International Journal of Pattern Recognition
and Artificial Intelligence
Vol. 29, No. 5 (2015) 1555007 (15 pages)
© World Scientific Publishing Company
DOI: 10.1142/S0218001415550071

World Scientific
www.worldscientific.com

Group Activity Recognition with Group Interaction Zone Based on Relative Distance Between Human Objects

Nam-Gyu Cho[*,¶], Young-Ji Kim[†,‖], Unsang Park[‡,**], Jeong-Seon Park[§,††]
and Seong-Whan Lee[*,†,‡‡]

[*]*Department of Brain and Cognitive Engineering*
Korea University, Anam-dong 5-ga, Seongbuk-gu Seoul, 136-713, Korea

[†]*Department of Computer Science and Engineering*
Korea University, Anam-dong 5-ga, Seongbuk-gu Seoul, 136-713, Korea

[‡]*Department of Computer Science and Engineering*
Sogang University, 5 Baekbeom-ro (Sinsu-dong)
Mapo-gu, Seoul, 121-742, Korea

[§]*Department of Multimedia*
Chonnam National University, Yeosu
Jeollanam-do 550-749, Korea

[¶]*ngcho@image.korea.ac.kr*
[‖]*yjkim@image.korea.ac.kr*
[**]*unsangpark@sogang.ac.kr*
[††]*jpark@chonnam.ac.kr*
[‡‡]*sw.lee@korea.ac.kr*

Received 26 November 2014
Accepted 27 February 2015
Published 21 May 2015

In this paper, we address the problem of recognizing group activities of human objects based on their motion trajectory analysis. In order to resolve the complexity and ambiguity problems caused by a large number of human objects, we propose a Group Interaction Zone (GIZ) to detect meaningful groups in a scene to effectively handle noisy information. Two novel features, Group Interaction Energy (GIE) feature and Attraction and Repulsion Features, are proposed to better describe group activities within a GIZ. We demonstrate the performance of our method in two ways by (i) comparing the performance of the proposed method with the previous methods and (ii) analyzing the influence of the proposed features and GIZ-based meaningful group detection on group activity recognition using public datasets.

Keywords: Human group activity recognition; visual surveillance; machine vision; pattern recognition.

[‡‡] Corresponding author.

1. Introduction

Human activity recognition is one of the most important problems in computer vision and has many practical applications such as human–computer interaction and video surveillance.[19] Researchers have conducted research in roughly four sub-problems: (i) individual action recognition, analyzing the behavior of a single person either by decoding changes of local or global movements[13,30] or decomposing semantic body parts' movements,[3,6] (ii) interaction recognition, interpreting two people's movements,[21,26,29] (iii) crowd activity recognition, mainly analyzing trajectory (or flow) information of a large number of people,[22,28] and (iv) group activity recognition, analyzing interaction more than two people and less than crowd, which was first addressed by Ni *et al.*,[24] then further studied by many researchers.[7,16,17,36]

Particularly, recognizing a group activity is a challenging and important problem — not only because of its technical difficulties, but also because of its increasing requirements in practical applications such as public security. In this paper, we focus on the group activity recognition problem. In general, a group activity consists of multiple individual activities. For example, an "approaching" activity consists of multiple individual's "walking" activities. Therefore, in order to recognize group activities, both local (individual) and global (group) information need to be considered together.

Previous work on group activity recognition can be categorized into two groups: image feature-based and trajectory-based approaches. The image feature-based approaches[1,8,26] describe an activity as a collection of motion gradient features[18,33] and their statistics or sparse representation.[31] Therefore, sometimes a few dominant features represent each group activity. However, because of strong dependency on feature extraction, they are vulnerable to situations where feature extraction fails, mainly due to occlusion or low resolution. Meanwhile, trajectory-based approaches[4,20,24,27,35] focus on analyzing human activities in terms of interactions between individual trajectories. Thus, they are more robust to occlusions or low resolution. Zhou *et al.*[35] analyzed interactions between two individuals using the Granger Causality Test (GCT).[12] However, due to the limitations of GCT, they only focused on the pair-activity recognition problem. In order to deal with more complex situations, Ni *et al.*[24] analyzed self, pair, and group causalities using local trajectory information. However, they assumed only one group in a scene. Therefore, these methods cannot be generalized for more complex situations such as a group of people participating in an activity where other individuals pass by, e.g. the BEHAVE dataset.[2] In order to cope with this situation, Yin *et al.*[32] first clustered individuals into several sub-groups by minimum spanning tree algorithm, and then constructed a network and extracted a histogram feature. Zhang *et al.*[34] found sub-groups with K-means algorithm, and then regarded group activities as a combination of sub-groups by characterizing four types of causalities: individual, pair, behavior, and inter-group. Although finding a sub-group is helpful, how to detect a sub-group remains a challenging problem. In particular, assuming a fixed number of groups in a scene is not a robust solution.

Proxemics theory[14] has been recently used to fit realistic social relationship into data in various fields such as architecture and psychology. It has been also used to analyze human interaction in surveillance environments. Cristani *et al.*[9] and Rota *et al.*[25] used proxemics to infer social relation as F-formation in generated human relationships. They analyzed the interaction of people considering the configuration of objects around the bubble zone of proxemics in every time step. Gan *et al.*[11] proposed an extended F-formation based on a heat map[21] to take temporal information into consideration to find groups suitable for a photo shot. In a similar manner, Cui *et al.*[10] proposed Interaction Energy Potential to detect normal and abnormal events based on the Social Force Model[15] which models pedestrian's behavior. They assumed that people are aware of other people's positions and velocities and can predict whether they will meet other in the near future. However, they considered only an approaching event despite the fact that people are not only meeting but also departing in the real world.

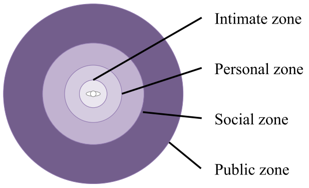
We tackle the problem of recognizing group activity by detecting meaningful groups (e.g. where a few people are actually involved in an activity while the rest are not) and describing it. We argue that detecting and describing a group activity need to be carried out based on a better understanding of human behavior. Our contributions are twofold: (i) We propose a novel meaningful group detection method by modeling proxemics. Based on this, we define a Group Interaction Zone (GIZ), and then detect and update it in a scene so that we can suppress noisy information incurred by human objects not involved in the target activity. (ii) We optimally describe a group activity in a GIZ using attraction and repulsion properties inspired by Sethi and Roy-Chowdhury[27] which considered an interaction in terms of "getting close", "away", and "keeping the same distance together". As a part of the proposed method, we also propose two novel features, Group Interaction Energy (GIE) feature and Attraction and Repulsion Features (ARF).

The rest of this paper is organized as follows. In Sec. 2, we introduce our proposed methods, and Sec. 3 demonstrates experimental results. In Sec. 4, we conclude the paper and discuss future work.

2. Method for Group Activity Recognition

2.1. *GIZ detection*

We first introduce a GIZ between human objects by modeling proxemics. Proxemics is defined as "*the interrelated observations and theories of man's use of space as a specialized elaboration of culture*".[14] According to this, an area around a person can be divided into four categories: intimate, personal, social, and public (Fig. 1). In this paper, we assume that an interaction between people will occur within a certain distance. Thus, we define an Interaction Potential Zone (IPZ) according to the proxemic's personal distance to represent a possible range of an interaction. An IPZ is a basic unit for detecting a GIZ, as described by the following four steps (Fig. 2).



Category	Distance (m)	Possible interaction
Intimate zone	0 – 0.4	Embracing, touching, or whispering
Personal zone	0.4 – 1	Interactions among good friends or family members
Social zone	1.2 – 3.7	Interactions among acquaintances (touching body does not happen)
Public zone	Above 3.7	Public speaking

Fig. 1. Interpersonal zones based on proxemics.[14]

First, we draw an IPZ around each human object (Fig. 2(a)). Second, we calculate the overlapping area between IPZs — the larger the overlapping region, the more likely a group activity occurs (Fig. 2(b)). Third, we compute the ratio of the overlapping area to total area covered by interacting human objects as,

$$\gamma^t = \frac{\bigcap_{i=1}^{N_C} \Omega(\mathbf{x}_i^t)}{\bigcup_{i=1}^{N_C} \Omega(\mathbf{x}_i^t)}, \tag{1}$$

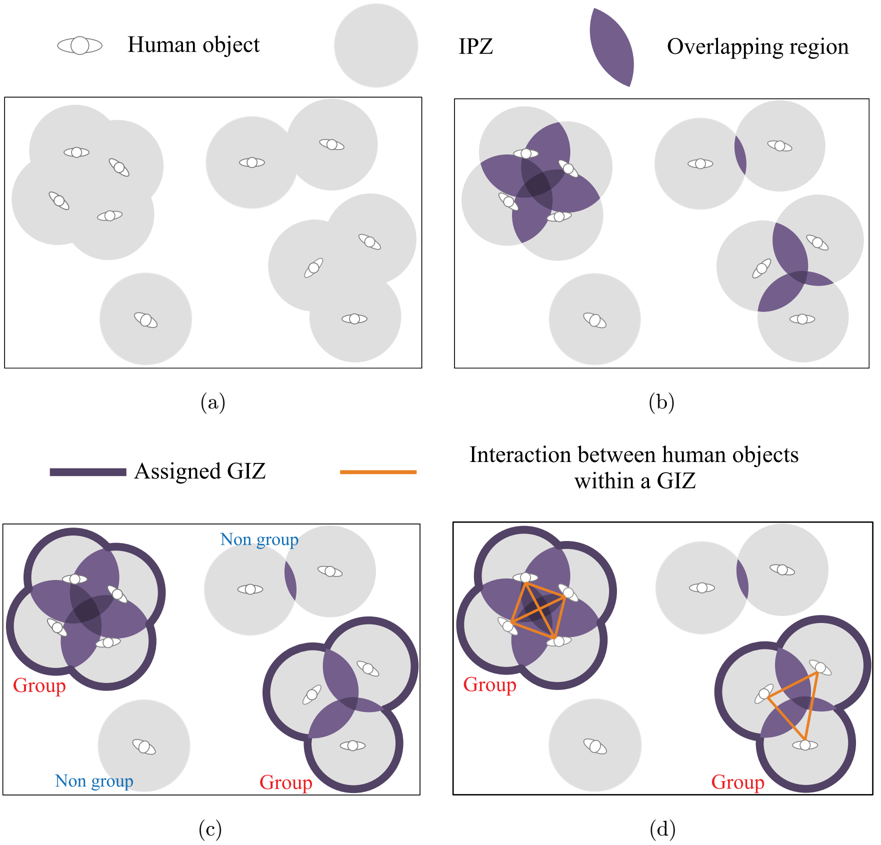


Fig. 2. Constructing a GIZ. (a) Drawing IPZs, (b) calculating overlapping area, (c) assigning GIZs and (d) calculating features within a GIZ.

where $\Omega(\mathbf{x}_i^t)$ represents an IPZ of the ith human object, $\mathbf{x}_i^t$ is a tuple of an image coordinate (x, y) at time step t, and N_C is the number of people having overlapping IPZs. Then, we assign a GIZ ID to a set of human objects if $\gamma^t \geq \tau_{\mathrm{GIZ}}$ where τ_{GIZ} is a threshold (Fig. 2(c)) that controls the likelihood of a set of human objects fall into the same GIZ. Finally, the interaction features between every possible pairs within a GIZ are calculated (Fig. 2(d)).

Our method is different from previous group detection approaches. Extended F-formation method,[11] closely related to ours, models personal interaction area as a fan shape with regard to the gaze direction to find group formations such as staying in a circular (facing each other) or side by side position. Thus, it is suitable for static situations such as "stand talking" or "gathering". In the meanwhile, our method connects people loosely with a circular interaction area — it is not restricted for a specific viewing direction. Thus, it is suitable for dynamic situations such as "fighting" where people alternate between facing and passing other people frequently. Our method is also efficient in terms of determining the number of existing groups automatically while clustering-based methods[32,34] assume fixed number.

2.2. *Extracting interaction features from GIZ*

Attraction and repulsion properties are considered for feature extraction. The attraction property captures the tendency of people to get close to each other whereas the repulsion property captures the opposite case. These properties are closely connected with relative distance changes. Thus, we first consider a subset of trajectory information as follows,

$$\xi_i^{T,k} = [\mathbf{x}_i^{T-(k-1)}, \mathbf{x}_i^{T-(k-2)}, \ldots, \mathbf{x}_i^{T}], \tag{2}$$

where $\xi_i^{T,k}$ is a variable that consists of a subset of the object i's trajectory information during k time steps. Then we calculate the relative distance between objects i and j as,

$$\alpha_{ij}^{T,k} = \mathbf{d}(\xi_i^{T,k}, \xi_j^{T,k}), \tag{3}$$

where $\mathbf{d}(,)$ returns a distance vector. The mean and variance of $\alpha_{ij}^{T,k}$ are denoted as $\hat{\alpha}_{ij}^{T,k}$ and $\tilde{\alpha}_{ij}^{T,k}$. To model a GIE, we extend the Interaction Energy Potential by Cui *et al.*[10] that models the likelihood of people's encounters in a near future (i.e. the attraction property). However, inspired by Sethi and Roy-Chowdhury[27] (please refer to Sec. 1), we consider both attraction and repulsion properties together in our GIE. We define a state variable ω to reflect the influence of past trajectories as follows,

$$\omega = \begin{cases} 1 & \text{when } \hat{\alpha}_{ij}^{T,k} > \alpha_{ij}^{T,0}, \\ -1 & \text{when } \hat{\alpha}_{ij}^{T,k} < \alpha_{ij}^{T,0}, \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

Thus, ω represents the attraction (1) and repulsion (-1). We then construct a GIE function to model these properties with their strengths as follows,

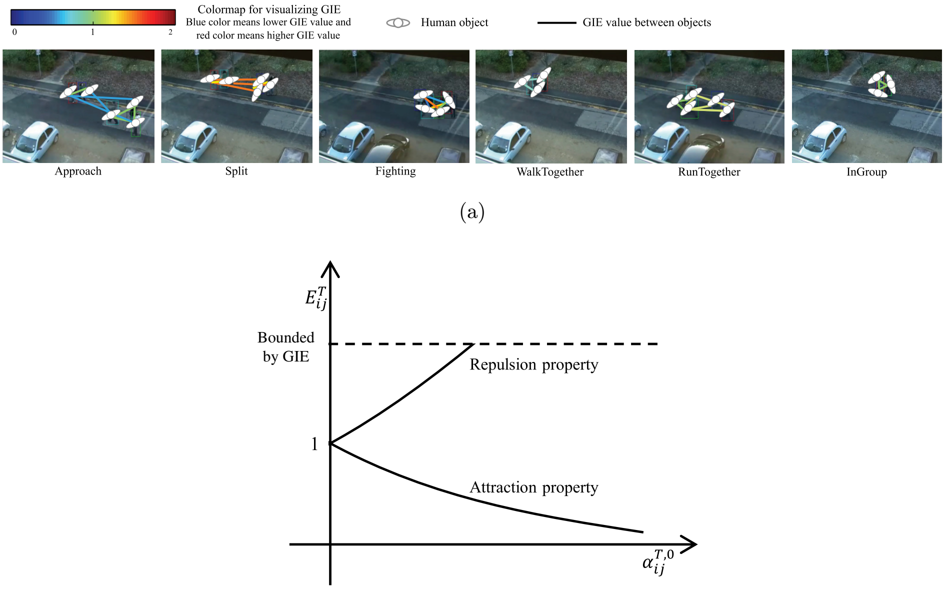
$$E_{ij}^{T} = \exp\left(-\omega\frac{(\alpha_{ij}^{T,0})^2}{\sigma^2}\right), \tag{5}$$

where σ is an affection distance between objects. We set this value same as the distance of IPZ. The value of E_{ij}^{T} is correlated with the current relative distance $\alpha_{ij}^{T,0}$. The response of GIE feature is visualized in Fig. 3. To calculate ARF, we first calculate relative distance changes as,

$$\begin{aligned} \delta_{ij}^{T,k} &= \alpha_{ij}^{T,0} - \alpha_{ij}^{T-(k-1),0}, \\ \hat{\delta}_{ij}^{T,k} &= \hat{\alpha}_{ij}^{T,k} - \alpha_{ij}^{T-(k-1),0}, \end{aligned} \tag{6}$$

where $\delta_{ij}^{T,k}$ represents the two properties with their magnitudes and signs, and $\hat{\delta}_{ij}^{T,k}$ is used to handle outliers. We calculate two summary values of the dynamics as,

$$\begin{aligned} \psi^{+} &= \sum_{t} I^{+}(\delta_{ij}^{T-(k-t),1}), \\ \psi^{-} &= \sum_{t} I^{-}(\delta_{ij}^{T-(k-t),1}), \end{aligned} \tag{7}$$



(a)

(b)

Fig. 3. (a) Visualization of GIE to real data. GIE responds differently to different activities. For example, it has lower value to Approach where people get close to each other (attraction property). It has higher value to Split where people get far away from each other (repulsion property). For Fighting, it shows two different responses. Thus, with ARF, we can distinguish these different activities and (b) GIE's response to $\alpha_{ij}^{T,0}$ (color online).

where $I^+(n)$ and $I^-(n)$ are the indicator functions that return 1 when the value n is greater than 0 and vice versa. With two additional variables $\nu_{ij}^{T,k}$ and $\phi_{ij}^{T,k}$, the magnitude and orientation of mean velocity during k time steps, we get a seven-dimensional feature for ARF as,

$$\lambda_{ij}^{T,k} = \left[\delta_{ij}^{T,k}, \hat{\delta}_{ij}^{T,k}, \nu_{ij}^{T,k}, \phi_{ij}^{T,k}, \psi^+, \psi^-, \sum_t \delta_{ij}^{T-(k-t),1}\right]. \tag{8}$$

2.3. *Causality feature*

We use GCT[35] feature to represent causality between human objects in temporal space. The process of getting the causality ratio and feedback ratio features is as follows.

Given two stationary motion trajectories, whose overall shape does not change within a short time period, M_t and N_t, two prediction functions are modeled as $P(m_t|M_{t-l})$ and $P(m_t \mid M_{t-l}, N_{t-l})$. We use kth order linear predictor,

$$\begin{aligned} m_t &= \sum_{s=0}^{k-1} \alpha_s m_{t-l-s} + \varepsilon, \\ m_t &= \sum_{s=0}^{k-1} \beta_s m_{t-l-s} + \sum_{s=0}^{k-1} \gamma_s n_{t-l-s} + \zeta, \end{aligned} \tag{9}$$

where α_s, β_s, and γ_s are the regression coefficients, ε and ζ are the Gaussian noise, k is the number of sample, and l is the non-negative time lag to avoid overfitting. We used Least Square Error (LSE) method for the parameter estimation. The prediction error is assumed to be of Gaussian noise with standard deviation of $\varphi(m_t \mid M_{t-l})$ and $\varphi(m_t \mid M_{t-l}, N_{t-l})$. Now we calculate the Causality Ratio and Feedback Ratio by getting ratios of the prediction errors as

$$\begin{aligned} \text{Causality Ratio} : g_c &= \frac{\varphi(m_t \mid M_{t-l})}{\varphi(m_t \mid M_{t-l}, N_{t-l})}, \\ \text{Feedback Ratio} : g_f &= \frac{\varphi(n_t \mid N_{t-l})}{\varphi(n_t \mid M_{t-l}, N_{t-l})}. \end{aligned} \tag{10}$$

In addition to the proposed features, we use Additional Features (AF): The mean and variance of the magnitudes and orientations of absolute and relative velocity, $|\max(\alpha_{ij}^{T,k}) - \min(\alpha_{ij}^{T,k})|$, $\hat{\alpha}_{ij}^{T,k}$, and $\tilde{\alpha}_{ij}^{T,k}$. Finally, we get a 25-dimensional feature. We calculate an average of the features from every existing pair within a GIZ.

In order to describe a group activity, we first accumulate the extracted features within a time window of size ρ. We then learn a bag-of-words model for each group activity class by clustering features with the K-means algorithm. Finally, we train classifiers using linear SVM in "one versus all others" manner.

3. Experimental Results and Analysis

In this section, we demonstrate the performance of our method on two benchmark datasets: the BEHAVE[2] and NUS-HGA.[24]

3.1. *BEHAVE dataset*

This dataset provides 10 groups activity classes: InGroup, Approach, WalkTogether, Split, Ignore, Following, Chase, Fight, RunTogether, and Meet. The groundtruth is provided for five sequences and each of them includes several classes. Zhang *et al.*,[34] Munch *et al.*,[23] and Yin *et al.*[32] used a subset of classes to demonstrate their method. We first compare the performance on the same subset of classes, then evaluate our method on six group activity classes (Approach (A), Split (S), WalkTogether (W), RunTogether (R), Fighting (F), and InGroup (I)). The rest of the classes are excluded because they do not include group activities or only contain few short sequences. We use the trajectory information provided by the dataset.

Parameters for the implementation are chosen as follows. The personal distance for an IPZ and σ is 58 pixels, and the threshold τ_{GIZ} is 0.1. For feature extraction, the time interval k is 13. For group activity description, the window size ρ is set as 3 frames and the cluster size for K-means algorithm is 100. We evaluated our method via three-folds cross-validation process due to the small number of instances for each class (see Sec. 3.4 for detailed discussion).

We first compare our method with Refs. 23 and 34 (Table 1). Zhang *et al.*[34] divided Approach and Split into two sub-classes such as ApproachOne and ApproachBoth — we used the “Both” in this comparison. Our method have achieved better performance than those of Refs. 23 and 34 for all the classes. For the comparison with Ref. 32, we considered Split, WalkTogether, Fighting, and InGroup activities (Table 2). Our method showed better performances for Split and InGroup activities and slightly lower for WalkTogether and Fighting. However, the overall performance of our method was still the best.

3.2. *Influence of proposed features and GIZ*

In this subsection, we analyze the influence of proposed features and GIZ used in our method on six activity classes of the BEHAVE dataset. First we compared four types

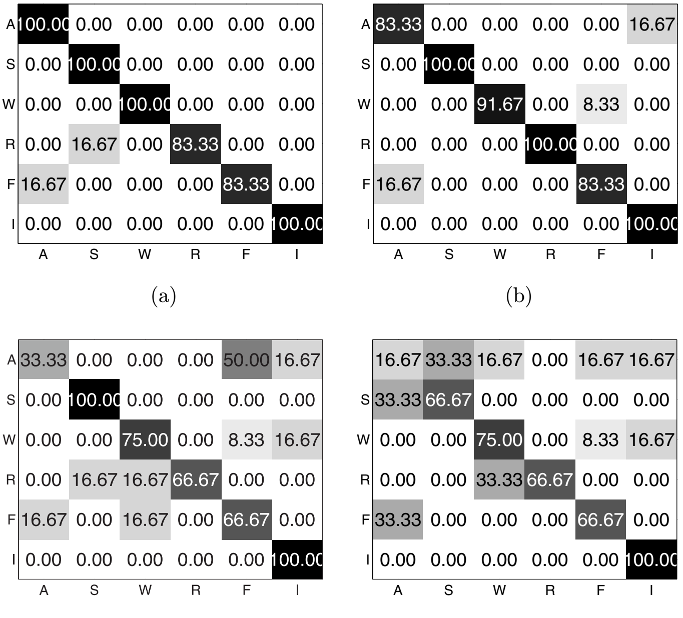
Table 1. Performance comparison with Refs. 23 and 34.

	Ours	Ref. 34	Ref. 23
Approach	83.33	71	60
Split	100	79	70
WalkTogether	91.66	88	45
InGroup	100	88	90
Average	**93.74**	81.5	66.25

Table 2. Performance comparison with Ref. 32.

	Ours	Ref. 32
Split	100	93.1
WalkTogether	91.66	92.1
Fighting	83.33	95.1
InGroup	100	94.3
Average	**93.74**	93.65

of features, GIE, ARF, GCT, and AF with four combinations (Fig. 4): (a) GIE + ARF + GCT + AF (25-dim.), (b) ARF + GCT + AF (24-dim.), (c) GIE + GCT + AF (18-dim.), and (d) GCT + AF (17-dim.). Please note that the proposed features, GIE and ARF together, significantly improved the performance on classifying confusing classes such as Approach–Split and WalkTogether–RunTogether. Thus, as we argued in Sec. 1, using GIZ combined with proposed features helps to better describe group activities (Figs. 5(a) and 5(b)). We also evaluated the average precision and recall of GIZ on the BEHAVE dataset, and the results are 0.78 and 0.82, respectively. We believe that there still exists a possibility to further improve the performance of GIZ — thus improve the whole system.



	A	S	W	R	F	I
A	100.00	0.00	0.00	0.00	0.00	0.00
S	0.00	100.00	0.00	0.00	0.00	0.00
W	0.00	0.00	100.00	0.00	0.00	0.00
R	0.00	16.67	0.00	83.33	0.00	0.00
F	16.67	0.00	0.00	0.00	83.33	0.00
I	0.00	0.00	0.00	0.00	0.00	100.00

(a)

	A	S	W	R	F	I
A	83.33	0.00	0.00	0.00	0.00	16.67
S	0.00	100.00	0.00	0.00	0.00	0.00
W	0.00	0.00	91.67	0.00	8.33	0.00
R	0.00	0.00	0.00	100.00	0.00	0.00
F	16.67	0.00	0.00	0.00	83.33	0.00
I	0.00	0.00	0.00	0.00	0.00	100.00

(b)

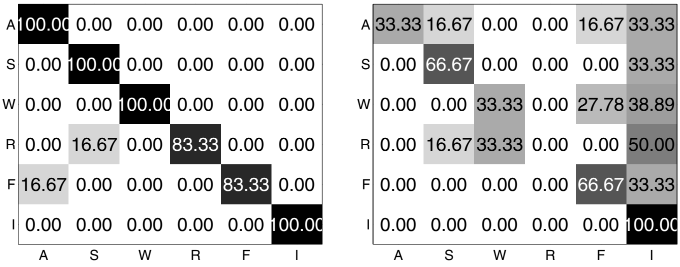
	A	S	W	R	F	I
A	33.33	0.00	0.00	0.00	50.00	16.67
S	0.00	100.00	0.00	0.00	0.00	0.00
W	0.00	0.00	75.00	0.00	8.33	16.67
R	0.00	16.67	16.67	66.67	0.00	0.00
F	16.67	0.00	16.67	0.00	66.67	0.00
I	0.00	0.00	0.00	0.00	0.00	100.00

(c)

	A	S	W	R	F	I
	16.67	33.33	16.67	0.00	16.67	16.67
S	33.33	66.67	0.00	0.00	0.00	0.00
W	0.00	0.00	75.00	0.00	8.33	16.67
R	0.00	0.00	33.33	66.67	0.00	0.00
F	33.33	0.00	0.00	0.00	66.67	0.00
I	0.00	0.00	0.00	0.00	0.00	100.00

(d)

Fig. 4. Confusion matrices for analyzing the influence of features. (a) GIE + ARF + GCT + AF, (b) ARF + GCT + AF, (c) GIE + GCT + AF and (d) GCT + AF.



(a) (b)

Fig. 5. Analyzing the influence of using GIZ. (a) With GIZ and (b) without GIZ.

3.3. *NUS-HGA dataset*

NUS-HGA dataset provides 476 video clips containing six group activity classes: WalkInGroup, Gather, RunInGroup, Fight, StandTalk, and Ignore. Each instance involves 4–8 people. Parameters for the implementation are chosen as follows. The threshold τ_{GIZ} is 0.3. For group activity description, the window size ρ is set as 4 frames and the cluster size for K-means algorithm is 1600. The rest are same as the setting of the BEHAVE dataset (see Sec. 3.4 for detailed discussion). Since this dataset provides enough number of instances, different from the BEHAVE dataset, we randomly divide the dataset into three subsets proportionally: training (40%), validation (20%), and test (40%). The performance reported in Table 3 is obtained as the average of the results of five rounds.

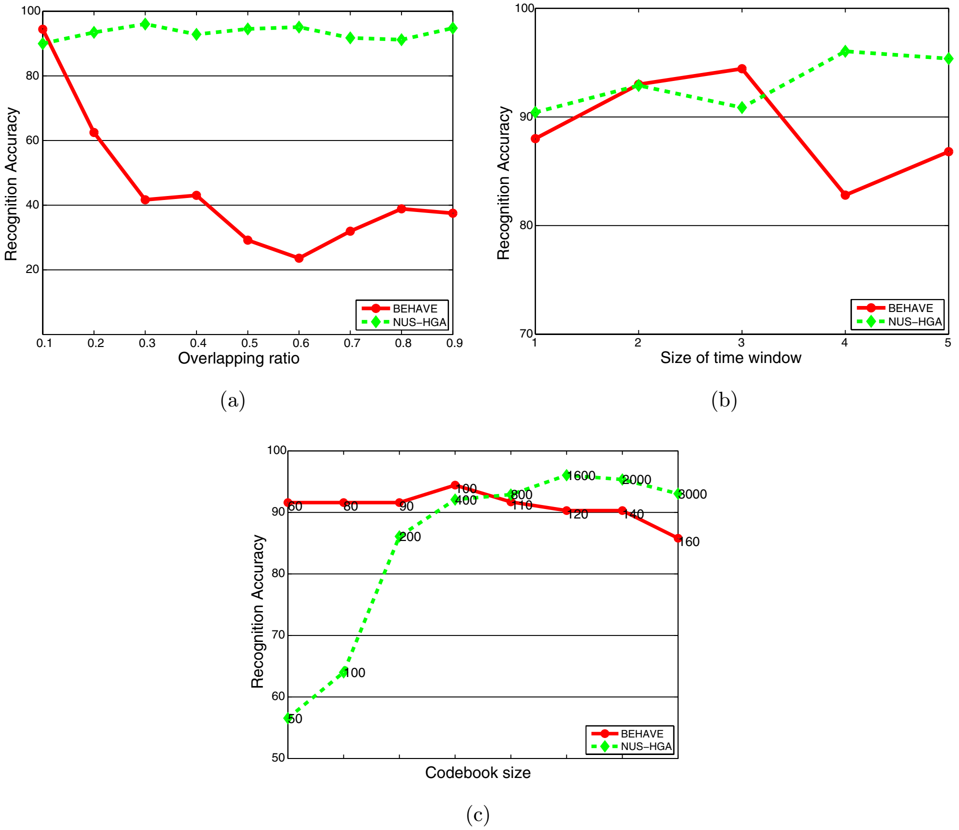
We compared the performance with Refs. 5 and 24, the state-of-the-art methods on this dataset. Please note that Cheng *et al.*[5] demonstrated their performance with several variations. We compare with top two versions, Motion Features (MF) and Motion and Appearance Fusion with adaptive weights (MAF). Our method outperforms MF, which means that the proposed motion features GIE and ARF have discriminative power, and is competitive with MAF.

3.4. *Influence of parameters*

Figure 6 shows the influence of major parameters on the performance: the overlapping ratio (Fig. 6(a)), the size of time window (Fig. 6(b)), and the size of codebook

Table 3. Performance comparison on the NUS-HGA dataset.

	Ours	Ref. 24	Ref. 5 (MF)	Ref. 5 (MAF)
WalkInGroup	96.62	74	—	—
Gather	96.4	40	—	—
RunInGroup	94.8	89	—	—
Fight	97.5	89	—	—
StandTalk	96.11	89	—	—
Ignore	94.78	64	—	—
Average	96.03	74.16	93.2	**96.2**



(a)

(b)

(c)

Fig. 6. Analyzing the influence of major parameters setting. (a) Influence of the overlapping ratio τ_{GIZ}, (b) Influence of the time window size ρ and (c) Influence of the codebook size K (the size of codebook is depicted on each point).

(Fig. 6(c)). The influence of the overlapping ratio τ_{GIZ} is critical on the BEHAVE dataset. Please note that the BEHAVE dataset contains situations we described in Sec. 1 (e.g. where a few people are actually involved in an activity while the rest are not); thus finding sub-groups, which is decided by τ_{GIZ}, is important for this dataset. However, the influence of τ_{GIZ} on the NUS-HGA dataset is not significant since most of instances of the NUS-HGA dataset have a single group. The influence of codebook size K is different on each dataset. This is because of a small codebook is insufficient to represent the diversity of interactions in the NUS-HGA dataset[5] while the BEHAVE dataset can be relatively well represented by a small codebook.

4. Conclusions and Future Work

In this paper, we proposed a novel GIZ detection method based on proxemics to better describe human social behavior to detect meaningful groups. When compared to previous methods, the novelty of our method is to allow people to be connected loosely. Thus, it is more suitable for dynamic group acitivites. Then, we proposed GIE and ARF features for representing group activities in terms of attraction and repulsion properties. We demonstrated our method on two benchmark datasets: the BEHAVE and NUS-HGA. Experimental results showed that our method is more effective for group activity recognition on these datasets compared to other methods. As a future work, we plan to extend our method to handle more complex activities such as bullying and apply to other datasets such as Refs. 7 and 36.

Acknowledgments

The research was supported by the Implementation of Technologies for Identification, Behavior, and Location of Human based on Sensor Network Fusion Program through the Ministry of Trade, Industry and Energy (Grant No. 10041629) and the 2014 R&D Program for S/W Computing Industrial Core Technology through the MSIP (Ministry of Science, ICT and Future Planning)/KEIT (Korea Evaluation Institute of Industrial Technology) (Project No. 14-824-09-005).

References

1. M. R. Amer and S. Todorovic, A chains model for localizing participants of group activities in videos, in *Proc. IEEE Int. Conf. Computer Vision*, Barcelona, Spain (2011), pp. 786–793.
2. S. Blunsden and R. Fisher, The behave video dataset: Ground truthed video for multi-person behavior classification, in *Proc. British Machine Vision Conf.*, Vol. 4, Aberystwyth, UK, August 2010, pp. 1–12.
3. X. Chen and A. L. Yuille, Articulated pose estimation by a graphical model with image dependent pairwise relations, in *Advances in Neural Information Processing Systems* (2014), pp. 1736–1744.
4. Z. Cheng, L. Qin, Q. Huang, S. Jiang and Q. Tian, Group activity recognition by gaussian processes estimation, in *Proc. IEEE Int. Conf. Pattern Recognition*, Istanbul, Turkey (2010), pp. 3228–3231.
5. Z. Cheng, L. Qin, Q. Huang, S. Yan and Q. Tian, Recognizing human group action by layered model with multiple cues, *Neurocomputing* **136** (2014) 124–135.
6. N.-G. Cho, A. L. Yuille and S.-W. Lee, Adaptive occlusion state estimation for human pose tracking under self-occlusions, *Pattern Recogn.* **46**(3) (2013) 649–661.
7. W. Choi and S. Savarese, A unified framework for multi-target tracking and collective activity recognition, in *Proc. European Conf. Computer Vision*, Florence, Italy (2012), pp. 215–230.
8. W. Choi, K. Shahid and S. Savarese, Learning context for collective activity recognition, in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, Colorado Springs, CO, USA, June 2011, pp. 3273–3280.
9. M. Cristani, G. Paggetti, A. Vinciarelli, L. Bazzani, G. Menegaz and V. Murino, Towards computational proxemics: Inferring social relations from interpersonal distances, in *Proc. IEEE Int. Conf. Privacy, Security, Risk and Trust*, Boston, MA, USA (2011), pp. 290–297.

10. X. Cui, Q. Liu, M. Gao and D. N. Metaxas, Abnormal detection using interaction energy potentials, in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, Colorado Springs, CO, USA, June 2011, pp. 3161–3167.
11. T. Gan, Y. Wong, D. Zhang and M. S. Kankanhalli, Temporal encoded f-formation system for social interaction detection, in *Proc. 21st ACM Int. Conf. Multimedia, MM '13* (ACM, New York, NY, USA, 2013), pp. 937–946.
12. C. W. J. Granger, Investigating causal relations by econometric models and cross-spectral methods, *Econometrica* (1969).
13. T. Guha and R. K. Ward, Learning sparse representations for human action recognition, *IEEE Trans. Pattern Anal. Mach. Intell.* **34**(8) (2012) 1576–1588.
14. E. Hall, *The Hidden Dimension* (Anchor Books, 1966).
15. D. Helbing and P. Molnar, Social force model for pedestrian dynamics, *Phys. Rev. E* **51**(5) (1995) 4282–4286.
16. T. Lan, L. Sigal and G. Mori, Social roles in hierarchical models for human activity recognition, in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, Providence, RI, USA, June 2012, pp. 1354–1361.
17. T. Lan, Y. Wang, W. Yang, S. Robinovitch and G. Mori, Discriminative latent models for recognizing contextual group activities, *IEEE Trans. Pattern Anal. Mach. Intell.* **34**(8) (2012) 1549–1562.
18. I. Laptev, M. Marszalek, C. Schmid and B. Rozenfeld, Learning realistic human actions from movies, in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, June 2008, pp. 1–8.
19. T. Li, H. Chang, M. Wang, B. Ni, R. Hong and S. Yan, Crowded scene analysis: A survey, *IEEE Trans. Circuits Syst. Video Technol.* **99** (2014).
20. W. Lin, Y. Chen, J. Wu, H. Wang, B. Sheng and H. Li, A new network-based algorithm for human activity recognition in videos, *IEEE Trans. Circuits Syst. Video Technol.* **24**(5) (2014) 826–841.
21. W. Lin, H. Chu, J. Wu, B. Sheng and Z. Chen, A heat-map-based algorithm for recognizing group activities in videos, *IEEE Trans. Circuits Syst. Video Technol.* **23**(11) (2013) 1980–1992.
22. R. Mehran, A. Oyama and M. Shah, Abnormal crowd behavior detection using social force model, in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, June 2009, pp. 935–942.
23. D. Munch, E. Michaelsen and M. Arens, Supporting fuzzy metric temporal logic based situation recognition by mean shift clustering, in *Advances in Artificial Intelligence*, Saarbrucken, Germany, Lecture Notes in Computer Science, Vol. 7526, June 2012, pp. 233–236.
24. B. Ni, S. Yan and A. Kassim, Recognizing human group activities with localized causalities, in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, Miami, FL, USA, June 2009, pp. 1470–1477.
25. P. Rota, N. Conci and N. Sebe, Real time detection of social interactions in surveillance video, in *Proc. European Conf. Computer Vision*, Florence, Italy (2012), pp. 111–120.
26. M. S. Ryoo and J. K. Aggarwal, Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities, in *Proc. IEEE Int. Conf. Computer Vision*, Kyoto, Japan (2009), pp. 1593–1600.
27. R. J. Sethi and A. K. Roy-Chowdhury, Individuals, groups, and crowds: Modelling complex, multi-object behaviour in phase space, in *Proc. IEEE Conf. Computer Vision Workshops*, Barcelona, Spain (2011), pp. 1502–1509.

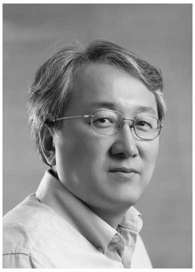
28. B. Solmaz, B. E. Moore and M. Shah, Identifying behaviors in crowd scenes using stability analysis for dynamical systems, *IEEE Trans. Pattern Anal. Mach. Intell.* **34**(10) (2012) 2064–2070.
29. H.-I. Suk, A. Jain and S.-W. Lee, A network of dynamic probabilistic models for human interaction analysis, *IEEE Trans. Circuits Syst. Video Technol.* **21**(7) (2011) 932–945.
30. J. Wang, Z. Liu, Y. Wu and J. Yuan, Mining actionlet ensemble for action recognition with depth cameras, in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, Providence, RI, USA (2012), pp. 1290–1297.
31. J. Xu, G. Yang, H. Man and H. He, L1 graph based on sparse coding for feature selection, in *Advances in Neural Networks*, Lecture Notes in Computer Science, Vol. 7951 (2013), pp. 594–601.
32. Y. Yin, G. Yang, J. Xu and H. Man, Small group human activity recognition, in *Proc. IEEE Int. Conf. Image Processing*, Lake Buena Vista, FL, USA (2012), pp. 2709–2712.
33. M. Zhang and A. A. Sawchuk, A feature selection-based framework for human activity recognition using wearable multimodal sensors, in *Proc. 6th Int. Conf. Body Area Networks*, Brussels, Belgium (2011), pp. 92–98.
34. C. Zhang, X. Yang, W. Lin and J. Zhu, Recognizing human group behaviors with multigroup causalities, in *Proc. IEEE/WIC/ACM Int. Conf. Web Intelligence and Intelligent Agent Technology Workshops*, Macau, China (2012), pp. 44–48.
35. Y. Zhou, T. S. Huang, B. Ni and S. Yan, Recognizing pair-activities by causality analysis, *ACM Trans. Intell. Syst. Technol.* **2**(5) (2011) 1–20.
36. Y. Zhu, N. Nayak and A. Roy-Chowdhury, Context-aware activity recognition and anomaly detection in video, *IEEE J. Sel. Topics Signal Process.* **7**(1) (2013) 91–101.

Nam-Gyu Cho received his B.S. degree in Information and Communication Engineering from University of Incheon, Incheon, Korea and his M.S. degree in Computer Science and Engineering from Korea University, Seoul, Korea, in 2009 and 2011, respectively. He is currently a Ph.D. candidate in the Department of Brain and Cognitive Engineering in Korea University. His research interests include computer vision, machine learning, and computational models of vision.

Young-Ji Kim received her B.S. degree in Computer Science from Women's University of Dongduk, Seoul, Korea and her M.S. degree in Computer Science and Engineering from Korea University, Seoul, Korea, in 2012 and 2014, respectively. Her research interests include computer vision, machine learning, and human activity analysis.

Unsang Park received his B.S. and M.S. degrees from the Department of Materials Engineering, Hanyang University, South Korea, in 1998 and 2000, respectively. He received his M.S. and Ph.D. degrees from the Department of Computer Science and Engineering, Michigan State University, in 2004 and 2009, respectively. Since 2012, he has been an Assistant Professor in the Department of Computer Science and Engineering at Sogang University. His research interests include pattern recognition, image processing, computer vision, and machine learning.

Jeong-Seon Park received her Ph.D. from the Department of Computer Science of Korean University, Seoul, Korea in 2005. Since 2005, she has been a Professor in the Department of Multimedia, Chonnam National University, Chonnam, Korea. Her research interests include image processing, pattern recognition, and IT convergence application.



Seong-Whan Lee received his B.S. degree in Computer Science and Statistics from Seoul National University, Seoul, in 1984, and his M.S. and Ph.D. degrees in Computer Science from the Korea Advanced Institute of Science and Technology, Seoul, Korea, in 1986 and 1989, respectively. From February 1989 to 1995, he was an Assistant Professor at, Chungbuk National University, Cheongju, Korea. In March 1995, he joined the faculty of the Department of Computer Science and Engineering at Korea University, Korea, and currently, he is the Hyundai-Kia Motor chair Professor and the head of the Department of Brain and Cognitive Engineering at Korea University. He is a fellow of the IEEE, IAPR, and the Korean Academy of Science and Technology. His research interests include pattern recognition, artificial intelligence and brain engineering.