Temporal Encoded F-formation System for Social Interaction Detection

Tian Gan
School of Computing
National University of Singapore
gantian@comp.nus.edu.sg

Daqing Zhang
Télécom SudParis, France
CNRS UMR 5157 SAMOVAR
daqing.zhang@it-sudparis.eu

Yongkang Wong
Interactive & Digital Media Institute
National University of Singapore
yongkang.wong@nus.edu.sg

Mohan S Kankanhalli
School of Computing
National University of Singapore
mohan@comp.nus.edu.sg

ABSTRACT
In the context of a social gathering, such as a cocktail party, the memorable moments are generally captured by professional photographers or by the participants. The latter case is often undesirable because many participants would rather enjoy the event instead of being occupied by the photo-taking task. Motivated by this scenario, we propose the use of a set of cameras to automatically take photos. Instead of performing dense analysis on all cameras for photo capturing, we first detect the occurrence and location of social interactions via F-formation detection. In the sociology literature, F-formation is a concept used to define social interactions, where each detection only requires the spatial location and orientation of each participant. This information can be robustly obtained with additional Kinect depth sensors. In this paper, we propose an extended F-formation system for robust detection of interactions and interactants. The extended F-formation system employs a heat-map based feature representation for each individual, namely Interaction Space (IS), to model their location, orientation, and temporal information. Using the temporally encoded IS for each detected interactant, we propose a best-view camera selection framework to detect the corresponding best view camera for each detected social interaction. The extended F-formation system is evaluated with synthetic data on multiple scenarios. To demonstrate the effectiveness of the proposed system, we conducted a user study to compare our best view camera ranking with human’s ranking using real-world data.

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

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

ACM ’13 Barcelona
Copyright 2013 ACM 978-1-4503-2404-5/13/10 ...$15.00.
http://dx.doi.org/10.1145/2502081.2502096.

Keywords
Social Interaction, Social Computing, F-formation, Video Analytics, Behaviour Modeling

1. INTRODUCTION
In social gatherings such as cocktail parties, conference receptions, etc., the interactions between the event participants are often captured with multiple cameras or smartphones. In many scenarios, the event participants play the role of the photographer, which forces them to become passive observers of the event. This goes against the primary purpose of socializing where the participants ought to enjoy the events. Furthermore, the participants may not capture all the important shots due to the fact that no one is able to observe the whole event [2]. Therefore, it would be desirable to have the photos taken by the professional photographer or by an automated photo-capture system. Hiring a professional photographer is generally expensive and hence not affordable for many types of informal social gatherings. With an automated photo-capture system, the cost can be negligible. Moreover, these approaches can be scaled to support closed-door events with privacy concerns, or a live streaming system that shows the latest photo on a public display, or to automatically annotate videos capturing a social event.

One potential solution for the automated photo-capture system is to configure a set of cameras to record the entire event. The recorded videos can be manually edited after the event, or analyzed using video post-processing [10, 18, 21]. Such works have been proposed for various applications in the literature, such as video summarization for video conferencing [13], lecture webcasting [10, 18], sport events [20], and video mash-up for live performance [21, 22]. However, these approaches require large storage capacity for the videos, as well as computationally expensive vision algorithms to analyze the footages. Therefore, these approaches cannot be scaled for large scale deployment. In addition, the aforementioned approaches can only be applied to specific predefined actions/tasks [10, 13, 18, 20, 21]. In practice, one cannot predict a priori where the interesting “events” will occur so it is difficult to zoom and take good photos by a priori set-up.

In contrast to the aforementioned methods, another approach is to employ the F-formation concept for detecting the social interaction [1, 5, 12]. In the sociological literature,
F-formation is defined as a set of spatial patterns maintained during social interactions by two or more interactants, where the spatial and orientation relationship among multiple persons forms an interaction space [5, 9]. The F-formation is formalized into three social spaces: o-space, p-space, and r-space (see Figure 1). The o-space, also known as the joint transaction space, is the interaction space between the interactants. In practical systems, we can conclude that a social interaction is formed whenever an o-space is created [9]. The p-space and r-space are the area occupied by the interactants and the area that surrounds the interactants, respectively. Examples of various interaction patterns are shown in Figure 1.

The F-formation based approach has two main benefits. Firstly, a social interaction can be easily identified from the detection of o-space, which is derived from the orientations and spatial locations of the interactants. Secondly, the computational resources can be utilized only on the detected interaction regions. This also increases the likelihood of capturing photos that are more “interesting” without recourse to dense analysis on all video streams. However, most of the existing F-formation based approaches do not incorporate the temporal information. This gives negative classification result for some interaction arrangements. For example, two persons walking past each other would be immediately considered as a valid F-formation. This is intuitively against the idea of having a social interaction.

Recently, a heat map based approach is proposed to recognize the type of human group activity [4]. The heat map based approach models the human movement trajectory as a heat map with thermal diffusion. The resulting heat map is used to classify the query activity as one of the predefined activities (e.g., gather, follow, separate, etc.) with a surface fitting process [4]. We argue that the surface fitting approach is not suitable for the aforementioned social events. This is because the number of participants in social events is generally high, which results high intraclass variance for each type of group activity. Despite that, we acknowledge that heat map based approach is an effective method to incorporate the temporal information.

In this paper, we propose an extended F-formation system for robust interaction and interactant detection. The extended F-formation system employs a heat map based feature representation for each unique individual, namely Interaction Space (IS), to model their respective location, orientation, and temporal information. In our work, the individual’s spatial location and orientation are retrieved with the Kinect depth sensors. Given the IS of all individuals at the t-th frame, we detect the interaction centers (i.e., o-space) and the respective interactants, as well as the location of the best view camera. The proposed temporal-encoded IS based approach is evaluated on both the synthetic data and real-world experimental environment.

For the real-world scenario, we configure a test environment with four Pan-Tilt-Zoom (PTZ) cameras and three Kinect depth sensors. The snapshot of the test environment is shown in Figure 2. To the best of our knowledge, this is the first time F-formation is used for automated social event photo-capture application.

The rest of this paper is organized as follows. The background and related works are discussed in Section 2. The details of proposed method are given in Section 3. Extensive evaluations on both the synthetic and real-world data are presented in Section 4. The main findings and possible future directions are covered in Section 5.

2. PREVIOUS WORK

In recent years, there is growing interest in the detection of social group behavior [1, 5, 7]. Social interaction detection requires modeling of two components: (1) individual activities, and (2) social relationships between individuals. The literature can be categorized into three approaches. The first category relies on the visual information and statistical models [15, 16, 19], however, its efficacy in real world application is questionable due to the uncontrolled nature of human behavior. The second category utilizes visual and audio data collected from various sensors, and performs multimodal processing to detect interaction [3]. The third category analyzes the social interactions using social behavioral cues [1, 5, 7, 24]. In [24], Vinciarelli et al. organized the social behavioral cues into five categories: (i) physical appearance, (ii) gesture and posture, (iii) face and eyes behavior, (iv) vocal behavior, and (v) space and environment. These cues have been recognized in Psychology literature as the most important factors in human judgments [24].

In this paper, we focus on the space and environment social behavioral cue for social interaction detection. A popular sociological concept to exploit this behavioral cue is the F-formation system [9]. This concept is commonly used in computer-supported cooperative work, where the interaction is established with an appropriate spatial relationship between participants. For example, Yamashita et al. [26] examined how changes in seating position across different sites affect the video-mediated communication by exploring the F-formation. While in [17], the F-formation knowledge is used to navigate the robot to join an interaction group using a socially adapted behavior with lower risk of collision.
and disturbance. Despite being tangentially relevant to social interaction detection, it inspired us to make use of the F-formation to explore and analyze social interactions.

There are a number of methods to detect F-formation. Marquardt et al. [11] used the ubiquitous computing environment to sense the social proximity of people in the form of F-formation. Their goal is to motivate group interactions. Specifically, they define two persons to be in an F-formation if the following conditions are met: (1) they are not standing behind each other; (2) the angle between their orientation vectors is smaller than 180 degrees; (3) the distance between them is small enough. After the three conditions are met, the algorithm iterates over all pairs of people, calculates the distance and angle between them, and assigns an F-formation type (i.e., side-by-side, L-shaped, face-to-face, or none) based on tolerance thresholds. This work is intended to prove that small-group interaction can be sensed in the form of F-formation. Cristani et al. [5] designed an F-formation recognizer based on the Hough-voting strategy. First, they take a certain number of candidate sample interaction centers for each subject, then the candidate positions are voted by weighted samples. The interaction center is selected as the position which has the highest value. Their method incorporates the uncertainty by assuming the position and orientation of each subject as Gaussian random variables. However, this method detects the F-formation for each frame independently. Therefore, the temporal information, or the continuous group interactions, are not explored in this work. In [1], the social interaction is detected by taking temporal information into consideration. They determine whether two persons are interacting with each other when the following three conditions are satisfied: (1) the distance between the subjects is closer than 2 meters; (2) their Field of View (FoV) are overlapped; (3) their heads are positioned inside the reciprocal FoVs. Then they accumulate the existence of this relationship over a period of time. These conditions assume that each person should have at least one person to be related with, in terms of visual attention, within a single social group. However, the three simple rules define the interaction as the pairwise relationship.

It cannot characterize many types of interaction spatial arrangements, such as "side-by-side" (refer to Figure 1(c)), in which each person need not to be in the reciprocal FoVs. This is a common scenario in social interaction, that is all people look towards a certain direction. Different from F-formation based analysis, the heat map, a kind of graphical representation of data, has been employed to analyze some types of social interaction [23, 4]. It highlights the "hot" data regions in a visually pleasant way. The heat map can also be interpreted as a kind of knowledge accumulation. Heat map can be created with various types of information, by which rich information might be retained in the heat map for further analysis. For example, Singh et al. [23] aggregated social multimedia data spatiotemporally to derive semantic situation information. The result of the aggregated data is one kind of heat map. Chu et al. [4] proposed a heat-map based algorithm for group activity recognition. By using the heat map feature to represent activities, the temporal information can be modeled effectively. The recognition of group activity is based on this heat map feature with the surface fitting process [4].

3. EXTENDED F-FORMATION SYSTEM

We propose an extended F-formation system which uses a heat map based representation to encode the spatial location, orientation, and temporal information. In this work, we consider a video sequence of a social event, where the spatial coordinate and orientation for person \( k \) at \( t \)-th frame, \( p_k^t = \langle x_k^t, y_k^t, \theta_k^t \rangle \), is first obtained from multiple Kinect depth sensors with Kinect for Windows SDK\(^1\). The \( t \)-th frame is represented as \( P^t = \{ p_1^t, p_2^t, \ldots, p_k^t | k \} \) where \( |k| \) is the cardinality of \( t \)-th frame. The aim of this work is to identify all possible interaction centers, \( \{ I_1^t, I_2^t, \ldots, I^n_t \} \), and their respective interactants \( P_{I^t}^t \subset P^t, t = 1, \ldots, n \).

We continue this section by first giving an overview of the proposed framework, followed by describing the heat map based F-formation detection algorithm. We then elaborate on the algorithm to detect the interactants for each F-formation and their respective best view camera.

\(^1\)http://www.microsoft.com/en-us/kinectforwindows/
3.1 Framework

A conceptual diagram of the proposed framework is shown in Figure 3. Given the spatial coordinates and orientations for each individual, we first compute the individual Interaction Space (iIS), where the Interaction Space (IS) is restricted by the individual’s field of attention (see Figure 4). The IS is modeled as a heat map where the highest energy point is selected with prior knowledge obtained from sociology study [8]. For each time frame, a global Interaction Space (gIS) is computed by averaging the overlapped iIS. We then compute the temporal encoded IS for each individual and the global view (denoted as TiIS and TgIS, respectively). The computed TiIS and TgIS are used to detect the F-formation(s), interactants, and the respective best view camera(s).

3.2 F-formation Detection

3.2.1 Individual Interaction Space

Given person \(k\) at \(t\)-th frame, \(p_k^t = (x_{k,t}, y_{k,t}, \theta_{k,t})\), we first represent its iIS as a heat map, where the point with the highest energy is called the individual’s interaction center, denoted by \(s_k^t\). The spatial coordinate of \(s_k^t\) is defined as:

\[
s_k^t = (x_{k,t}, y_{k,t}) = (x_k^t + r \cos \theta_k^t, y_k^t + r \sin \theta_k^t)
\]

(1)

where \(r\) and \(\theta\) represent the distance from \(p_k^t\)'s spatial location and its orientation, respectively. The \(p_k^t\)'s iIS has the highest energy at \(s_k^t\) and diffuses towards the neighboring region. Furthermore, \(p_k^t\)'s field of view is restricted between \([-\beta, \beta]\) degrees with respect to its orientation and a radius of \(r^t\). The field of view forms an active IS for each individual.

The value of iIS\(_k^t\) is assumed to have a Gaussian distribution, we apply the two-dimensional Gaussian function on \(s_k^t\) to compute iIS\(_k^t\) as follows:

\[
iIS_k^t(x, y) = \exp\left(-\frac{(x-x_{k,t})^2}{2\sigma_x^2} - \frac{(y-y_{k,t})^2}{2\sigma_y^2}\right)
\]

for \(F_k^t(x, y) = 1\)

\[
iIS_k^t(x, y) = 0
\]

otherwise

(2)

where \(\sigma_x^2\) and \(\sigma_y^2\) are the variance on x-axis and y-axis, respectively. \(F_k^t\) represents the binary mask for \(p_k^t\)'s field of view. A conceptual example is shown in Figure 4.

3.2.2 Global Interaction Space

Given the iIS for all individuals detected at \(t\)-th frame, iIS\(_k\) = \([iIS_1^t, iIS_2^t, \ldots, iIS_{k|t|}^t]\), the gIS is computed to represent the common interaction space for all individuals. The gIS for pixel located at \((x, y)\) is computed as:

\[
gIS^t(x, y) = \left\{
\begin{array}{ll}
\frac{1}{|\text{gIS}(x,y)|} \sum_{k=1}^{k|t|} \text{iIS}_k^t(x, y) & \text{if } |\text{iIS}_k^t(x, y)| \geq 2, \\
0 & \text{otherwise}
\end{array}
\right.
\]

where the notation \(|\text{iIS}_k^t(x, y)|\) counts the number of nonzero entries of iIS\(_k^t\) at location \((x, y)\). Example of iIS and gIS are shown in Figure 5.

3.2.3 Temporal encoded Interaction Space

To address the missing element of motion trajectory in the original F-formation system, the temporal information is encoded in both iIS and gIS using an energy decay based accumulation approach. In the following discussions, we elaborate on the temporal encoding algorithm with gIS, where the same method is applied to iIS. Consider the gIS at frame \(t_{cur}\), the corresponding Temporal encoded gIS (TgIS) is modelled as

\[
TgIS^{t_{cur}} = \int_0^{t_{cur}} (1 - e^{-K t}) \cdot gIS_t \cdot e^{-K (t_{cur} - t)} dt
\]

(4)

where the term \(1 - e^{-K t}\) is a scale factor to keep TgIS\(_{t_{cur}}\) in the range of \([0, 1]\). The weight decay term \(e^{-K (t_{cur} - t)}\) controls the contribution of gIS\(_t\) whereas the most recent frame has the highest weight. The constant \(K\) controls the rate of decay.

The example of IS and temporal encoded IS are shown in Figure 5. We demonstrate two unique scenarios here. In scenario 1, person 1 and person 2 initiate the first frame of the social interaction\(^2\). The gIS (top) shows an IS with high energy level. Based on the proposed interaction center detection algorithm (see Section 3.2.4), this will be classified as a valid F-formation. On the other hand, the energy level in TgIS (bottom) is much lower. In scenario 2, both person 1 and person 2 have maintained the social interaction for a period of time. Now, the energy level of TgIS has risen to a high level (similar to gIS). This is indeed a desired property. Consider a scenario where multiple persons are constantly walking pass each other, the original F-formation would give many false alarms.

In the preliminary experiment, we also observed that the temporal encoded iIS can stabilize the detection error (a side effect from hair style, clothing or accessories) from the Kinect depth sensors. In this case, the orientation of some individual gives the shaking effect over a period of time. Based on our observation, the temporal encoding can smooth the interaction space.

3.2.4 Interaction Centers Detection

The energy level in TgIS characterize the location of social interactions as several “hot spots”. To locate these “hot spots”, we first apply the Interaction threshold, \(T_i\), to the heat map. Then, we apply a smoothing function \(f()\), e.g. the Gaussian filter, to the thresholded TgIS. This is because the temporal encoding step (i.e., Eq. 4) introduces a “staircase step” effect to the heat map. We note that this effect is largely influenced by the moving speed of each person and the selected frame rate.

Given the thresholded and smoothed TgIS, we locate all the local maxima in the heat map, which gives us a set of \(\text{maxima}\)

\(^2\)This scenario is the same as the passing by scenario
Figure 5: Example of individual Interaction Space (iIS) and global Interaction Space (gIS) in two scenarios. Top row: IS computed for each static frame; Bottom row: Temporal information is encoded for each IS. Red and blue indicate high and low energy level, respectively. Scenario 1 can represent the first frame when two persons form an interaction (similar to the passing by scenario). Scenario 2 represent a social interaction after a period of time.

Algorithm 1 Pseudo code for interaction centers detection

Input: Global Interaction Space \( gIS \in \mathbb{R}^{M \times N} \), Interaction threshold \( T_i \in [0, 1] \), Interaction center radius \( \tau_{center} \), and Smoothing function \( f(\cdot) \).

Output: Interaction centers \( I = \{I_1, I_2, \ldots, I_i\} \)

1: for all \( (x, y) \in gIS \) do
2: if \( gIS(x, y) < T_i \) then
3: \( gIS(x, y) \leftarrow 0 \)
4: end if
5: end for
6: \( gIS_{smooth} \leftarrow f(gIS) \)
7: CandiCenters \( \leftarrow \) findLocalMaxima(\( gIS_{smooth} \))
8: while \( |\text{CandiCenters}| > 0 \) do
9: Center\(_{max}\) \( \leftarrow \) findMaxCenters(CandiCenters)
10: mergeList \( \leftarrow \) \( \emptyset \)
11: for all \( i = 1, 2, \ldots, |\text{CandiCenters}| \) do
12: if \( dist(\text{CandiCenters}_i, \text{Center}_{max}) \leq \tau_{center} \) then
13: \( \text{mergeList} = \text{mergeList} \cup \{\text{CandiCenters}_i\} \)
14: end if
15: end for
16: newCenter \( = \) merge(mergeList)
17: \( I_c = I_c + \{\text{newCenter}\} \)
18: CandiCenters \( = \) CandiCenters \( - \) mergeList
19: end while

3.3 Interactant Detection

The detection of interactant is performed by analyzing the contribution of each individual with respect to the interaction center. Given a detected interaction center \( I_i \) and a binary mask \( M_i \) for its o-space, we compute the contribution score \( S_{k}^{c} \) for person \( k \) at \( t \)-th frame via:

\[
S_{k}^{c}(k, i) = \sum_{x,y} [TiIS_{k}(x, y) \times M_i^{c}(x, y)]
\]

(5)

The mask \( M_i^{c} \) has the value of 1 for a pixel within \( 2\tau_i \) radius from \( I_i \). We consider a person as the interactant of \( I_i \) if and only if \( S_{k}^{c}(k, i) \) is smaller than a predefined contribution threshold \( T_c \). In other words, a person will be considered to be an interactant if the individual has stayed in the o-space for a period of time. We note that this is only valid for the TiIS. For the non-temporal encoded IS, each individual will be considered as an interactant when they enter the o-space.

3.4 Best View Camera Selection

We formulate the best view camera selection method as a ranking system. For each detected interaction space and the corresponding interactants, we compute the camera selection score for each camera and rank the camera based on the scores. As discussed in Section 1, the F-formation has three interaction spaces: o-space, p-space, and r-space. We define a ring region of camera ranking zone, \( A \), on r-space, where the zone is equally divided in to \( N \) sub-zones.

The selection score for sub-zone \( n \), \( s(A_n) \), and interaction center \( I_i \) at \( t \)-th frame is computed as

\[
s(A_n) = \frac{1}{|P_i^{t}|} \sum_{k \in P_i^{t}} \sum_{(x, y) \in A_n} TiIS_{k}(x, y)
\]

(6)

where \( |P_i^{t}| \) is the cardinality of the interactant set \( P_i^{t} \).

For each camera, we assign the selection score of the sub-zone that is located between the camera and \( I_i \). Note that if the number of sub-zones of the camera ranking zone is small, the number of cameras assigned to each sub-zone would be higher. We argue that there is no rule of thumb for the selection of this value, the selection should be based on the target application and the number of available cameras, or be learnt for a particular application. A conceptual example are shown in Figure 6.

---

*The merge function can be a choice of mean, max, medium, etc. We use the max function in this paper because it gave the best results in our preliminary tests.*
Figure 6: Conceptual diagram for the best view camera selection method. The interaction space covers both the o-space and p-space.

4. EXPERIMENTS

In this section, we examine the performance of the proposed extended F-formation system. We first evaluate the accuracy for both the interaction center detection and interactant detection. Then, the output of the best view camera selection algorithm is “visually inspected” on real-world recording and also evaluated with a user study.

Experiments were conducted on synthetic data and real-world video recording. For the synthetic data, we simulated 10 scenarios of social interaction with two variables (i.e., the number of unique individual and the concurrent interaction centers). Each scenario is denoted by a standard name scenario_#people_#center. For each scenario, we randomly generate 5 sequences where each sequence consists of 600 frames with the frame rate of 5 fps. Each frame consists of the individual’s ID, spatial location and orientation, as well as the spatial location of the interaction centers. The ground truth data consists of the number and location of each interaction centers, and its corresponding interactants. It was generated by the simulation script.

In order to collect video sequences from real-world environment, we set up a set of cameras, including three Kinect depth sensors and four PTZ cameras, in an indoor lab environment. The snapshot of the lab environment and the floor plan are shown Figure 2 and 7, respectively. All 7 cameras are calibrated to the ground plane. In addition, the Kinect depth sensor is used to extract the location and the orientation of all persons. Note that the extracted orientation information is more reliable if the person’s orientation deviates within 30 degrees from the Kinect’s principal axis. Therefore, we only consider the data that fall within this range. On some scenarios, where the Kinect depth sensor could not distinguish the frontal and back view, manual correction is applied. Furthermore, we manually correlate the label for each person across the seven cameras. For this work, we record three video sequences of four persons with eight unique group interactions.

4.1 Parameters Selection

In our application, we can define some of the parameters with the study from the sociological literature. Hall [8] introduced proxemics as a theory to study the interpersonal spatial relationships. The physical distance and the social distance between individuals can be correlated and categorized into four discrete zones: (1) intimate (0m - 0.45m), (2) personal (0.45m - 1.2m), (3) social (1.2m - 3.5m), and (4) public (> 3.5m). In this work, we set r = 0.45m as the distance between person’s current location and his interaction center, r' = 3.5m as the maximum distance for this person’s influence, 2β = 90 degrees as the individual Interaction Space angle, and σx = σy = 0.6 to constrain the heat energy distribution. Based on the preliminary experiments, the remaining parameters are as follows: \( K_t = 10 \), \( T_i = 0.65 \), and \( T_c = 0.22 \).

4.2 Interaction Detection Experiments

In this subsection, we evaluate the accuracy for detecting the interaction center and the respective interactants. We quantitatively report the results with the F-measure metric, which is

\[
F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}
\]

where the precision and recall are defined as \( \frac{tp}{tp + fp} \) and \( \frac{tp}{tp + fn} \), respectively. The notations tp, fp, and fn are the total number of true positive, false positive, and false negative (in terms of center/interactant detection), respectively.

We employ a similar evaluation metric as in [5] for the interaction center detection. The interaction center is considered correctly detected if the distance between the detected interaction center and the ground-truth data is smaller than r (2m in our experiments), and at least two-thirds of the participants of the ground truth are correctly identified. For the interactants detection, we evaluate the performance only when the interaction center is valid for frame t.

Two variants of our proposed method are evaluated. The first is the heat map based F-formation system without encoding the temporal information (denoted as IS), while the second is the temporal encoded F-formation system (denoted as Temporal encoded IS). We also compare our
method with Bazani et al.’s approach [1] (see Section 2 for more details). We denote their method as Bazani et al.

The complete average precision, recall, and F$_1$ scores are shown in Figure 8. The left column shows the results for interaction center detection, and the right column indicates the performance of interactants detection. The average performance for all scenarios are shown in Table 1 and 2. As shown in the figure and tables, our approach outperforms Bazani et al. [1] with a noticeable margin. For the interaction center detection, the F$_1$ score of the Temporal encoded IS and IS outperformed Bazani et al. [1] by 16.2% and 18.2%, respectively. We observed that all the results for scenario$_{10}$ are very low for all approaches. This is because the ratio between number of people and center is too high. The scenario is generally very crowded, therefore the algorithms which study the spatial relationship between the interactants are not suitable. Another observation is that the recall rate of the proposed method (for both variance) outperform Bazani et al. [1] with a significant margin. For scenario$_{2,1}$, the improvement is about 43.4% and 51.2% for Temporal encoded IS and IS, respectively.

For the interactants detection experiment, the difference in performance is even more obvious. In particular, we fix the number of people and increase the number of interaction centers (e.g., scenario$_{4,1}$ and scenario$_{4,2}$). The difference of performance can be explained as follows. Our method models the interaction space as a common interaction area, and it can robustly handle group interaction with various spatial arrangements for a group of people. In contrast, Bazani et al. [1] define the interaction as a pairwise relationship, where each person should be in the reciprocal visual field of view and the group is established based on this pairwise relationship. This method would fail to detect the common side-by-side interaction pattern (refer to Figure 1(c)), where each person is not within the reciprocal visual field of view of the corresponding interactant. This phenomenon is more obvious when the number of interaction centers increases. Under such a scenario, the distribution of the group is more sparse and the likelihood of the aforementioned problem is relatively higher.

During the comparison between the accuracy of IS and Temporal encoded IS, we find that the performance of Temporal encoded IS is generally worse than IS. This is contradictory to our expectation and we note that this is a problem of our ground truth data, where a frame is considered as having valid interaction center when two persons meet. The Temporal encoded IS can only identify an interaction center after a period of time (a side effect of the energy decay based accumulation approach). Despite that, we cannot determine a reasonable frame duration to form a mutual social interaction. Therefore, modifying the ground truth to accommodate this scenario is not reasonable. To establish
Figure 9: Experimental result with real-world video recording. Each column represents a unique social interaction. (a) The spatial locations and the orientations of the detected interactants, as well as the camera ranking zone; (b) Temporal encoded global Interaction Space; (c-f) The snapshots obtained from the top 4 ranked cameras with (c) having the highest rank.
our hypothesis, we generated 10 sets of simulated sequences with 4 individuals and 1 interaction center. Each sequence has a spatial dimension of 1000 × 1000 and 1000 frames in total. No interaction is allowed in these sequences and only precision is reported. The results are shown in Table 3. The results agree with our hypothesis where Temporal encoded IS gives a precision of 0.999 and IS gives 0.800.

4.3 Best View Camera Selection Experiments

In this subsection, we demonstrate an initial investigation of the effectiveness of the best view camera selection method. The snapshots of the top 4 ranked cameras in three unique social interactions are shown in Figure 9. Row (a) shows the interactants’ spatial location and the respective orientation. The camera ranking zone is shown around the interactants. Row (b) are the TgIS. Row (c-f) are the snapshots obtained for each sequence where row (c) indicates the top rank image and row (f) to be the lowest rank. Each column shows a unique interaction. This preliminary experiment shows that the camera ranking zone with the highest selection score is indeed corresponding to TgIS. For the first and the third interactions, the top two ranked images also show more front-tier view when compared to snapshots located in row (f).

To validate the efficacy of the best view camera rank, we conducted a user study to compare our camera ranking with human expectations. This study was conducted on fifty individuals (34 males and 16 females). The participants were asked to rank the camera views from eight detected social interactions. Each interaction consists of six views which were captured by different cameras at the same time. In order to compare our ranking with the users’ camera view ranking, we calculate the average matching accuracy of our top-N rank with with K variation of users ranking. For each sequence, our top-N ranked cameras and users’ top-K ranked cameras are considered as matched if one of the camera view was presented in both ranking. The results are presented with Cumulative Match Characteristic (CMC) curve.

As shown in Figure 10, the top-1 rank from our algorithm only agrees with 33% and 56% of users’ top-1 and top-2 rank, respectively. We argue that the low accuracy for our top-1 rank is reasonable as the users’ top-1 ranked camera are not consistent. When we consider the top-2 rank from our algorithm, the matching accuracy raised significantly to 65% for users’ top-1 rank and 86% for users’ top-2 rank. This indicates that our method generally agrees with users’ expectation. Further investigation of the data shows that the performance is heavily biased by one specific detected social interaction. In this sequence, best view camera ranked by our algorithm contains a person who is partially cropped from the view (due to camera placement and interaction spatial location). Although the frontal face of all three persons were visible in this view, most user ranked this view as the worst. We acknowledge this problem in our algorithm and highlight that this can be further addressed with automated PTZ camera control [14] to provide visually satisfying snapshots.

5. CONCLUSIONS AND FUTURE WORKS

In this paper, we proposed an extended F-formation system for robust interaction and interactant detection. Inspired by the heat map based method for human group activity recognition [4], we define the individual Interaction Space (iIS) and global Interaction Space (gIS) to model the objects’ spatial locations and orientation. In order to address the problem of unintentional F-formation detection, such as two persons passing by or a person walking past a social interaction, we encoded the temporal information via an energy decay based accumulation function. The heat map based Interaction Space is used to detect the interaction center and the corresponding interactants. In addition, we further utilized it to detect the camera with high probability to capture good photos. We also propose a camera configuration for automated photo capturing application. In addition to the standard PTZ cameras, we added a number of Kinect depth sensors to obtain accurate spatial locations and the respective orientations.

We evaluate our proposed method with both the synthetic data and real-world video recording. Experiments on 10 unique scenarios show that the proposed method outperforms the rule based F-formation system in [1]. The results on interaction center detection in the Precision, Recall, and F1 score show improvement of 7.1%, 22.0%, and 16.1%, respectively. The results on interactant detection are even more convincing. We evaluate the best view camera selection with the real-world video recording. The results on visual analytic and user study agree with our expectation.

For future work, we aim to apply machine learning techniques to model the prior knowledge with training data. In addition, we would like to explore the possibility of combining the proposed extended F-formation system with two components. First, we would like to apply automated PTZ camera control algorithm [14] to improve the image resolution, as well as to maximize the photo capturing range. Second, we would like to investigate the ability to perform automated photo capturing via key-frame selection [6, 25], or video summarization [10, 18, 27] to actively analyze the content of the video.
6. 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.

7. REFERENCES


