Social Saliency Prediction

Hyun Soo Park and Jianbo Shi
University of Pennsylvania

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.

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. 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.

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.

Social Formation Feature We represent a social formation using the social dipole moment, $\mathbf{p}$, inspired by electric dipole moments:

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

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 $\mathbf{s}$. This social dipole moment allows us to encode 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:

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

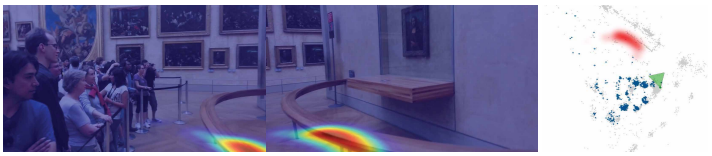
$$\mathbf{f}^{\mathbf{c}}_k = \frac{1}{J^{\mathbf{c}}_k}\sum_j^{J^{\mathbf{c}}_k} \bar{r}^{\mathbf{c}}_j \quad \text{for } \theta_k \le \theta^{\mathbf{c}}_j < \theta_{k+1}$$

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 $\theta^{\mathbf{c}}_j = \angle(\mathbf{m}_j - \mathbf{c}) - \angle\mathbf{p}$.

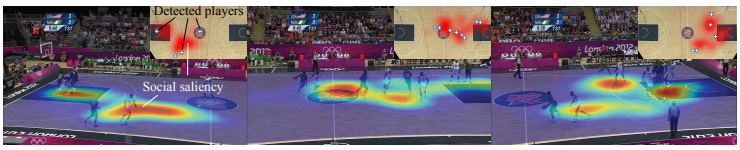
Social Saliency Prediction We learn the geometric relationship between joint attention and its members using the social formation feature and predict social saliency of a target scene. A binary ensemble classifier is trained by leveraging a collection of social formation features of the social interaction data. These data are generated by first person cameras that involve with social interactions. We measure joint attention and the locations of associated members via 3D reconstruction of first person cameras [1]. These datasets consist of 49,490 social formations for social interactions and 140,028 formations for basketball games.

Social Group Detection 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, the group detection must be carried out to isolate each social group. We present a method to identify the membership of social groups based on the locations of members. First, 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. Second, we find the minimal subset that covers all members and has the desired properties between groups. This is equivalent to the set cover problem 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 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.

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 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. The results from the Louvre and basketball scenes are shown in Figure 2. The heatmap indicates the predicted social saliency.



(a) Social interaction scenes



(b) Basketball scene

Figure 2: We apply our method to predict social saliency on third person videos. (a) We identify social space (space near Mona Lisa painting) in the Louvre scene. (b) We use the modified social formation feature to predict social saliency in a basketball game.

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