2014 22nd International Conference on Pattern Recognition

Group Activity Recognition with Group Interaction Zone

Young-Ji Kim*, Nam-Gyu Cho†, Seong-Whan Lee*†
*Department of Computer Science and Engineering, Korea University, Seoul, Korea
†Department of Brain and Cognitive Engineering, Korea University, Seoul, Korea
Email: {yjkim, ngcho, swlee}@image.korea.ac.kr

***Abstract*—In this paper, we address the problem of recognizing group activities that include interactions between 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 so as to be robust against noisy information. Two novel features, Group Interaction Energy feature and Attraction and Repulsion Features, are proposed to better describe group activities within a GIZ. We demonstrate the effectiveness of our method with other methods on the public BEHAVE dataset.**

I. Introduction

Human activity recognition is one of the important issues in computer vision and has many practical applications such as human-computer interaction and video surveillance. The field of human activity recognition can be divided into three sub-fields: i) recognition of individual activity such as gesture and action [1]–[3], ii) recognition of complex activity which includes interaction between people [4]–[6], and iii) recognition of group activity [7]–[11]. While most of previous works focused on solving an individual action or complex activity recognition problem, recognizing a group activity remains a challenging and important problem – not only due to its technical difficulties, but also increasing practical requirements for 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 "walking" individual activities. Therefore, in order to recognize group activity, both local (individual) and global (group) information needs to be considered together.

Previous works on group activity recognition can be categorized into two approaches: image feature-based and trajectory-based approaches. The image feature-based approaches [4], [7], [8] describe an activity as a collection of motion gradient (spatio-temporal) features and their statistics. Therefore, 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 [6], [9]–[11] focus on analyzing human activities in terms of interactions between individual trajectories. Thus, they are more robust to occlusions or low-resolution. Reference [6] analyzes interactions between two individuals using the Granger Causality Test (GCT) [12]. However, owing to the limitations of GCT, they only focused on the pair-activity recognition problem. In order to deal with more complex situations, [10] 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., BEHAVE dataset [13] where 2 to 5 people perform ten group activities in aerial views (Fig. 1). In order to cope with this situation, [14] and [15] proposed methods which first detect sub-groups then recognize activities of each group. Reference [14] first cluster individuals into several sub-groups by the minimum spanning tree algorithm and then construct a network and extract a histogram feature. Reference [15] regard group activity as a combination of sub-groups and characterize four types of causalities: individual, pair, behavior, and inter-group. They cluster individuals by k-mean algorithm. Although [14], [15] demonstrated that using a sub-group information can be used to recognize group activity from a complex situation, how to detect a sub-group remains a challenging problem. In Particular, assuming a fixed number of group in a scene is not a robust solution.

In this paper, we tackle a problem of recognizing group activity by detecting a 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 needs to be carried out based on a better understanding of human behavior. Our contributions are two fold: i) We propose a novel meaningful group detection method by modeling proxemics [16]. 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 induced by human objects that do not participate an occurring activity. ii) We optimally describe a group activity in a GIZ by using an attraction and repulsion properties inspired by [11] which considered an interaction in terms of getting close, away, and keeping the same distance together.

The rest of this paper is organized as follows. In section II, we introduce our proposed methods. In section III, we demonstrate experimental results. Section IV, discusses conclusion and future work.

II. Method for Group Activity Recognition

In this paper, we represent a i-th human trajectory from time step 1 to T as,

$$\mathbf{z}_i = [\mathbf{x}_i^1, ..., \mathbf{x}_i^T], \tag{1}$$

where $\mathbf{x}_i^1$ is a tuple of an image coordinate (x, y) of human object i at the time step 1. We first define a Group Interaction

1051-4651/14 $31.00 © 2014 IEEE
DOI 10.1109/ICPR.2014.605

IEEE computer society

Fig. 1: Example of group activities in the BEHAVE dataset.

Zone (GIZ) between human objects by modeling proxemics, and then extract Attraction and Repulsion Features (ARF). Each group activity class is described as a bag-of-words and classified by linear SVM.

A. Detecting a Group Interaction Zone

Proxemics is defined as *"The interrelated observations and theories of man's use of space as a specialized elaboration of culture"* [16]. In other words, human has an inherent space of certain distance based on personal relationship; one tends to maintain a close distance with a familiar person. Otherwise, he feels discomfort and embarrassment when stranger approaches close. According to this, a distance around a person can be divided into four categories: intimate, personal, social, and public (Fig. 2). 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 personal distance of proxemics to represent a possibility of an interaction. An IPZ is a basic unit for detecting a GIZ, as given by the following four steps (Fig. 3). First, we draw an IPZ around each human object (Fig. 3-(a)). Second, we calculate the overlapping area between IPZs (Fig. 3-(b)) – the larger the overlapping region, the more likely group activity occurs. Third, we compute the ratio of the overlapping area to total area covered by interacting human objects as,

$$\gamma = \frac{\bigcap_{i=1}^{N_C} \Omega(\mathbf{x}_i)}{\bigcup_{i=1}^{N_C} \Omega(\mathbf{x}_i)} \tag{2}$$

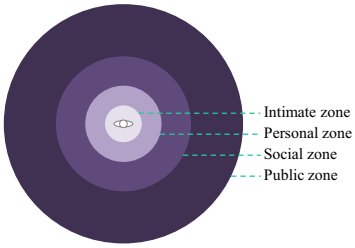
where $\Omega(\mathbf{x}_i)$ represents an IPZ of i-th human object and N_C is the number of people having overlapping IPZs. Then, we assign a GIZ ID to a set of human objects (Fig. 3-(c)) according to the following criterion,

$$\begin{array}{ll} \text{Assign the same GIZ ID}, & if \quad \gamma \geq \tau_{GIZ} \\ \text{None}, & \text{otherwise} \end{array} \tag{3}$$

where τ_{GIZ} is a threshold that controls the tendency of how likely it is that a set of human objects fall into the same GIZ. Finally, interaction features between every possible pairs within a GIZ are calculated (Fig. 3-(d)). Details of extracting features are described at the next subsection.

B. Extracting Interaction Features in a GIZ

In order to describe interactions within a GIZ, we propose a new feature that considers two properties: attraction and repulsion. The attraction property captures the tendency of people to get close to each other whereas the repulsion property captures



Category	Distance	Possible interaction
Intimate distance	0 - 46 cm	Embracing, touching, or whispering
Personal distance	46 - 120 cm	Interactions among good friends or family members
Social distance	1.2 - 3.7 m	Interactions among acquaintances (touching body doesn't happen)
Public distance	above 3.7 m	Public speaking

Fig. 2: Interpersonal zones based on proxemics [16].

the opposite case. These properties are closely connected with relative distance changes between human objects (Fig. 4). During a specific time period, from t_a to t_b, relative distance changes differently for these two properties.

$$\begin{array}{ll} \text{Attraction property}, & if \quad a > b \\ \text{Repulsion property}, & if \quad b > a \end{array} \tag{4}$$

where a and b are relative distance at t_a and t_b respectively. Thus, we first consider a subset of a trajectory information as follows,

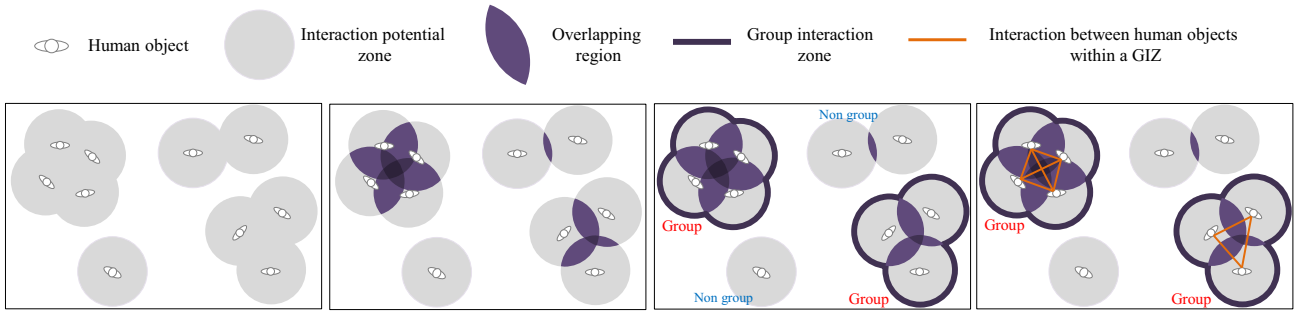
$$\xi_i^{T,k} = [\mathbf{x}_i^{T-(k-1)}, \mathbf{x}_i^{T-(k-2)}, ..., \mathbf{x}_i^{T}] \tag{5}$$

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 object i and j as,

$$\alpha_{ij}^{T,k} = \mathbf{d}(\xi_i^{T,k}, \xi_j^{T,k}) \tag{6}$$

where $\mathbf{d}(,)$ returns a distance vector where each element is the Euclidean distance between corresponding elements of two vectors. The mean and variance of $\alpha_{ij}^{T,k}$ is denoted as $\hat{\alpha}_{ij}^{T,k}$ and $\tilde{\alpha}_{ij}^{T,k}$. Now we calculate relative distances as,

$$\begin{aligned} \delta_{ij}^{T,k} &= \alpha_{ij}^{T,0} - \alpha_{ij}^{T-(k-1),0}, \\ \hat{\delta}_{ij}^{T,k} &= \hat{\alpha}_{ij}^{T,k} - \alpha_{ij}^{T-(k-1),0}, \end{aligned} \tag{7}$$



(a) Generating an interaction potential zone for each human objects (b) Calculating overlapping regions between human objects (c) Assigning a GIZ (d) Calculating interaction features within a GIZ

Fig. 3: Constructing a Group Interaction Zone (GIZ).

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

$$\psi^{+} = \sum_{t} I^{+}(\delta_{ij}^{T-(k-t),1}) \\ \psi^{-} = \sum_{t} I^{-}(\delta_{ij}^{T-(k-t),1}) \tag{8}$$

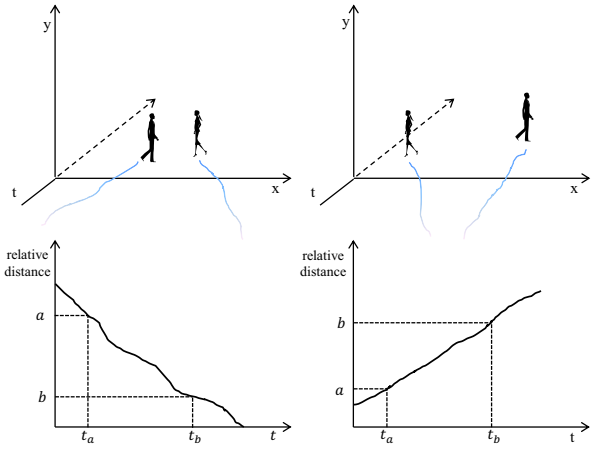
where $I^{+}(n)$ and $I^{-}(n)$ are the indicator functions that return 1 when value n is greater than 0 and vice versa. With additional $\nu_{ij}^{T,k}$ and $\phi_{ij}^{T,k}$, the magnitude and orientation of mean velocity during k time steps, we get a 7-dimensional feature for ARF as,

$$\lambda_{ij}^{T,k} = [\delta_{ij}^{T,k}, \hat{\delta}_{ij}^{T,k}, \nu_{ij}^{T,k}, \phi_{ij}^{T,k}, \psi^{+}, \psi^{-}, \sum_{t} \delta_{ij}^{T-(k-t),1}] \tag{9}$$

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 $\tilde{\alpha}_{ij}^{T,k}$. We also use the Causality and Feedback ratios from Granger Causality Test (GCT) features [6] to represent causality between objects. Given the trajectory information of two human objects i and j, GCT first defines predictor functions $P(x_i^t|\mathbf{x}_j^{t-l_i}(k))$ and $P(x_i^t|\mathbf{x}_i^{t-l_i}(k), \mathbf{x}_j^{t-l_i}(k))$ as follows,

$$x_i^t = \sum_{p=0}^{k-1} m^p x_i^{t-l_j-p} + \epsilon \\ x_i^t = \sum_{p=0}^{k-1} n^p x_i^{t-l_i-p} + \sum_{p=0}^{k-1} o^p x_i^{t-l_j-p} + \epsilon \tag{10}$$

where m, n, and o are the regression coefficients, l_i is the time lag, k is the length of trajectory subset, and ϵ is the Gaussian noise. $P(x_i^t|\mathbf{x}_i^{t-l_i}(k))$ and $P(x_j^t|\mathbf{x}_j^{t-l_j}(k))$ are modeled in a similar way. If the prediction error of $P(x_i^t|\mathbf{x}_i^{t-l_i}(k), \mathbf{x}_j^{t-l_i}(k))$ is smaller than $P(x_i^t|\mathbf{x}_i^{t-l_i}(k))$, we say the trajectory of object j is Granger causal to the trajectory of object i. If object i and j are Granger causal to each other, we say object i and j have feedback. Causality ratio is the ratio of the error of $P(x_i^t|\mathbf{x}_i^{t-l_i}(k))$ to the error of $P(x_i^t|\mathbf{x}_i^{t-l_i}(k), \mathbf{x}_j^{t-l_i}(k))$. Feedback ratio is the ratio of the error of $P(x_j^t|\mathbf{x}_j^{t-l_j}(k))$ to the error of $P(x_j^t|\mathbf{x}_i^{t-l_j}(k), \mathbf{x}_j^{t-l_j}(k))$ (for more details on the GCT, please refer to [6]). Finally, we get a 24-dimensional feature. We calculate an average of the features from every existing pairs within a GIZ.



(a) Attraction property (b) Repulsion property

Fig. 4: Example of trajectories and relative distance between objects in terms of two properties.

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

III. Experimental Results and Analysis

In this section, we demonstrate the performance of our method on the BEHAVE dataset [13] that provides various challenging then group activities. We evaluate the proposed method in two ways: i) comparison to previous works –

TABLE I: Group activity recognition comparison with [15] and [17].

	Ours	[15]	[17]
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 II: Group activity recognition comparison with [14].

	Ours	[14]
Split	100	93.1
WalkTogether	91.66	92.1
Fighting	83.33	95.1
InGroup	100	94.3
Average	**93.74**	93.65

how accurately our method can recognize presented group activities. ii) influence of features – how much the proposed feature can improve performance. We first describe the dataset, and then describe the the detail of our experiment.

A. Database

We use the public BEHAVE dataset which was captured as 25 frames per second frame rate and has 640×480 pixels image resolution. It consists of ten group activity classes performed by 2 to 5 people – (1) InGroup, (2) Approach, (3) WalkTogether, (4) Meet, (5) Split, (6) Ignore, (7) Chase, (8) Fight, (9) RunTogether, and (10) Following. However, please note that previous works demonstrating their performance on this dataset such as [14], [15], [17], used a subset of the dataset so as to best demonstrate their method. Similarly, we consider Approach (A), Split (B), WalkTogether (W), RunTogether (R), Fighting (F), and InGroup (I) to demonstrate our method. However, we also demonstrate our method on subsets used by [14], [15], [17]. The rest of activity classes were excluded because they do not include group activity or they only have few sequences, e.g., Meet class has only one instance. In this experiment, we use the trajectory information provided by the dataset.

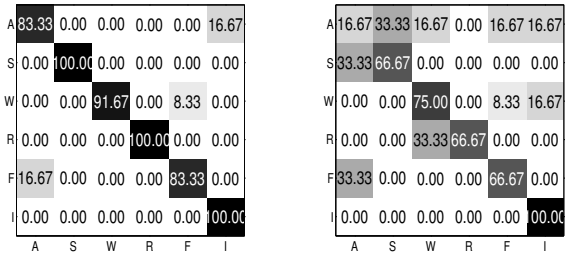
B. Implementation

The proposed method was implemented in MATLAB R2010a on a desktop PC (Intel Core i5-2500 3.30 GHZ CPU, 8 GB RAM, and 64 bit Windows 7 operating system). Parameters for the implementation were set as follows. The personal distance for an IPZ was 58 pixels, and the threshold τ_{GIZ} was 0.1. For feature extraction, the time interval k was 13. For group activity description, the window size ω was set as three frames and the cluster size for k-mean algorithm was 100. Classifiers were trained with MATLAB SVM toolbox. We evaluated our method via three-folds-cross-validation process due to the small number of instances for each class – the number varies from 4 to 35.

C. Comparison to Previous Works

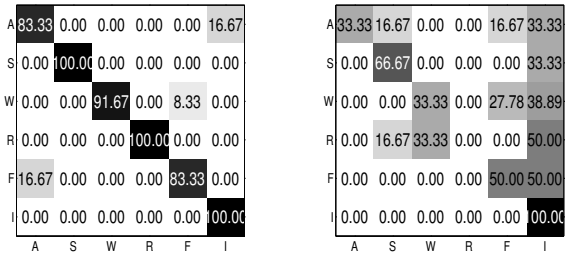
We first compare our method with [15], [17]. In this comparison we consider four classes – Approach, Split, WalkTogether, and InGroup. Reference [15] divided Approach activity into ApproachOne and ApproachBoth sub-classes, and also divided Split activity into same sub-classes. We compared activities with their "Both". As can be seen in Table I, our method achieves a performance better than [15] and [17].

For the comparison with [14], we consider Split, WalkTogether, Fighting, and InGroup activities. Table II shows the result. Our method shows better performances for Split and InGroup activities and slightly lower for WalkTogether and Fighting. However, it still outperforms in the average.



(a) ARF + GCT + AF (b) GCT + AF

Fig. 5: Confusion matrices representing influences of features of our method.



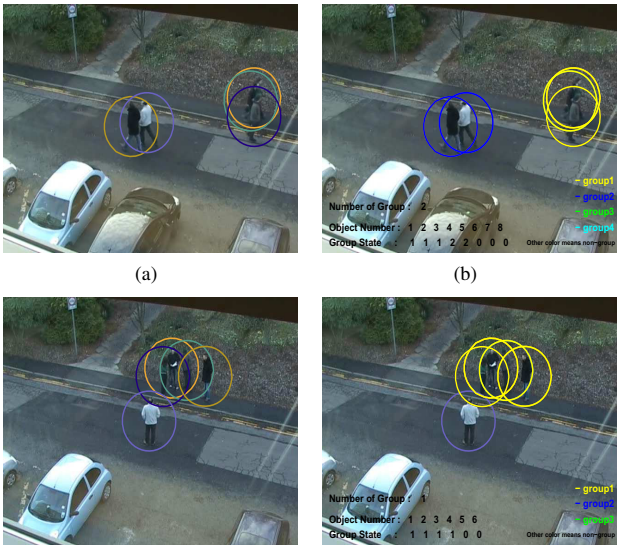
(a) With GIZ detection (b) Without GIZ detection

Fig. 6: Confusion matrices representing the influence of GIZ detection on the BEHAVE dataset.

D. Influence of Features and GIZ

In this subsection, we evaluate influence of the features used in our method. We use three types of features: ARF, GCT, and AF. We combined our features as follows: (a) all features (ARF + GCT + AF), 24-dimensions; (b) without proposed features (GCT + AF), 17-dimensions. As can be seen in Fig. 5, the proposed features that considers attraction and repulsion properties are considerably useful.

We also demonstrate the influence of the proposed GIZ group detection method. Fig. 6 shows the performance. In this comparison, the proposed method performs better than the method without GIZ. Some examples of detecting GIZs on the BEHAVE dataset are illustrated in Fig. 7.



(a) (b) (c) (d)

Fig. 7: Example of GIZ detection on the BEHAVE dataset. The first row shows the case when objects walk together within each group and the second row shows the case when four objects are in one group and the other object is not. The right column shows detected GIZ (assigned different color for different GIZ ID).

IV. Conclusions and Future Work

In this paper, we proposed a novel GIZ detection method based on proxemics to better represent human social behavior for detecting a meaningful groups from a group activity situation. Then we proposed ARF features for representing group activity in terms of attraction and repulsion properties. We demonstrated our method on the public BEHAVE dataset. The results showed that our method works better on this dataset compared to other methods. As future work, we plan to apply our method to situations where a large number of people exists in a scene with more complex activities such as bullying.

Acknowledgment

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 Number: 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. 2014-044-023-001).

References

[1] T. Guha and R. K. Ward, "Learning sparse representations for human action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 8, pp. 1576–1588, 2012.

[2] X. Wu, D. Xu, L. Duan, and J. Luo, "Action recognition using context and appearance distribution features," in *Proceedings of IEEE Computer Vision and Pattern Recognition*, Colorado Springs, CO, USA, June 2011, pp. 489–496.

[3] J. Wang, Z. Liu, Y. Wu, and J. Yuan, "Mining actionlet ensemble for action recognition with depth cameras," in *Proceedings of IEEE Computer Vision and Pattern Recognition*, Providence, RI, USA, 2012, pp. 1290–1297.

[4] M. S. Ryoo and J. K. Aggarwal, "Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities," in *Proceedings of IEEE International Conference on Computer Vision*, Kyoto, Japan, 2009, pp. 1593–1600.

[5] U. Gaur, Y. Zhu, B. Song, and A. Roy-Chowdhury, "A string of feature graphs model for recognition of complex activities in natural videos," in *Proceedings of IEEE International Conference on Computer Vision*, Barcelona, Spain, 2011, pp. 2595–2602.

[6] Y. Zhou, T. S. Huang, B. Ni, and S. Yan, "Recognizing pair-activities by causality analysis," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 5, pp. 1–20, 2011.

[7] W. Choi, K. Shahid, and S. Savarese, "Learning context for collective activity recognition," in *Proceedings of IEEE Computer Vision and Pattern Recognition*, Colorado Springs, CO, USA, June 2011, pp. 3273–3280.

[8] M. R. Amer and S. Todorovic, "A chains model for localizing participants of group activities in videos," in *Proceedings of IEEE International Conference on Computer Vision*, Barcelona, Spain, 2011, pp. 786–793.

[9] Z. Cheng, L. Qin, Q. Huang, S. Jiang, and Q. Tian, "Group activity recognition by gaussian processes estimation," in *Proceedings of IEEE International Conference on Pattern Recognition*, Istanbul, Turkey, 2010, pp. 3228–3231.

[10] B. Ni, S. Yan, and A. Kassim, "Recognizing human group activities with localized causalities," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshops*, Miami, FL, USA, 2009, pp. 1470–1477.

[11] R. J. Sethi and A. K. Roy-Chowdhury, "Individuals, groups, and crowds: Modelling complex, multi-object behaviour in phase space," in *Proceedings of IEEE Conference on Computer Vision Workshops*, Barcelona, Spain, 2011, pp. 1502–1509.

[12] C. W. J. Granger, "Investigating causal relations by econometric models and cross-spectral methods," *Econometrica*, 1969.

[13] S. Blunsden and R. Fisher, "The behave video dataset: Ground truthed video for multi-person behavior classification," in *Proceddings of The British Machine Vision Conference*, vol. 2010, no. 4, Aberystwyth, UK, August 2010, pp. 1–12.

[14] Y. Yin, G. Yang, J. Xu, and H. Man, "Small group human activity recognition," in *Proceedings of IEEE International Conference on Image Processing*, Lake Buena Vista, FL, USA, 2012, pp. 2709–2712.

[15] C. Zhang, X. Yang, W. Lin, and J. Zhu, "Recognizing human group behaviors with multi-group causalities," in *Proceedings of 2012 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology Workshops*, Macau, China, 2012, pp. 44–48.

[16] E. Hall, *The Hidden Dimension*. Anchor Books, 1966.

[17] D. Munch, E. Michaelsen, and M. Arens, "Supporting fuzzy metric temporal logic based situation recognition by mean shift clustering," in *Lecture Notes in Computer Science*, Saarbrucken, Germany, June 2012, pp. 233–236.