Social Saliency Prediction

Hyun Soo Park Jianbo Shi
University of Pennsylvania
{hypar,jshi}@seas.upenn.edu

Abstract

This paper presents a method to predict social saliency, the likelihood of joint attention, given an input image or video by leveraging the social interaction data captured by first person cameras. Inspired by electric dipole moments, we introduce a social formation feature that encodes the geometric relationship between joint attention and its social formation. We learn this feature from the first person social interaction data where we can precisely measure the locations of joint attention and its associated members in 3D. An ensemble classifier is trained to learn the geometric relationship. Using the trained classifier, we predict social saliency in real-world scenes with multiple social groups including scenes from team sports captured in a third person view. Our representation does not require directional measurements such as gaze directions. A geometric analysis of social interactions in terms of the F-formation theory is also presented.

1. Introduction

Imagine an artificial agent such as a service robot operating in a social scene as shown in Figure 1. It would detect obstacles such as humans in the scene and plan its trajectory to avoid collisions with the obstacles. It may plan a trajectory that passes through the empty space between the audiences and performer. This trajectory intrudes on the social space created by their interactions, e.g., occluding the sight of the audiences, and thus, it is socially inappropriate. We expect the artificial agent to respect our social space although the boundary of the social space does not physically exist. This requires social intelligence [31]—an ability to perceive, model, and predict social behaviors—to be integrated into its functionality.

Joint attention is the primary basis of social intelligence as it serves as a medium of social interactions; we interact with others *via* joint attention[1]. Understanding joint attention, specifically knowing where it is and knowing how it moves, provides a strong cue to analyze and recognize group behaviors. It has been recognized that computer vision solutions can provide a large scale measurement for developing a computational representation of joint attention. The challenges are: (1) human detection and tracking failure in the presence of occlusions, (2) scene variability, e.g., the different number, scale, and orientation of social groups, and (3) inaccurate measurements of gaze directions.

Figure 1. We present a method to estimate the likelihood of joint attention called *social saliency* from a spatial distribution of social members. The inset image shows the top view of the reconstructed scene. The blue points are the points belonging to humans. The heat map shows the predicted social saliency and we overlay this map by projecting onto the ground plane in the image.

While there are many factors involved in joint attention, our main question is: can one predict joint attention using social formation information alone, without the gaze information of each member?

In this paper, we show that it is possible to empirically learn the likelihood of joint attention called *social saliency* as a function of a social formation, a spatial distribution of social members, using data from first person cameras. Three key properties of our predictive joint attention model are: a) it is scale and orientation invariant to social formations; b) it is invariant to scene context, both indoors and outdoors; c) it is robust to missing data. Once this model is constructed, a sparse point cloud representation of humans can be used to predict the locations of joint attention as shown in the inset image of Figure 1, without any directional measurement such as gaze directions—we measure and learn this predictive model in the first person view, and apply in the third person view.

To construct such joint attention model, we use first person cameras. With multiple first person cameras, joint attention can be precisely measured in 3D since the ego-

[1]Gaze directions are correlated with joint attention in quasi-static social interactions while motion becomes a dominant factor in rapid dynamic interactions as shown in Figure 3(b).

motion of the cameras follows gaze behaviors of the wearers [2, 9, 23]. Furthermore, we can simultaneously compute the 3D positions of the wearers to provide precise measurements of social formations. These first person in-situ computational measurements of the geometric relationship between joint attention and social formation can be applied in a variety of third person social interaction scenes including a basketball game where the players strategically take advantage of spatial formations.

Contributions To our best knowledge, this is the first work that provides a predictive model encoding the geometric relationship between joint attention and social formation, using in-situ 3D measurements from first person cameras. This paper presents three core technical contributions: (1) a construction of a social formation feature that is scale and orientation invariant inspired by electric dipole moments; (2) a method of discovering multiple social groups using scale space extrema in a spatial distribution of social members; (3) a consolidation of social interaction data reconstructed in difference scenes, which allow us to learn and infer social saliency in a unified coordinate system. Our method can predict social saliency from a third person video or image of social interactions.

2. Related Work

A social formation in social interactions is characterized by the geometric configuration of people. For instance, a circular formation is created by the audiences around a street busker while an equilateral triangular formation is often observed in triadic interactions. Kendon's characterization [14] of social formations based on his F-formation theory states that a spatial distribution of social members evolves to afford equal accessibility. This characterization provides a computational template for modeling social interactions. Cristani et al. [7] identified social interactions in a crowd scene by fitting the location and orientation of social members into the F-formation template and Setti et al. [30] extended this framework to handle multiscale formations. Marshall et al. [21] analyzed how physical structures such as tables and chairs affect social formations and the efficiency of interactions. Choi et al. [5] generalized the template to model group activities. In addition to F-formation, other computational representations such as context [1, 6, 16, 27, 28] and proxemics [4, 8, 32] have been also used to detect social interactions and activities.

Social formations and location of joint attention are mutually dependent on each other according to the F-formation theory, and empirically measuring their relationship is essential to computationally model and understand social behaviors. Two main approaches are presented towards measuring their relationship: third person approach and first person approach. With a third person view, Hoffman et al. [11] combined visual saliency and gaze directions to detect joint attention and Marín-Jiménez et al. [20] found people looking at each other from rough estimates of gaze directions obtained by HOG features. Prabhakar et al. [24] learned the causal relationship that propagates through joint attention in turn-taking interactions. Ousley et al. [22] and Regh et al. [26] presented a multimodal dataset of dyadic interactions with human annotated joint attention in the course of a child autism assessment. Such annotations facilitated feature extraction from multimodal interaction signals and spatial and temporal recognition of joint attention.

Detecting accurate gaze directions in a third person view is a key challenge in measuring joint attention. While head localization has shown impressive performance via a cascade detector [29], estimating the head orientation from a third person view is unreliable; state-of-the-art face detection frameworks [3, 20, 29] can produce only limited degree of accuracy.

First person cameras observe other faces at a short distance thus allowing estimation of gaze directions of people around [17, 18]. Pusiol et al. [25] investigated the correlation between joint attention and a care-giver's location with respect to a child's first person camera. Fathi et al. [9] employed face detection from a first person view to infer the relative depth and orientation of social members with respect to the wearer. Park et al. [23] exploited structure from motion to localize first person cameras in a unified coordinate system and triangulate joint attention in 3D. This method is not biased by viewpoint and thus, produces accurate measurements of joint attention. Using the joint attention measurements, Arev et al. [2] showed an automatic video editing tool for multiple first person cameras.

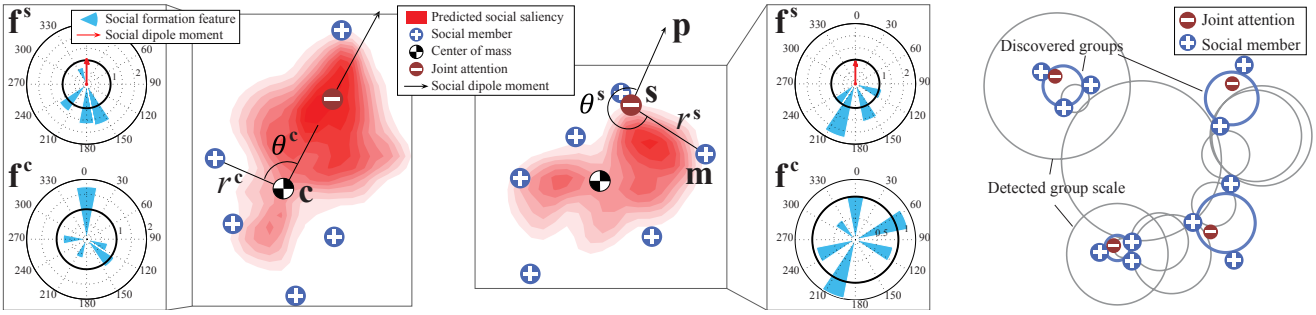
On the modeling side, our approach is to use first person view to obtain in-situ measurements of both social formation and joint attention in 3D. We leverage a large collection of social interaction data reconstructed by the first person cameras in 3D [23] to learn the geometric model between joint attention and its social formation. Once this model is learned, we applied it to a general third person video or image. Note that unlike all previous approaches, our method predicts the location of joint attention using the social formation without the need for gaze measurements.

3. Social Saliency Prediction

We predict social saliency, the likelihood of joint attention, using a *social formation feature* that is designed to capture the geometric relationship between joint attention and its members. We learn this relationship from the social interaction data described in Section 5.

3.1. Social Formation Feature

Let $\mathbf{m}_i$ denote the location of the i^{th} social member engaging to $\mathbf{s}$ joint attention, i.e., the gaze direction of the member is oriented towards $\mathbf{s}$. Throughout this paper, we use the 2D configuration of scenes by projecting onto the ground plane, i.e., $\mathbf{m}, \mathbf{s} \in \mathbb{R}^2$, for computational efficiency but the 3D configuration can be applied without any modification. We represent a social formation using the social



(a) Social dipole moment (b) Social group detection

Figure 2. (a) We represent a social formation using the social dipole moment that measures the spatial distribution of social members with respect to joint attention. This social dipole moment provides a scale and orientation invariant description of the social formation. We leverage this representation to predict social saliency based on their locations. (b) We detect social groups using scale space extrema (gray circles) on their spatial distribution. The detected memberships are used to optimally compute the social groups (blue circles) by solving Equation (2).

dipole moment, $\mathbf{p}$, inspired by electric dipole moments[2]:

$$\mathbf{p} = \mathbf{s} - \frac{1}{|\mathcal{S}|}\sum_{i\in\mathcal{S}}\mathbf{m}_i = \mathbf{s} - \mathbf{c}, \tag{1}$$

where $\mathbf{c}$ is the center of mass of the social members, $\mathbf{c} = \sum_{i\in\mathcal{S}}\mathbf{m}_i/|\mathcal{S}|$, and $\mathcal{S}$ is the set containing the indices of the social members engaging to $\mathbf{s}$. Note that the social dipole moment is normalized by the number of members whereas the electric dipole moment is not.

This social dipole moment characterizes the location of joint attention with respect to the first statistics of the social spatial distribution—the center of mass of social members. The direction of the social dipole moment indicates the dominant gaze direction of the group and the magnitude measures the alignment of their gaze directions. For instance, the direction of the social dipole moment of a side-by-side dyadic formation is oriented towards what they roughly look at with large magnitude as their gaze directions are well aligned.

We represent a spatial distribution of social members using a social formation feature, $\mathbf{f}$, that captures the geometric relationship between joint attention and its members. The social formation feature is represented by $\mathbf{f} = \begin{bmatrix}\mathbf{f}^{\mathbf{s}\mathsf{T}} & \mathbf{f}^{\mathbf{c}\mathsf{T}}\end{bmatrix}^{\mathsf{T}}$ where $\mathbf{f}^{\mathbf{s}}$ and $\mathbf{f}^{\mathbf{c}}$ are the spatial features centered at joint attention and the center of mass, respectively. The k^{th} element of these features is defined as:

$$\mathbf{f}^{\mathbf{s}}_k = \frac{1}{J^{\mathbf{s}}_k}\sum_j^{J^{\mathbf{s}}_k}\bar{r}^{\mathbf{s}}_j \quad \text{for } \theta_k \le \theta^{\mathbf{s}}_j < \theta_{k+1}$$
$$\mathbf{f}^{\mathbf{c}}_k = \frac{1}{J^{\mathbf{c}}_k}\sum_j^{J^{\mathbf{c}}_k}\bar{r}^{\mathbf{c}}_j \quad \text{for } \theta_k \le \theta^{\mathbf{c}}_j < \theta_{k+1}$$

where $J^{\mathbf{s}}_k$ and $J^{\mathbf{s}}_k$ are the number of members belonging to the k^{th} angular bin. $\bar{r}^{\mathbf{s}}_j = \|\mathbf{s} - \mathbf{m}_j\|/\bar{r}$ and $\bar{r}^{\mathbf{c}}_j = \|\mathbf{c} - \mathbf{m}_j\|/\bar{r}$ are normalized distance by average distance to the center of mass, i.e., $\bar{r} = \sum_{i\in\mathcal{S}}\|\mathbf{c} - \mathbf{m}_i\|/|\mathcal{S}|$. We also normalize the angle of each member based on the direction of the social dipole moment, i.e., $\theta^{\mathbf{s}}_j = \angle(\mathbf{m}_j - \mathbf{s}) - \angle\mathbf{p}$ and $\theta^{\mathbf{c}}_j = \angle(\mathbf{m}_j - \mathbf{c}) - \angle\mathbf{p}$.

Figure 2(a) illustrates the social formation features generated by two groups. Note that the scale and orientation of the features are normalized by $\bar{r}$ and $\angle\mathbf{p}$, respectively. Different locations of joint attention given a social formation yield different features, $\mathbf{f}^{\mathbf{s}}$ and $\mathbf{f}^{\mathbf{c}}$, due to angular normalization. This joint attention centric representation is designed to capture a geometric pattern of formation as a function of joint attention.

A social formation feature is scale and orientation invariant as it is normalized accordingly, which allows us to directly learn and infer their relationship from diverse formations across different scenes. Note that each social formation has the unique number, scale, and orientation of social members and this representation transforms them to a canonical form.

3.2. Prediction

Given social formation features extracted from the first person social interaction data described in Section 5, we learn the geometric relationship between joint attention and its members and predict social saliency of a target scene. We train a binary ensemble classifier from a collection of social formation features of the interaction data. 16 angular bins are used to represent $\mathbf{f}^{\mathbf{s}}$ and $\mathbf{f}^{\mathbf{c}}$. These features constitute the positive training data. We also generate negative training data by randomly sampling points that retain a distance from joint attention, i.e., $\|\mathbf{s}_n - \mathbf{s}\| \ge \epsilon_n$ where $\mathbf{s}_n$ is a negative sample point for joint attention, $\mathbf{s}$. The social formation features for these negative sampled points are extracted and used as negative training data. To encode the magnitude of social dipole moment, we discretize $\|\mathbf{p}\|$ into 50 magnitude bins and train an AdaBoost [10] classifier per bin. We experimented with several discriminative classifiers (KNN, Linear SVM, Random Forests, and AdaBoost) and compare top two classifiers (AdaBoost and Random Forest) in Section 6.1.

[2]The electric dipole moment of a molecule of water, H_2O, is 1.85 D due to uneven charge distribution, i.e., hydrogen atoms form 104.48° angle with respect to oxygen atom.

In the prediction stage, we compute the social formation feature of the target scene based on the estimated membership described in Section 4. We predict the binary label of discrete locations in the target scene with the trained classifier that falls into the same magnitude bin. We generate a continuous social saliency map by convolving with a Gaussian kernel. The resulting social saliency map and the ground truth locations of joint attention are shown in Figure 2(a). High social saliency forms near the ground truth locations of joint attention.

4. Social Group Discovery

In social scenes, multiple groups with diverse formations arise simultaneously from dyadic interactions to crowd interactions. To predict social saliency using a social formation feature presented in Section 3, the group detection must be carried out to isolate each social group. In this section, we present a method to identify the membership of social groups based on the locations of members. Note that all previous work [5, 7, 9, 23] discovered social groups based on the positions and orientations of members whereas we use the positions only.

We observe that social members often form a circular and coherent shape as shown in Figure 2(b). Also if the space allows, two groups do not tend to overlap each other because that would interfere their interactions. We encode such properties to detect the multiple groups based on the locations of the members.

4.1. Scale Space of Social Formation

We find candidates of social groups inspired by a scale space representation in signal processing. This representation allows us to discover circular and coherent structures formed by the spatial distribution of the members in a scene. Given the locations of the members, $\{\mathbf{m}_i\}$, we convolve their spatial distribution with a Gaussian kernel, $G(\mathbf{x};\sigma)$, where each member is modeled by the Dirac delta function, $\delta(\mathbf{m}_i)$:

$$L(\mathbf{x};\sigma) = \int_A G\left(\mathbf{x}-\mathbf{m}_i;\sigma\right) * \delta(\mathbf{m_i})d\mathbf{x} = \sum_{i=1}^{N} G(\mathbf{m}_i;\sigma).$$

L is the convolution between the spatial distribution of social members with the Gaussian kernel. We find the local extrema in the scale space approximated by the difference of Gaussians, $D(\mathbf{x},\sigma) = L(\mathbf{x},k\sigma) - L(\mathbf{x},\sigma)$ [19]. These extrema reflect underlying shape structures in different scales. The detected scale space extrema comprises the set of candidate social groups, $\mathcal{G}$, in the scene.

Each element of the set of candidate social groups, $\mathfrak{G} \in \mathcal{G}$, is represented by a triple, $\mathfrak{G} = \{\mathbf{g}, k, \mathcal{M}\}$ where $\mathbf{g}$ and k are the detected location and scale, and $\mathcal{M}$ is a set containing indices of its membership, i.e., $i \in \mathcal{M}$ if the i^{th} member belongs to the social group. Each member is assigned to groups by measuring the influence of the detected scale, i.e., $i \in \mathcal{M}$ if $G\left(\frac{\mathbf{g}-\mathbf{m}_i}{k\sigma};1\right) > \epsilon$.

4.2. Social Groups Detection

Given a set of social group candidates, $\mathcal{G} = \{\mathfrak{G}_j\}$, we find the minimal subset $\mathcal{G}^*$ that covers all members and has the desired properties between groups as noted above. This is equivalent to the set cover problem [13] that finds the minimum number of sets whose union constitutes the entire set. We modify the set cover problem to include the intergroup repulsive force [5] to retain no overlapping groups. We also penalize the double counted social members by multiple groups; each member must belong to no more than one group and therefore, groups do not overlap each other. The modified set cover problem is formulated as:

$$\begin{aligned} \underset{\mathbf{y},\boldsymbol{\xi}_1,\boldsymbol{\xi}_2}{\text{minimize}} \;& \frac{1}{2}\mathbf{y}^{\mathsf{T}}\mathbf{Q}\mathbf{y} + \lambda_{\mathbf{y}}\mathbf{1}^{\mathsf{T}}\mathbf{y} + \lambda_{\boldsymbol{\xi}}\left(\mathbf{1}^{\mathsf{T}}\boldsymbol{\xi}_1 + \mathbf{1}^{\mathsf{T}}\boldsymbol{\xi}_2\right) \quad (2) \\ \text{subject to} \;& \begin{array}{l} \mathbf{V}\mathbf{y} \geq \mathbf{1} - \boldsymbol{\xi}_1, \;\; \mathbf{V}\mathbf{y} \leq \mathbf{1} + \boldsymbol{\xi}_2 \\ \mathbf{y} \in \{0,1\}^Y, \;\; \boldsymbol{\xi}_1, \boldsymbol{\xi}_2 \in \{0,1\}^P \end{array}, \end{aligned}$$

where $\mathbf{y}$ is a binary indicator vector of the group candidates, i.e., $\mathbf{y}_j = 1$ if $\mathfrak{G}_j \in \mathcal{G}^*$ and zero otherwise. $\boldsymbol{\xi}_1$ and $\boldsymbol{\xi}_2$ are slack variables allowing social outliers who do not belong to any social group and double counted members, respectively. The number of social outliers and double counted members is minimized by $\mathbf{1}^{\mathsf{T}}\boldsymbol{\xi}_1$ and $\mathbf{1}^{\mathsf{T}}\boldsymbol{\xi}_2$ in the objective function. $\mathbf{Q}$ captures the pairwise relationship defined by the intergroup distance, $\mathbf{Q}_{ij} = 1/\|\mathbf{g}_i - \mathbf{g}_j\|$ if $i \neq j$ and zero otherwise. This quadratic term ensures that joint attention of each group does not form too close each other assuming joint attention can be approximated by the center of the circumcircle of the formation. A quantitative evaluation on this approximation will be discussed in Section 6.1. $\mathbf{1}^{\mathsf{T}}\mathbf{y}$ minimizes the number of groups. $\lambda_{\mathbf{y}}$ and $\lambda_{\boldsymbol{\xi}}$ control the balance between the objectives. $\mathbf{V} \in \{0,1\}^{P\times Y}$ is a binary matrix that indicates whether the i^{th} member belongs to the j^{th} candidate, i.e., $\mathbf{V}_{i,j} = 1$ if $i \in \mathcal{M}_j$, and zero otherwise. Therefore, the i^{th} element of $\mathbf{V}\mathbf{y}$ counts the number of times that the i^{th} member is included in the optimally estimated set $\mathcal{G}^*$. P and Y are the number of detected groups and members, respectively.

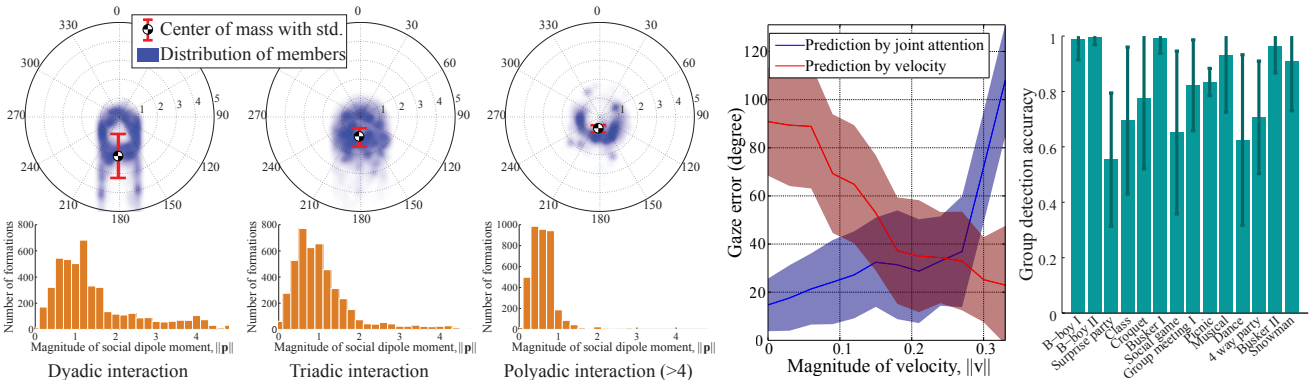
This optimization jointly finds the minimal set of social groups, $\mathcal{G}^*$, and social outliers. In Figure 2(b), we illustrate the optimally estimated social groups (blue circles) based on Equation (2) from all candidates (gray circles). The detected social groups coincide with membership of the ground truth joint attention.

Solving Equation (2) is NP-complete as it inherits from the set cover problem [13]. We solve this optimization using the commercial optimization software, Gurobi[3] that employs a branch and bound method.

5. Social Interaction Data

We measure joint attention and the locations of associated members via 3D reconstruction of first person cameras [23]. We manually synchronize cameras and recon-

[3] http://www.gurobi.com



(a) Formation analysis (b) Velocity analysis (c) Group detection accuracy

Figure 3. (a) We characterize social interactions based on the F-formation theory. The top row shows the spatial distribution of social members with respect to social formation features. The bottom row illustrates the histograms of social formations given the magnitude of social dipole moment that measures distance between joint attention and the center of mass. (b) We study a relationship between motion speed and joint attention. As the speed increases, the velocity direction becomes a strong predictor of gaze directions. (c) We detect the candidates of social groups using scale space extrema and optimally find the groups by solving Equation (2).

struct all first person cameras in a unified 3D coordinate system via the standard structure from motion pipeline. The fixed relationship between a camera and gaze direction is calibrated by a predefined sequence. We represent a gaze direction using a cone shaped gaze model and superimpose all gaze models to produce a social saliency field. The modes of the social saliency field that corresponds to joint attention are estimated using a meanshift algorithm [23] that automatically determines the number of joint attention. We estimate the ground plane via a plane RANSAC and project the locations of joint attention and its social members to the ground plane. The social formation feature is extracted in the projected locations.

Our dataset consists of 20 social interaction scenes that were partially collected by Park et al. [23] and Arev et al. [2][4]. We add 9 more scenes (B-boy II, B-boy III, B-boy IV, Class, Busker II, Card game, Hide and seek, Collaborative building, and Group meeting II). Three scenes (Card game, Hide and seek, and Collaborative building) captured triadic interactions between children aged 5-6. The entire dataset contains a total of 49,490 social formations.

We also collect data for basketball games where team players strategically take advantage of team formations. For each basketball data, we register all 3D reconstructions into an NBA standard court (dimension: 94 feet by 50 feet). This registration allows us to learn the geometric relationship directly in the canonical coordinate system. Two types of games were captured: one with amateur players and one with university team players directed by a professional coach. The entire dataset contains a total of 140,028 formations. The summary of the basketball dataset is listed in Table 1.

[4]A few scenes were captured by hand-held cameras that still follow the gaze behaviors of the camera operators.

Scene	N	T	F	Basketball	N	T	F
B-boy I	18	105	317	Amateur 1	10	1380	6750
B-boy II	18	450	1351	Amateur 2	8	900	4199
B-boy III	18	160	528	Amateur 3	10	1740	35843
B-boy IV	18	50	180	Univ. team 1	9	2516	25138
Surprise party	11	120	2227	Univ. team 2	10	2609	23150
Class	11	360	3590	Univ. team 3	10	2853	25335
Croquet	6	300	6000	Univ. team 4	10	2186	19613
Busker I	6	120	3566				
Busker II	6	180	5394				
Card game	3	180	768				
Hide and seek	3	180	214				
Block building	3	700	2702				
Social game	8	450	2086				
Meetings I	11	120	832				
Meetings II	5	44	1120				
Picnic	6	60	965				
Musical	7	180	2184				
Dance	6	180	5301				
4 way party	11	180	1909				
Snowman	4	753	8256				

Table 1. First person social interaction data (N: the number of members; T: duration (sec); F: the number of formations)

5.1. Data Analysis

Based on the social interaction data, we characterize social formations evolving in natural interactions. According to the F-formation theory, the social space evolves to afford equal accessibility to all social members. In Figure 3(a), we quantitatively identify such space from dyadic, triadic, and polyadic (more than four members) interactions. The top row illustrates the distribution of social members with respect to the social formation feature, $\mathbf{f^s}$, i.e, the joint attention located at the center and the spatial distribution of social members is normalized by its scale and orientation. For dyadic interactions, joint attention primarily forms near one of two in vis-a-vis interactions; many members are co-located with joint attention. The L-shape or side-by-side interactions are captured as the two downward parallel tails in the graph, i.e., joint attention forms out of the line connecting the two. This distribution changes as the number of interacting people increases. In the polyadic interactions, they form a circular shape around joint attention that affords

to equal accessibility to social members; few members are located at the center. In the bottom row, we show the histogram of social formations given the magnitude of a social dipole moment, $\|\mathbf{p}\|$, that measures the distance between joint attention and the center of mass. As the number of interacting people increases, the distance becomes shorter with low variance (see the standard deviation of the center of mass in the top row). This indicates that people form a circular formation centered at their joint attention, which affords equal accessibility and this analysis quantitatively confirms the F-formation theory.

In Figure 3(b), we show a predictive power of joint attention when social members are dynamic. Joint attention is a strong predictor of gaze directions of social members when they are quasi-static. Once they start to move, their gaze directions tend to deviate from the joint attention, i.e., they less likely look at what others are looking at but tend to align with the directions of their motion. This analysis concurs with a study on time scale dependency of visual perception [12].

6. Result

We apply our method to predict social saliency in real-world social scenes by leveraging the social interaction data captured by first person cameras. For the quantitative evaluation, we use the locations of joint attention in the data as the ground truth. For the qualitative evaluation, we apply our method to third person videos.

6.1. Quantitative Evaluation

In this section, we quantitatively evaluate our method in two criteria: group detection and social saliency prediction.

Group detection We detect social groups by solving a binary quadratic programming in Equation (2). The detected social groups are compared with the ground truth social groups obtained by the first person camera data. Accuracy of the group detection is defined by $A = \frac{1}{N_g}\sum_{i=1}^{N_g}\max_j\left\{\frac{n(\mathcal{M}_i^g\cap\mathcal{M}_j^d)}{n(\mathcal{M}_i^g\cup\mathcal{M}_j^d)}\right\}$ where $\mathcal{M}_i^g$ and $\mathcal{M}_j^d$ are the membership index set of the i^{th} ground truth group and the j^{th} detected group, respectively. $n(\cdot)$ counts the number of elements in the set and N_g is the number of ground truth groups. The mean accuracy of the group detection is 0.8169 with 0.0964 standard deviation for all first person camera data. In Figure 3(c), we show the mean accuracy of each scene with standard deviation (a few scenes are omitted as they are highly similar to other scenes). Accurate detection is achieved for the scenes with regular social interactions such as the Busker II, Musical, and Group meeting I. When a scene is chaotic such as the Surprise party, the detection accuracy is relatively low (53%).

Social saliency prediction We compare our method (SFF+Boosting) with four baseline methods: Random forests predictor with our social formation feature (SFF+RF), predictor using the center of circumcircle (CC), predictor using the center of mass (COM), and predictor with context feature (CF). We exclude the target scene when we train the classifiers. Note that no previous method used the locations of social members to predict joint attention, i.e., a comparison with a baseline method without a trivial modification is not available. Instead, we compare with geometric predictors (CC and COM) and a predictor (CF) based on context features that were used in group activity recognition [6, 16].

A context feature (CF) is a member centric representation while a social formation feature is a joint attention centric representation. The context feature is computed by pooling the number of members in each angular and radial bin for each member. We train AdaBoost classifiers for the context features as described in Section 3.

We evaluate our prediction with the ground truth joint attention. We convolve the ground truth joint attention with a Gaussian kernel to produce the ground truth social saliency. We measure the area in a scene that corresponds to higher social saliency than a certain threshold. The true positive and false positive rate are computed by changing the threshold in the ground truth and predicted social saliency.

Figure 4 shows ROC curves for each predictor. The inset images illustrate the configuration of social members, joint attention, the center of mass, the center of circumcircle, and predicted social saliency from the top view of the scene. The mean average precisions[5] are listed in Table 2. The predictors based on our social formation feature (SFF+Boosting and SFF+RF) outperform other predictors. Note that the center of circumcircle predictor (CC) is also a strong predictor for joint attention when the group forms a circular formation such as the Dance and B-boy I scenes.

For basketball scenes, we modify our social formation feature and prediction procedure because (1) the scenes are registered in a canonical basketball dimension unlike other scenes and thus, we can exploit the location feature, and (2) the formation of joint attention is often dominated by the motion of players. We augment instant velocity of the center of mass (2 dimensional vector) on the social formation feature, which can handle missing data due to player detection failures. Also we discretize the canonical court with 94 by 50 grids and train a classifier for each grid independently. This discretization exhibits stronger discriminative power on the prediction and efficient computation. We compare the modified social formation feature with the original feature and the geometric predictors[6] as shown in Figure 4 and Table 2. The modified feature outperforms other representations with large margin, which indicates motion plays a pivotal role to localize joint attention.

[5] The absolute value of the mean average precision is dependent on sparsity of data points. For instance, the Picnic scene is fairly large comparing to the interaction area that results in low precision for all methods.

[6] We do not compare our method with the ROI detection work [15] because they require a temporal association of each player to estimate a motion field.

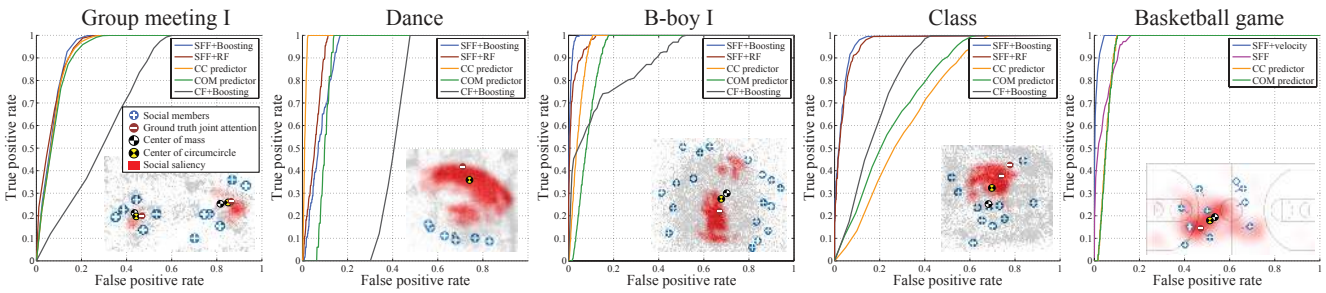


Figure 4. We compare our method with the geometric predictors and predictor using context feature. The center of circumcircle (CC) is a strong geometric predictor in regular social scenes. We also compare with the predictor using the center of mass (COM). The predictor based on the context feature (CF) which is a member centric representation is also compared. The inset image shows the configuration of social members, joint attention, the center of mass, the center of circumcircle, and predicted social saliency from the top view of the scene. For basketball scenes, we augment instant velocity of the center of mass on the social formation feature. This feature significantly improves the precision of the joint attention prediction.

Scene	SFF+Boosting	SFF+RF	CC	COM	CF
Dance	0.2769	0.1381	**0.3299**	0.0419	0.0106
Meeting I	0.2941	**0.3599**	0.2418	0.2350	0.0649
B-boy I	**0.7178**	0.6907	0.2078	0.1232	0.1225
Class	**0.7678**	0.7386	0.1445	0.2757	0.1873
Busker I	0.2919	0.2059	**0.3432**	0.1929	0.0103
Picnic	**0.1364**	0.1349	0.1115	0.1808	0.0244
Social game	**0.5425**	0.4419	0.3461	0.2463	0.0020
Scene	SFF+velocity	SFF	CC	COM	
Basketball	**0.7709**	0.1210	0.0987	0.0977	

Table 2. Mean average precision

6.2. Qualitative Evaluation

We apply our method on real-world examples involved with various social interactions. Given a video or a set of images, we reconstruct the scene in 3D using structure from motion. A main benefit of using the social formation feature is that it does not require directional measurements such as gaze directions where a sparse point cloud representation of humans can be used for prediction. We use a point cloud associated with heads identified by the head detector [20] to predict social saliency. The 3D reconstructed point cloud is projected to the ground plane and the projected point cloud is used to discover groups and predict social saliency.

Figure 1 and Figure 5(a) illustrate our results on social saliency prediction. We collect five social interaction scenes using a cellphone camera from third person view including the Halloween show, Cafeteria, Busker, Classroom, and Flash mob scenes. For all scenes, we overlay social saliency by projecting onto the ground plane. The configurations of the scenes from the top view are shown on the right column. Two social groups and three groups are detected in the Cafeteria and Flash mob scenes, respectively. From the Cafeteria and Classroom scenes, our prediction allows us to identify structure associated with social interactions such as the space near couches and podium. In the Busker scene, joint attention forms near the center of circumcircle where the busker was located. We also apply our method to YouTube videos captured at Time Square[7] and Louvre museum[8]. In the Time Square scene, our method correctly recognize the social space created by the photographers and subject with Muppet characters. Also we predict social saliency that forms around the Mona Lisa painting in the Louvre scene.

We also present our results on basketball scenes captured by third person view[9] in Figure 5(b). We register each frame to the canonical court. The modified social saliency feature (Section 6.1) is extracted by the locations of the feet of the players detected by [33]. The detected players and predicted social saliency are shown in the inset image and overlaid on the image. Our method correctly localizes social saliency in the presence of missing data.

7. Discussion

In this paper, we present a method to predict joint attention with a social formation feature that encodes the geometric relationship between joint attention and its members. We detect social groups using scale space extrema. We leverage the social interaction data captured by first person cameras that precisely measures joint attention to predict social saliency in real-world videos and images captured by third person views.

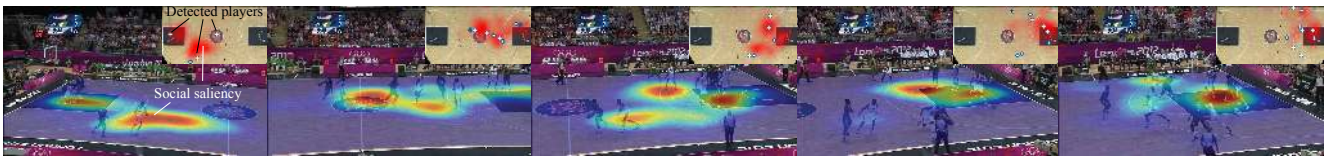
This work introduces a new way of using the data produced by first person cameras. We primarily use them in terms of *social statics* that describes how a social formation exerts a force on the joint attention. As demonstrated by the basketball scenes, motion is another driving force to form joint attention. One future direction is understanding *social dynamics* regarding motion, inertia, and stability of joint attention.

Figure 6. Prediction fails due to detection failure or unstructured formation.

Limitations Our method is primarily dependent on localization of social members. Failures on detection and unstructured formations cause erroneous prediction as shown in Figure 6.

[7]https://www.youtube.com/watch?v=ezyrSKgcyJw
[8]https://www.youtube.com/watch?v=VPjgsgLDu08
[9]https://www.youtube.com/watch?v=f6a3B499nwY

(a) Social interaction scenes



(b) Basketball scene

Figure 5. We apply our method to predict social saliency on third person videos. (a) We identify social space (space around the photographers and subject with the Muppet characters and space near Mona Lisa painting) in the Time Square and Louvre scenes obtained from YouTube. Multiple social groups are detected in the Cafeteria and Flash mob scenes. In the Busker and Classroom scenes, joint attention forms around the busker and podium. (b) We use the modified social formation feature to predict social saliency in a basketball game.

References

[1] A. Alahi, V. Ramanathan, and L. Fei-Fei. Socially-aware large-scale crowd forecasting. In *CVPR*, 2014. 2

[2] I. Arev, H. S. Park, Y. Sheikh, J. K. Hodgins, and A. Shamir. Automatic editing of footage from multiple social cameras. *SIGGRAPH*, 2014. 2, 5

[3] B. Benfold and I. D. Reid. Unsupervised learning of a scene-specific coarse gaze estimator. In *ICCV*, 2011. 2

[4] I. Chakraborty, H. Cheng, and O. Javed. 3D visual proxemics: Recognizing human interactions in 3D from a single image. In *CVPR*, 2013. 2

[5] W. Choi, Y. Chao, C. Pantofaru, and S. Savarese. Discovering groups of people in images. In *ECCV*, 2014. 2, 4

[6] W. Choi, K. Shahid, and S. Savarese. What are they doing? : Collective activity classification using spatio-temporal relationship among people. In *ICCV Workshop*, 2009. 2, 6

[7] 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 *BMVC*, 2011. 2, 4

[8] M. Cristani, G. Paggetti, A. Vinciarelli, L. Bazzani, G. Menegaz, and V. Murino. Towards computational proxemics: Inferring social relations from interpersonal distances. In *SocialCam*, 2011. 2

[9] A. Fathi, J. K. Hodgins, and J. M. Rehg. Social interaction: A first-person perspective. In *CVPR*, 2012. 2, 4

[10] Y. Freund and R. E. Schapire. A short introduction to boosting. 1999. 3

[11] M. W. Hoffman, D. B. Grimes, A. P. Shon, and R. P. Rao. A probabilistic model of gaze imitation and shared attention. *Neural Networks*, 2006. 2

[12] A. O. Holcombe. Seeing slow and seeing fast: two limits on perception. *Trends in Cognitive Sciences*, 2009. 6

[13] R. M. Karp. Reducibility among combinatorial problems. *Complexity of Computer Computations*, 1972. 4

[14] A. Kendon. *Conducting Interaction: Patterns of Behavior in Focused Encounters*. Cambridge University Press, 1990. 2

[15] K. Kim, D. Lee, and I. Essa. Detecting regions of interest in dynamic scenes with camera motions. In *CVPR)*, 2012. 6

[16] T. Lan, Y. Wang, W. Yang, S. Robinovitch, and G. Mori. Discriminative latent models for recognizing contextual group activities. *PAMI*, 2012. 2, 6

[17] Y. J. Lee, J. Ghosh, and K. Grauman. Discovering important people and objects for egocentric video summarization. In *CVPR*, 2012. 2

[18] Y. Li, A. Fathi, and J. M. Rehg. Learning to predict gaze in egocentric video. In *ICCV*, 2013. 2

[19] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 2004. 4

[20] M. Marín-Jiménez, A. Zisserman, M. Eichner, and V. Ferrari. Detecting people looking at each other in videos. *IJCV*, 2014. 2, 7

[21] P. Marshall, Y. Rogers, and N. Pantidi. Using F-formations to analyse spatial patterns of interaction in physical environments. In *CSCW*, 2011. 2

[22] O. Y. Ousley, R. Arriaga, G. D. Abowd, and M. Morrier. Rapid assessment of social-communicative abilities in infants at risk for autism. *Technical Report, Georgia Tech*, 2012. 2

[23] H. S. Park, E. Jain, and Y. Shiekh. 3D social saliency from head-mounted cameras. In *NIPS*, 2012. 2, 4, 5

[24] K. Prabhakar and J. M. Rehg. Categorizing turn-taking interactions. In *ECCV*, 2012. 2

[25] G. Pusiol, L. Soriano, L. Fei-Fei, and M. C. Frank. Discovering the signatures of joint attention in child-caregiver interaction. In *CogSci*, 2014. 2

[26] J. M. Rehg, G. D. Abowd, A. Rozga, M. Romero, M. A. Clements, S. Sclaroff, I. Essa, O. Y. Ousley, Y. Li, C. Kim, H. Rao, J. C. Kim, L. L. Presti, J. Zhang, D. Lantsman, J. Bidwell, and Z. Ye. Decoding children's social behavior. In *CVPR*, 2013. 2

[27] M. Rodriguez, J. Sivic, I. Laptev, and J.-Y. Audibert. Density-aware person detection and tracking in crowds. 2

[28] M. Rodriguez, J. Sivic, I. Laptev, and J.-Y. Audibert. Data-driven crowd analysis in videos. In *ICCV*, 2011. 2

[29] H. Schneiderman and T. Kanade. Object detection using the statistics of parts. *IJCV*, 2002. 2

[30] F. Setti, O. Lanz, R. Ferrario, V. Murino, and M. Cristani. Multi-scale F-formation discovery for group detection. In *ICIP*, 2013. 2

[31] A. Vinciarelli, M. Pantic, and H. Bourlard. Social signal processing: Survey of an emerging domain. *Image and Vision Computing*, 2009. 1

[32] Y. Yang, S. Baker, A. Kannan, and D. Ramanan. Recognizing proxemics in personal photos. In *CVPR*, 2012. 2

[33] Y. Yang and D. Ramanan. Articulated pose estimation using flexible mixtures of parts. In *CVPR*, 2011. 7