Recovering Social Interaction Spatial Structure from Multiple First-Person Views

Tian Gan[1,3], Yongkang Wong[2], Bappaditya Mandal[3], Vijay Chandrasekhar[3], Liyuan Li[3], Joo-Hwee Lim[3], Mohan S. Kankanhalli[1]

[1]School of Computing, National University of Singapore
[2]Interactive & Digital Media Institute, National University of Singapore
[3]Institute for Infocomm Research, Singapore

gantian@comp.nus.edu.sg, yongkang.wong@nus.edu.sg,
{bmandal,vijay,lyli,joohwee}@i2r.a-star.edu.sg, mohan@comp.nus.edu.sg

ABSTRACT

In a typical multi-person social interaction, spatial information plays an important role in analyzing the structure of the social interaction. Previous studies, which analyze spatial structure of the social interaction using one or more third-person view cameras, suffer from the occlusion problem. With the increasing popularity of wearable computing devices, we are now able to obtain natural first-person observations with limited occlusion. However, such observations have a limited field of view, and can only capture a portion of the social interaction. To overcome the aforementioned limitation, we propose a search-based structure recovery method in a small group conversational social interaction scenario to reconstruct the social interaction structure from multiple first-person views, where each of them contributes to the multifaceted understanding of the social interaction. We first map each first-person view to a local coordinate system, then a set of constraints and spatial relationships are extracted from these local coordinate systems. Finally, the human spatial configuration is searched under the constraints to "best" match the extracted relationships. The proposed method is much simpler than full 3D reconstruction, and suffices for capturing the social interaction spatial structure. Experiments for both simulated and real-world data show the efficacy of the proposed method.

Categories and Subject Descriptors

I.2.10 [**Vision and Scene Understanding**]: Video analysis; I.4.8 [**Scene Analysis**]: Motion; I.5.4 [**Applications**]: Signal processing

General Terms

Algorithms

SAM'14, November 07 2014, Orlando, FL, USA.

http://dx.doi.org/10.1145/2661126.2661134 .

Keywords

Social Interaction, Social Computing, Video Analytics, Egocentric Vision

1. INTRODUCTION

Human social interactions play an important role in our daily lives. In a typical social interaction, the spatial information plays an important role as social signal [12], which helps people understand and structure the ongoing social interaction. When the social interaction is observed by several cameras, the overall social interaction spatial structure is useful in organizing and retrieval of multiple videos, as well as multiple-view video summarization. In prior work, social interaction has been studied with static third-person view cameras (e.g., surveillance cameras, Kinect depth sensors, *etc.*) [1, 2, 5, 9]. However, the static nature of these cameras restricts their usage and potential to recover the full spatial information of a social interaction. In addition, the "outsider" nature of the third-person view camera results in several occlusions due to the unrestricted human movement. In contrast, first-person view camera, such as Google Glass or GoPro camera, does not have the aforementioned restrictions and can be used for better spatial information retrieval.

Recently, the advances in sensor technologies and wearable computing device has motivated studies on first-person view cameras. In [3], Fathi *et al.* proposed to use single first-person view video to detect several types of social interactions, such as dialogue, discussion, and monologue. Though the "group videographer" can fully participate in the group experience, the "videographer" has never been observed because only a single camera-view has been considered in the work. Park *et al.* used multiple head-mounted cameras to estimate 3D social saliency [11]. They assume all the camera poses are registered in 3D via structure from motion, which is impractical in real world scenario.

In this work, we propose a method to recover the social interaction spatial structure from multiple first-person views. Different from [11], our work reconstructs the human social interaction spatial structure without employing full 3D reconstruction. In particularly, we use the constraints from the available first-person view cameras to estimate the spatial location and orientation of each observed individuals in a global space. The proposed method also fundamentally different from [3], where each person who wears a wearable

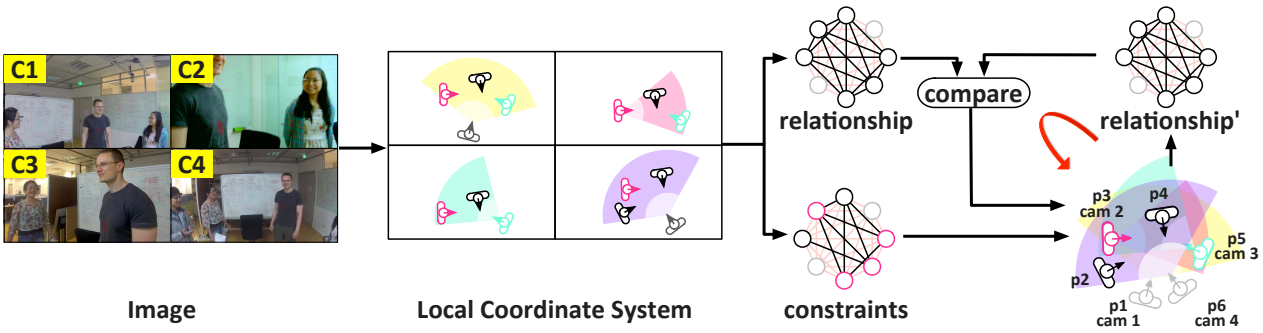

Figure 1: An overview of the proposed structure recovering method. Given multiple video sequences captured from multiple first-person view cameras, the 2D Local Coordinate Systems are constructed. Then a set of constraints between the camera wearer and the observed individuals are derived for each view; similarly, a set of pairwise relationships are built for all visible people. In the final stage, all the possible configurations (i.e., combination of all the people's spatial information) are enumerated with the constraints, and the configuration which generates the spatial relationship with minimum matching cost is selected as the recovered spatial structure.

camera sensors is acting as a "videographer". In this scenario, each videographer will be observed in another videographers' video. Additionally, multiple first-person views provide better overall understanding of the social interaction. In this work, we focus on small group social interactions scenario with face-to-face conversation, as the small group interaction tend to have more social dynamics [4, 6].

The contributions of this work are as follows: we combine multiple first-person view cameras to recover social interaction spatial structure. This populates the interaction group with multiple views, which is useful for understanding the complete interaction structure. We propose a search-based reconstruction method, which is simpler and computational efficient when compared with 3D reconstruction-based method. To the best of our knowledge, this is the first time that the multiple first-person view cameras are combined to analyze the social interaction spatial structure without performing pre-calibration.

The remaining of this paper is organized as follows. The details of proposed method are given in Section 2. Experiment and evaluation are presented in Section 3. The main findings and possible future directions are discussed in Section 4.

2. PROPOSED APPROACH

Given video sequences captured from multiple first-person view cameras, our goal is to recover the global spatial structure of human social interaction from these local observations. Our proposed approach consists of the following three stages:

1. For each local observation, we construct a 2D Local Coordinate System (LCS) with the spatial location and viewing direction of the camera positioned at origin and 90 degrees axis, respectively. Automated face detection and identity inference are applied to locate the observed people in the corresponding LCSs.

2. Given the constructed LCSs from each view, a set of constraints are derived based on the relative positions between the camera wearer and the observed individuals. Similarly, a set of pairwise relationships are built for all visible people.

3. By discretizing the person's location and orientation in LCSs, all possible configurations (i.e., combination of all the people's spatial information) are enumerated with the constraints. The configuration which generates the spatial relationship with smallest matching cost is selected as the recovered spatial structure.

An overview of the proposed method is shown in Figure 1.

2.1 Problem Formulation

Assume that $\boldsymbol{P} = \{p_1, p_2, \ldots, p_N\}$ is the people set consisting of N unique individuals. In the common 2D Global Coordinate System (GCS), each observed person p_n is represented as a four-tuple: $< x_n, y_n, \alpha_n, I_n >$, where x_n, y_n, and α_n are the spatial location and orientation, respectively. I_n represents the identity of p_n. Let the relationships among all observed person in $\boldsymbol{P}$ as $\mathcal{R}(\boldsymbol{P}) = \{\beta_{i,j}\}, i, j = 1, \ldots, N$, where $\beta_{i,j}$ is the angle for p_j with respect to p_i's location and orientation:

$$\beta_{i,j} = arctan\frac{y_j - y_i}{x_j - x_i} - \alpha_i \tag{1}$$

We assume that the first M people in $\boldsymbol{P}$ are equipped with wearable cameras. Given the m-th camera c_m worn by person p_m in GCS, the p_m's spatial location and view direction defines the Local Coordinate System (LCS) for p_m, termed as LCS_m. LCS_m contains a set of people $\boldsymbol{P}_m \subseteq \boldsymbol{P}$ observed by p_m, with p_m being positioned at the origin and 90 degrees anticlockwise with respect to x-axis positive direction.

The goal of this work is to estimate the spatial locations and orientations $\hat{\boldsymbol{P}}$ for all observed people in $\boldsymbol{P}$ in GCS, such that the matching cost between $\mathcal{R}(\hat{\boldsymbol{P}})$ and $\mathcal{R}(\boldsymbol{P})$ can be minimized via:

$$\arg\min_{\hat{\boldsymbol{P}}} \sum_{m} C(\mathcal{R}(\hat{\boldsymbol{P}}), \mathcal{R}(\boldsymbol{P})) \tag{2}$$

where

$$C(\mathcal{R}(\hat{\boldsymbol{P}}), \mathcal{R}(\boldsymbol{P})) = \sum_{i,j=1\ldots N} ||\hat{\beta}_{i,j} - \beta_{i,j}|| \tag{3}$$

2.2 Image to Local Coordinate System

Given an image captured from camera, face detection and identity inference are applied to extract the information

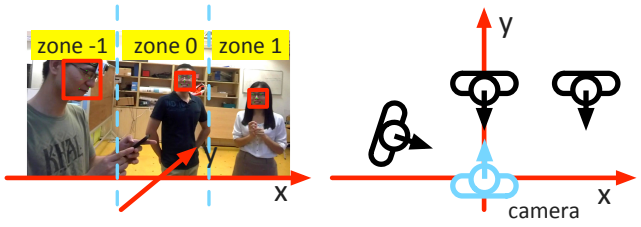

Figure 2: Illustration of transformation from image to local coordinate system.

about the observed people in the image. We assume the camera viewing direction is the same as the camera wearer c_m's orientation, thus we can create LCS_m which the camera wearer is at the origin with 90 degrees anticlockwise in terms of x-axis positive direction. In order to describe all the visible people in the image in LCS_m, we divide the image into sz_{grid} zones in the horizontal direction. The center zone is the 0-th zone, and its number increases/decreases along the positive/negative x-axis. This zone number is used as the x coordinate for each person based on his corresponding zone. As for the y coordinate, we calculate each person's face size and set a series of threshold $\{\sigma_1, \sigma_2, \ldots, \sigma_S\}$ to estimate the distance. The value of y is based on the index for its nearest threshold. In addition, we set each person's orientation with respect to x-axis positive direction as the orientation α. Figure 2 shows a visual illustration of the transformation process from observed image to LCS.

2.3 Spatial Constraint Extraction

Given the extracted LCSs from all first-person views, we now create a spatial relationship matrix. A relative position table $\mathcal{T}_{\text{rel}}$ is defined for all visible people. Suppose $p_r =< x_r, y_r, \alpha_r >$ and $p_a =< x_a, y_a, \alpha_a >$ are in the same coordinate system. The item $\mathcal{T}_{\text{rel}}(p_r, p_a) =< x_a^r, y_a^r, \alpha_a^r >$ is p_a's relative location and orientation with respect with p_r:

$$\begin{bmatrix} x_r^a \\ y_r^a \end{bmatrix} = \begin{bmatrix} \cos\alpha_r & \sin\alpha_r \\ -\sin\alpha_r & \cos\alpha_r \end{bmatrix} \left(\begin{bmatrix} x_a \\ y_a \end{bmatrix} - \begin{bmatrix} x_r \\ y_r \end{bmatrix} \right) \quad (4)$$

$$\alpha_r^a = \alpha_a - \alpha_r$$

For each unique pair of camera wearer p_m and the respective observed person p_o, we can generate two items, $\mathcal{T}_{\text{rel}}(p_m, p_o)$ and $\mathcal{T}_{\text{rel}}(p_o, p_m)$, based on their spatial coordinate and orientation in LCS_m. Therefore, we can generate $\mathcal{T}_{\text{rel}}$ to encodes all the pairwise spatial relationships between the camera wearer and each observed individuals. We do not compute the relative positions among the observers, because their relationships are inferred through the camera viewer, which are less reliable.

2.4 Search of Configuration

As the objective of this work is to recover the social interaction spatial structure, rather than the exact locations for all observed individual, we formulate our problem as a search problem instead of 3D reconstruction. We first discretize the space and person's orientation, and assume that different people must be in different grid locations; however, the overall solution space is still relatively large. In order to address this issue, we limit the visible range of all cameras, and specify the search space for each individual with

Algorithm 1 Pseudo code for the Search of Configuration

```
1: procedure SEARCH(confirmed,cost,sSpace,relation)
   ▷ This is a recursive function: confirmed is the current confirmed people's configuration; cost is the cost for matching confirmed with relation; sSpace is the current search space; relation is the spatial relationship constraints; result and bestCost are the global variables which store the results.
2:     if NOT(confirmed)=0 then
3:         result ← confirmed;
4:         bestCost ← cost;
5:     else
6:         newIdx ← SelectConfirmation(relation)
7:         for all newLoc in newIdx's sSpace do
8:             if IsOccupied(newLoc) then
9:                 continue
10:            end if
11:            newConfirmed ← confirmed + newIdx
12:            newCost ← CalcCost(newIdx,relation)
13:            if newCost < bestCost then
14:                newsSpace ← UpdateSolutionSpace(sSpace, newIdx, relation)
15:                SEARCH(newConfirmed, newCost, newsSpace, relation)
16:            end if
17:        end for
18:    end if
19: end procedure
```

the obtained structure constraints (See Section 2.3). The structure constraints reduce the search space significantly. For example, suppose we fix a location and orientation for person p_r, and we have the spatial constraint that person p_a is in front of camera wearer person p_r, then we will only search the area in front of p_r, rather than the entire space. The more constraints extracted from the local observations, the smaller the search space is. In this way, the problem is formulated as to find a configuration (combination of all the people's spatial coordinates and orientations) in a finite search space such that: **(1)** no more than one person is in the same location; and **(2)** the pairwise relationship generated from the result configuration best matches (with least matching cost as defined in Eq. 2 and 3) the observed relationship in every local coordinate system.

We propose an algorithm to estimate the location and orientation of all individual for the formulated search problem. The pseudo code is presented in Alg. 1. The underlying idea is to enumerate all possible configurations to find the "best" configuration. The search space is reduced by the "Spatial Relationship" constraints $\mathcal{T}_{\text{rel}}$ as defined in Section 2.3. Since $\mathcal{T}_{\text{rel}}$ only encodes the relationship between the camera wearer and the observed individuals, we further calculate the overall relationship matrix $\mathcal{R}_{overall}(\text{P})$ which encodes the pairwise relationship between all the people. Note that a people pair (p_a,p_b) may have more than one relationship generated from different cameras, we keep all of them to calculate the average relationship difference. The function *SelectConfirmation* in Alg. 1 chooses an unconfirmed person prioritized by **(1)** a smaller search space and **(2)** more constraints. The function *CalcCost* in Alg. 1 calculates the cost of matching the relationship $\mathcal{R}_{confirmed}$ generated by the confirmed people's information and the relationship gen-

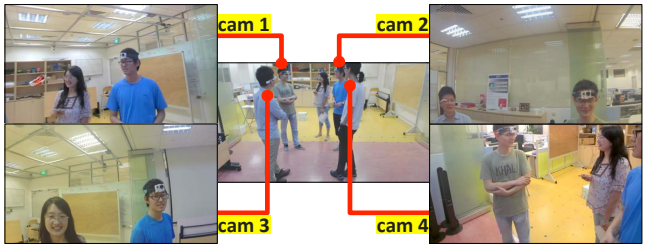

Figure 3: Experiment setup for real world experiment. The center image is captured by the static web camera; the rest are four first-person views from the corresponding wearers' cameras.

erated by $\mathcal{R}_{overall}$. The function $UpdateSolutionSpace$ in Alg. 1 reduces the solution space using the constraints related to new confirmed person $newIdx$.

Suppose n people are observed in the available first-person view cameras, and the size of the search space for each person is d, the worst-case run time complexity is $O(d^n)$. In practice, given the small group interaction scenario (with less than 10 interactants for one interaction group), and the constraints which can significantly reduce the search space, the actual running time is acceptable. However, we still limit the maximum amount of time spent as the stopping criterion in case of the worst-case scenario.

3. EXPERIMENTS

In this section, we examine the performance of the proposed work on both synthetic data and real-world recording. For the synthetic data, we simulate social interactions with 2 to 10 individuals in a social group, with "valid social distance" from interpersonal distance proxemics study [8] and "people creating a shared space" from F-formation system constraints. Each interaction contains 100 test cases, in which each person inside can be treated as a camera-wearer with 120° field of view. Persons within 4 meters and ±90° from the frontal position with respect to the camera are regarded as visible to the camera. The video sequences from real-world environment are captured with four first-person view cameras (two Google Glasses and two GoPro cameras) and a static web camera in an indoor lab environment. A snapshot of the recording is shown in Figure 3: the center image is captured by the static web camera; the rest are four first-person views from the corresponding wearers' cameras.

The face information (location and orientation) in each image are detected using Face++ Research Toolkit [10]. The identity correspondence between different cameras is labeled by humans.

Following the work on spatial-similarity-based image retrieval [7], we evaluate the social interaction spatial structure similarity between the result and ground truth based on the spatial orientation relationship. Particularly, the social interaction structure generates a Spatial Orientation Graph (SOG), in which the node is the person and the edge is the spatial relationship between the people. The similarity between two social interaction structures is then quantified based on the number as well as the extent to which the edges of the resulting SOG conform to the corresponding edges of the ground truth SOG.

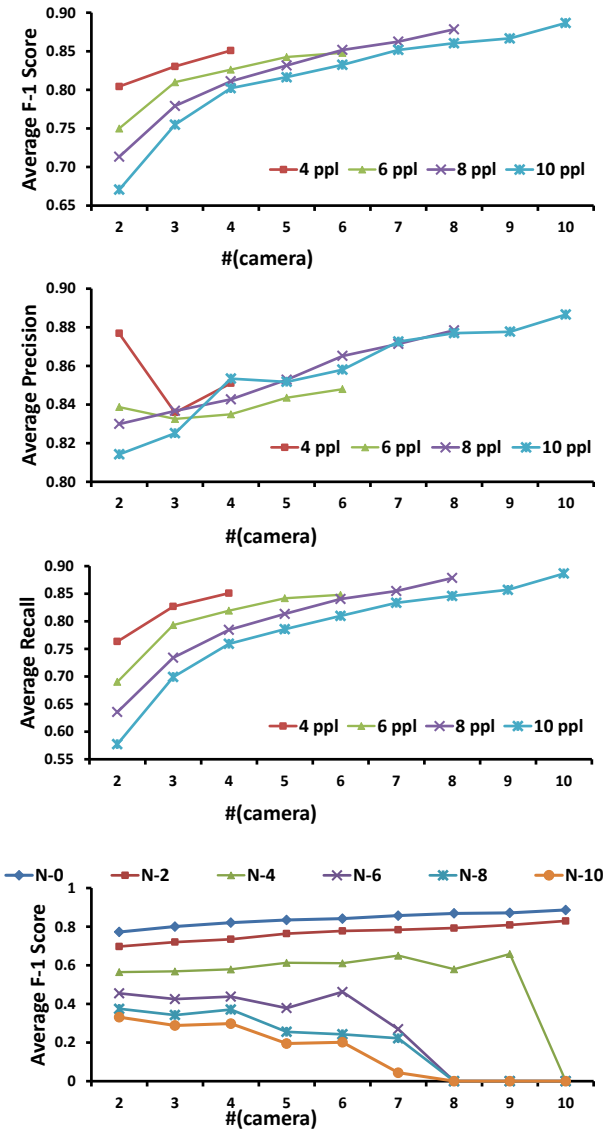

Figure 4: Experimental results on simulation data. The first three figures show the average F-1 score, precision, and recall with the configuration of noise level 0, grid size level 3 and 30° tolerance. The last figure shows the average F-1 score of different noise level with the configuration of grid size level 3 and 30° tolerance.

Formally, consider the example where p_i and p_j are two nodes. The edge $e_{ij} = \alpha_{ij}$ is defined as the angle for p_j using p_i as reference. If the difference of angle against the ground truth is less than a predefined threshold $\sigma_{tolerance}$, these two edges are regarded as similar. In this way, we quantitatively report the results with the F-measure metric. The overall precision and recall are the average of all the nodes' precision and recall. For each node, the precision is $P_i = \frac{tp}{tp+fp}$ and the recall is $R_i = \frac{tp}{tp+fn}$, where the true positive tp and false positive fp are the number of the "similar" and "dissimilar" edges in the result with respect to the ground truth, the false negative fn is the number of edges which are "dissimilar" or "missing" with respect to the result.

Figure 5: Experimental results on real-world data. PX indicates view from person X. The last column is the result of our proposed work.

3.1 Evaluation on Simulation Data

Three types of factors affect the result: **(1)** the number of cameras and the number of visible people in the raw image data, **(2)** the grid size and the extent of noise in the conversion process of image to local coordinate system, and **(3)** the tolerance threshold for the evaluation. In order to evaluate the effects of these factors, the grid size sz_{grid} is quantified into 4 levels: the i-th level means the field of view of a local coordinate system is divided into a $(2 * i + 1)^2$ space; the noise is divided into 10 levels: the reference is the uniform noise with the range $[-d_{max}, d_{max}]$ for x and y, $[-90°, +90°]$ for orientation, where d_{max} is the maximal visible distance from the camera. The noise in each level is set to 10% of the reference noise; The tolerance threshold are 15°, 30°, and 45°.

We first evaluate the effect of changing the number of cameras and number of people. The first three figures in Figure 4 show the average F-1 score, precision, and recall with the configuration of noise level 0, grid size level 3 and 30° tolerance. We can see that given the number of cameras, the average precision does not change too much (within the range of [0.8,0.9]). The small drop for precision from 2 cameras to 3 cameras is due to the conflicting information brought about by different cameras. The performance will increase when more cameras' observation are considered to resolve ambiguities. Compared to precision, there is a distinct increase in recall when the number of cameras is increased. Using the same parameter configuration, we evaluate the performance for various grid size levels and tolerance levels, which are discussed in Section 3.2 together with the performance on real-world data. The last figure in Figure 4 shows the influence of the noise level.

3.2 Evaluation on Real-world Data

Figure 5 shows three visual examples obtained from real-world data. PX indicates view from person X. The last column is the result of our proposed work. We can see that although the person who wears the camera is not shown in his or her own image, he or she can show up in other camera views (e.g., person 4 in P2). The last two examples are consecutive frames with the same social interaction spatial structure. But we notice a different result between these two examples in terms of person 6. From the raw camera data we can see that in example 3 person 6 turned his head towards another side, resulting in additional constraints (person 1 with person 6, and person 5 with person 6) from image, which improves the result compared to example 2.

Table 1: Comparison of real-world and simulated data using various grid sizes

	Simulation data			Real-world data		
Grid Size	F1	Pre	Rec	F1	Pre	Rec
1	0.700	0.720	0.683	0.504	0.602	0.433
2	0.750	0.765	0.729	0.511	0.612	0.439
3	0.820	0.844	0.802	0.538	0.644	0.463
4	0.840	0.865	0.823	0.537	0.645	0.460

Table 2: Comparison of real-world and simulated data using various tolerance values

	Simulation data			Real-world data		
Tolerance	F1	Pre	Rec	F1	Pre	Rec
15°	0.602	0.618	0.588	0.340	0.404	0.294
30°	0.822	0.844	0.802	0.538	0.644	0.463
45°	0.898	0.921	0.876	0.686	0.817	0.591

We also did quantitative experiments on the real-world data. We use 10 scenarios of real-world data consisting of 2 to 10 people equipped with 4 cameras. Each scenario contains 100 consecutive frames with the frame rate of 5fps. We assume each scenario is with the same social interaction structure, and manually label the ground truth. Table 1 compares the performance on real-world data and simulation error free data with 4 cameras, 2 to 10 people, and 30° tolerance value. Table 2 changed this configuration by fixing grid size as 3 and enumerate the tolerance value. As we can see from these tables, the performance on real-world data follows similar trends as the simulation data, which increases as the value of the grid size and tolerance go up. The degradation in precision of real-world data compared to simulation data comes from **(1)** error from real-world im-

age to local coordinate mapping, and **(2)** missing temporal dynamics in the ground truth for real-world data. Also, recall for the real-world data is much worse than simulation data. This is due to the simplified simulation data not accounting for occlusion, motion blur present in raw image data (e.g., the image from P8 in Figure 5). Overall, the performance for real-world data is acceptable, and it can be further improved if temporal information is considered.

4. CONCLUSION

In this paper, we combined multiple first person view cameras for social interaction spatial structure reconstruction. Our proposed search-based method is much simpler than 3D reconstruction, and achieves good performance for recovering the spatial social interaction structure. For future work, we aim to add temporal information to improve performance. We also plan to analyze other multi-modal social signals from first person views.

5. ACKNOWLEDGEMENT

This research was carried out at the NUS-ZJU Sensor-Enhanced Social Media (SeSaMe) Centre. It is supported by the Singapore National Research Foundation under its International Research Centre @ Singapore Funding Initiative and administered by the Interactive Digital Media Programme Office.

6. REFERENCES

[1] L. Bazzani, M. Cristani, D. Tosato, M. Farenzena, G. Paggetti, G. Menegaz, and V. Murino. Social interactions by visual focus of attention in a three-dimensional environment. *Expert Systems*, 30(2):115–127, 2013.

[2] M. Cristani, L. Bazzani, G. Paggetti, A. Fossati, D. Tosato, A. D. Bue, G. Menegaz, and V. Murino. Social interaction discovery by statistical analysis of F-formations. In *Proceedings of British Machine Vision Conference*, pages 1–12, 2011.

[3] A. Fathi, J. K. Hodgins, and J. M. Rehg. Social interactions: A first-person perspective. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1226–1233, 2012.

[4] N. Fay, S. Garrod, and J. Carletta. Group discussion as interactive dialogue or as serial monologue: The influence of group size. *Psychological Science*, 11(6):481–486, 2000.

[5] T. Gan, Y. Wong, D. Zhang, and M. S. Kankanhalli. Temporal encoded f-formation system for social interaction detection. In *Proceedings of ACM International Conference on Multimedia*, pages 937–946, 2013.

[6] D. Gatica-Perez. Automatic nonverbal analysis of social interaction in small groups: A review. *Image Vision Computing*, 27(12):1775–1787, 2009.

[7] V. N. Gudivada and V. V. Raghavan. Design and evaluation of algorithms for image retrieval by spatial similarity. *ACM Transactions on Information Systems*, 13(2):115–144, 1995.

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

[9] H. Hung and B. J. A. Kröse. Detecting F-formations as dominant sets. In *Proceedings of the International Conference on Multimodal Interfaces*, pages 231–238, 2011.

[10] M. Inc. Face++ research toolkit. www.faceplusplus.com, December 2013.

[11] H. S. Park, E. Jain, and Y. Sheikh. 3D social saliency from head-mounted cameras. In *Advances in Neural Information Processing Systems*, pages 431–439, 2012.

[12] A. Vinciarelli, M. Pantic, and H. Bourlard. Social signal processing: Survey of an emerging domain. *Image Vision Computing*, 27(12):1743–1759, 2009.