4091–4099, 2015.
6. Arash and-Hexiang Deng, Zhiwei and Greg Mori. Structure inference machines: Recurrent neural networks for analyzing relations in group activity recognition. In *CVPR*, 2016.
7. Z Deng, M Zhai, Yuhao Chen, L, S Muralidharan, M Roshtkhari, and G Mori. Deep structured models for group activity recognition. *arXiv preprint arXiv:1506.04191*, 2015.
8. Zhen Dong, Yu Kong, Cuiwei Liu, Hongdong Li, and Yunde Jia. Recognizing human interaction by multiple features. In *ACPR*, pages 77–81, 2011.
9. Clive WJ Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society*, pages 424–438, 1969.
10. Mostafa S. Ibrahim, Srikanth Muralidharan, Zhiwei Deng, Arash Vahdat, and Greg Mori. A hierarchical deep temporal model for group activity recognition. In *CVPR*, June 2016.
11. Binlong Li, Mustafa Ayazoglu, Teresa Mao, Octavia Camps, Mario Sznaier, et al. Activity recognition using dynamic subspace angles. In *CVPR*, pages 3193–3200, 2011.
12. David Münch, Eckart Michaelsen, and Michael Arens. Supporting fuzzy metric temporal logic based situation recognition by mean shift clustering. In *AI*, pages 233–236. 2012.
13. U. Park J.-S. Park N.-G. Cho, Y.-J. Kim and S.-W. Lee. Group activity recognition with group interaction zone based on relative distance between human objects. *Pattern Recognition and Artificial Intelligence*, 29(05), 2015.
14. Bingbing Ni, Shuicheng Yan, and Ashraf Kassim. Recognizing human group activities with localized causalities. In *CVPR*, pages 1470–1477, 2009.
15. Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. *arXiv preprint arXiv:1506.02640*, 2015.
16. Kyle Stephens and Adrian G Bors. Group activity recognition on outdoor scenes. In *Advanced Video and Signal Based Surveillance (AVSS)*, pages 59–65, 2016.
17. YW Teh, MI Jordan, MJ Beal, and DM Blei. Hierarchical dirichlet processes. *JASA*, 101(476), 2006.
18. Yafeng Yin, Guang Yang, Jin Xu, and Hong Man. Small group human activity recognition. In *Image Processing (ICIP)*, pages 2709–2712. IEEE, 2012.

19. Cong Zhang, Xiaokang Yang, Weiyao Lin, and Jun Zhu. Recognizing human group behaviors with multi-group causalities. In *WI-IAT*, volume 3, pages 44–48, 2012.
20. Yue Zhou, Bingbing Ni, Shuicheng Yan, and Thomas S Huang. Recognizing pair-activities by causality analysis. *ACM - TIST*, 2(1):5, 2011.
21. Guangyu Zhu, Shuicheng Yan, Tony X Han, and Changsheng Xu. Generative group activity analysis with quaternion descriptor. In *Multimedia Modeling*, pages 1–11. Springer, 2011.