Predicting Primary Gaze Behavior using Social Saliency Fields *

Hyun Soo Park
Carnegie Mellon University
hyunsoop@cs.cmu.edu

Eakta Jain
Texas Instruments
e-jain@ti.com

Yaser Sheikh
Carnegie Mellon University
yaser@cs.cmu.edu

Abstract

We present a method to predict primary gaze behavior in a social scene. Inspired by the study of electric fields, we posit "social charges"—latent quantities that drive the primary gaze behavior of members of a social group. These charges induce a gradient field that defines the relationship between the social charges and the primary gaze direction of members in the scene. This field model is used to predict primary gaze behavior at any location or time in the scene. We present an algorithm to estimate the time-varying behavior of these charges from the primary gaze behavior of measured observers in the scene. We validate the model by evaluating its predictive precision via cross-validation in a variety of social scenes.

1. Introduction

Humans interact, in part, by transmitting and receiving social signals, such as gaze direction, voice tone, or facial expression [36, 48]. These signals convey the transmitter's attention, emotion, or intent, and enable us to perceive social context [1]. Thus, translating these signals into a form that machines can interpret is an important step towards enabling artificial agents to appropriately interact with people in social environments. In this paper, we focus on interpreting one such signal: the primary gaze direction, i.e., the ray emitted from the center of the eyes oriented towards the neutral eye gaze direction. This signal is a strong directional indicator of what an individual is attending to [9]—we tend to face what we are interested in.

Inspired by Coulomb's law, which describes the electrostatic interaction between charged particles, we present a model to describe the primary gaze behavior of individuals in a social scene. We posit "social charges" that attract the attention of the individuals in the scene, and we analyze the time-varying behavior of these charges (i.e., their emergence, transition, and dissolution). We characterize how information of the time-varying location and charge of multiple moving social charges is combined to induce a social saliency field analogous to an electric field. Under our model, this field is used to predict a distribution over primary gaze directions at any time and at any location in the scene.

Social charges are latent entities as their location or magnitude cannot be measured directly. However, using head-mounted or scene-mounted cameras, we can obtain measurements of the 3D head orientation of members in the scene. We present an algorithm that can use these measurements to infer the time-varying behavior of multiple social charges. We identify a charge by virtue of its membership (i.e., the set of members whose primary gaze direction is aligned to the same charge). We use this feature to also establish correspondence of social charges over time. The induced spatiotemporal field can then be used to predict the distribution of primary gaze direction at any proximal location or time.

Attentive behavior is an early indicator in the diagnoses of behavioral disorders (e.g., autism [5]). Predictive models of primary gaze behavior will enable anomaly detection and hold the promise of automated diagnoses and monitoring. Such models can also be used within a filtering framework to more effectively track primary gaze directions in a social scene. In augmented reality applications, predictive models of primary gaze behavior will enable the insertion of believable virtual characters into social scenes that respond to the social dynamics of a scene. Finally, such models can also be used in human-robot interaction scenarios to appropriately direct sensors and to limit the extent of the scene that the system needs to process and react to.

We validate our social charge model on four real world sequences where various human interactions occur, including a social game, office meetings, and an informal party. We evaluate our gaze prediction with ground truth data via a cross validation scheme against a baseline regression algorithm. Finally, we demonstrate the potential of gaze prediction as a prior for head tracking and anomaly detection.

Contributions. There are two core contributions in this paper: (1) a predictive field model of primary gaze direction based on the concept of social charges, inspired by electric field theory; (2) a method to estimate the latent behavior of the social charges in this model from observed primary gaze behavior in the scene, using the concept of membership features for establishing correspondences over time.

*http://www.cs.cmu.edu/~hyunsoop/socialcharge.html

2. Related Work

We review prior research on representing social scenes and predicting gaze directions.

2.1. Social Scene Representation

Models of social interaction have been proposed in psychology, sociology, and computer science. We categorize existing models for the social scene as spatial, temporal, or spatiotemporal models.

A social scene can be represented by a spatial arrangement of interactions. Two representations have been used to model a scene: microscopic and macroscopic representations. The seminal work by Hall [16] introduced the concept of proxemics, a categorization of human individual (microscopic) interactions based on spatial distances. Cristani et al. [8] applied proxemics to infer relationships of people in an image and Yang et al. [52] exploited the touch code in proxemics to estimate body poses of interacting people. A macroscopic representation was introduced by Kendon [23, 24] who modeled group spatial arrangements via F-formations. He showed that similar patterns of the group spatial arrangement (position and orientation of each member) are repetitively observed as the members in the group share their attention. Marshall et al. [30] further studied how a physical environment can affect F-formations. Cristani et al. [7] used the F-formation model to detect human interactions in a single image. A generalized F-formation concept has been applied to estimate social attention where gaze directions intersect in the scene [2, 10, 29, 34].

Time is another axis to represent a social scene because the social scene often includes dynamic human interactions. Each interaction at each time instant is associated with other interactions at different time instances. The causality test, introduced by Granger [14], is widely used as a measure of causality of two social interactions. Zhou et al. [53] directly applied the causality measure on a pair of trajectories of human activities to detect and recognize interactions. Prabhaker et al. [39] represented a video sequence using a multivariate point-process over activities and estimated the correlation between the activities via the causality measure. Parabhaker and Rehg [40] further extended their work to characterize temporal causality emerging in turn-taking activities. Instead of the causality measure, Gupta et al. [15] showed a method to learn a storyline graph structure and Wang et al. [49] modeled a quasi-periodicity measure to extract a repetitive pattern of activities in a social game.

The full dynamics of a social scene can be modeled by spatiotemporal representations. The social force model proposed by Helbing et al. [18] has successfully emulated crowd dynamics. Each individual experiences repulsive and attractive forces by neighbors and environments. The net force applied to the individual induces motion by Newtonian physics. Johansson et al. [22] and Pellegrini et al. [35] applied the social force model to track pedestrians in a video, and Mehran et al. [31] and Raghavendra et al. [41] detected abnormal events in a crowd scene. A similar social force concept has been used for distributed robot control [13,25,43]. Kim et al. [26] represented a dynamic scene with a dense motion field estimated by trajectories of individual players in sports and Wang et al. [50] tracked the ball using the gaze directions of players. Oliver et al. [33] integrated spatiotemporal behaviors into a coupled hidden Markov model to recognize a few types of interactions. Ryoo and Aggarwal [44] proposed a spatiotemporal feature to match between videos of interactions, Choi et al. [6] exploited spatiotemporal relationship of interactions to characterize the scene, and Ramanathan et al. [42] identify the social role via modeling time-varying interactions with a Conditional Markov Random field.

2.2. Gaze Prediction

A large body of research has studied the human ability to predict the gaze direction. Koch and Ullman [27] proposed a computational foundation for visual saliency detection. They modeled selective visual attention using a hierarchical structure of neurons that are sensitive to low level features such as color, orientation, motion, and disparity. This framework was implemented by Itti et al. [20] via their Winner-Take-All networks and showed that their detected visual saliency is matched with eye tracking results. In conjunction with the low level features, Sophie et al. [46] showed that faces are highly salient features in both static and dynamic scenes and adding face features improves the gaze prediction accuracy. However, this feature-based gaze prediction does not generalize to all scenes. Simons and Chabris [45] and Peter and Itti [37] showed that humans are blind to inattentional structure in a scene. When a task is involved, only task related locations are fixated and remembered regardless of visual saliency. For these scenes, data-driven approaches are more accurate predictors [28, 32]. Bernhard et al. [3] learned the gaze patterns from multiple game users to construct an importance map in a 3D game scene to predict visual attention of gamers at run-time. A detailed review of the feature based and data driven approaches can be found in [19].

Social saliency is another stimulus that drives attention. This social saliency states that you are likely to look at what others look at [38]. Friesen and Kingstone [11] showed that gaze is a strong social attention stimulus that can trigger attention shift. A study by Birmingham and Kingstone [4] further confirmed that gaze is more likely fixated to the eyes than any other stimuli such as low level features. Also the direction of the eyes influences the fixation points higher than other directional stimuli such as an arrow.

We present a novel predictive representation based on the concept of latent social charges for any 3D location or time, and validate it on real measurements of 3D gaze behavior.

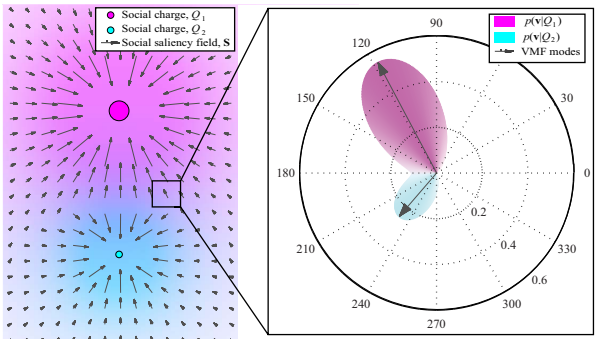


Figure 1. We model the relationship between a primary gaze direction and a social charge via a social saliency field inspired by Coulomb's law. The two social charges (the blue and green points) generate the social saliency saliency field on the left figure. The size of the social charges is proportional to social saliency. In the right figure, we show the probability distribution over gaze direction modeled by a mixture of von-Mises Fisher distributions in Equation (7).

3. Primary Gaze Behavior Prediction

A social *member* is a participant in a social scene in which multiple members interact with each other. Let $\mathbf{p}_j \in \mathbb{R}^3$ represent the center of the eyes of the j^{th} member and $\mathbf{v}_j \in \mathbb{R}^3$ represent the primary gaze direction, i.e., the ray emitted from $\mathbf{p}_j$ oriented towards the neutral gaze direction [21]. The set $\{(\mathbf{v}_j, \mathbf{p}_j)\}_{j=1}^J$ is the set of primary gaze directions and locations for the J members in the scene. Note that each $\mathbf{v}_j(t)$ and $\mathbf{p}_j(t)$ is time-varying, as the attention or location of each member can change over time. In this paper, we predict the primary gaze direction at any 3D location and time, given the observed gaze behavior of the members. We estimate $\mathbf{v}$ at any 3D location $\mathbf{p}$ in the scene, given $\{(\mathbf{v}_j, \mathbf{p}_j)\}_{j=1}^J$ by optimizing the following probability,

$$\mathbf{v}^* = \underset{\mathbf{v}}{\operatorname{argmax}}\, p\left(\mathbf{v}|\mathbf{p}, \{(\mathbf{v}_j, \mathbf{p}_j)\}_{j=1}^J\right), \tag{1}$$

where $p\left(\mathbf{v}|\mathbf{p}, \{(\mathbf{v}, \mathbf{p}_j)\}_{j=1}^J\right)$ is the probability of the primary gaze direction at $\mathbf{p}$ given the observed primary gaze directions, $\{(\mathbf{v}_j, \mathbf{p}_j)\}_{j=1}^J$.

One approach would be to directly regress $\mathbf{v}$ from $\{(\mathbf{v}_j, \mathbf{p}_j)\}_{j=1}^J$ [10, 26]. Inspired by Coulomb's law, we generatively model the relationship between primary gaze directions via latent *social charges* that drive attention of members in the scene — we show that this approach demonstrates superior predictive precision in the presence of missing and noisy measurements.

According to Coulomb's law, the force exerted on an electric charge due to the presence of another electric charge is directed along the line that connects these two charges. We represent a social charge as $Q = (q, \mathbf{r})$ where $q \in \mathbb{R}$ is a measure of social saliency, i.e., how strongly the social charge draws attention, and $\mathbf{r} \in \mathbb{R}^3$ is the 3D location of the charge. The decay of the spatial influence of the social charge is modeled as an inverse squared function (as with classic electric field model). A social charge is a quantity that changes over time because the scene includes dynamic human interactions. There may exist multiple social charges, $\{Q_i\}_{i=1}^I$ when multiple social groups are formed, where I is the number of the charges.

The social charge, Q_i is a latent quantity, i.e., it cannot be observed directly, and can only be estimated by its observed influence on the primary gaze direction of the members in the scene. Estimating the social charges given the primary gaze directions of the members is equivalent to optimizing the following likelihood,

$$\{Q_i^*\}_{i=1}^I = \underset{\{Q_i\}_{i=1}^I}{\operatorname{argmax}}\, L\left(\{Q_i\}_{i=1}^I | \{(\mathbf{v}_j, \mathbf{p}_j)\}_{j=1}^J\right). \tag{2}$$

This estimates the optimal $\{Q_i^*\}_{i=1}^I$ such that each observed primary gaze direction is oriented towards one of the social charges.

From these social charges, we can predict the most likely primary gaze direction at $\mathbf{p}$ by maximizing the following probability,

$$\begin{aligned}\mathbf{v}^* &= \underset{\mathbf{v}}{\operatorname{argmax}}\, p\left(\mathbf{v}|\mathbf{p}, \{Q_i^*\}_{i=1}^I, \{(\mathbf{v}_j, \mathbf{p}_j)\}_{j=1}^J\right) \\ &= \underset{\mathbf{v}}{\operatorname{argmax}}\, p\left(\mathbf{v}|\mathbf{p}, \{Q_i^*\}_{i=1}^I\right).\end{aligned} \tag{3}$$

Our social charge model assumes that $\mathbf{v}$ is conditionally independent on $\{\mathbf{v}_j, \mathbf{p}_j\}_{j=1}^J$ given $\{Q_i\}_{i=1}^I$. We discuss this assumption in more detail in the discussion.

We will develop a computational representation for the relationship between social charges and primary gaze directions to predict the primary gaze behavior via optimizing Equation (3) in Section 4. Based on the relationship, we present a method to estimate the latent social charges given primary gaze behaviors of the observers via optimizing Equation (2) in Section 5.

4. Social Saliency Field

In this section, we present a computational model that captures the relationship between time-varying social charges and primary gaze behavior. The charges induce a social saliency field that enables us to define a probability of the primary gaze direction given a location and time in Equation (3). Comparison between the social saliency field and electric field can be found in Table 1. The interacting

Electric charge	Social charge, $Q = \{q(t) \in \mathbb{R}_{\geq 0}, \mathbf{r}(t) \in \mathbb{R}^3\}$
Electric field, $\mathbf{E}$	Social saliency field, $\mathbf{S}(\mathbf{x}, t) \in \mathbb{R}^3$
$\mathbf{E}_{\text{net}} = \sum_i \mathbf{E}_i$	$\mathbf{S}_{\text{net}} = \max \mathbf{S}_i$

Table 1. Electric field vs. social saliency field

force between two charges, $Q = (q, \mathbf{r})$ and $Q_{\mathbf{x}} = (q_{\mathbf{x}}, \mathbf{x})$, from Coulomb's law is:

$$\mathbf{F} = K\frac{qq_{\mathbf{x}}\left(\mathbf{r} - \mathbf{x}\right)}{\|\mathbf{r} - \mathbf{x}\|^3}, \tag{4}$$

where K is a normalizing constant. The force between two charges is proportional to their magnitude of charges and

inversely proportional to squares of distance. When two charges have opposite polarities, the attractive force applies along the line that connects those two charges.

A point in space that attracts attention of members is represented as a negative charge — the more attractive the point, the higher negative charge. A member in the space is represented as an infinitesimal positive charge. We posit that a negative social charge, q, exerts an attractive force on a member (with an infinitesimal positive charge), along the line connecting the two charges $(\mathbf{r}-\mathbf{x})/\|\mathbf{r}-\mathbf{x}\|$, and with spatial influence decaying according to an inverse squared function, $\|\mathbf{r}-\mathbf{x}\|^{-2}$, as in Equation (4).

Analogous to the electric field, a social saliency field is defined by the limiting process,

$$\mathbf{S}(\mathbf{x}) = \lim_{q_{\mathbf{x}}\to 0} \frac{\mathbf{F}}{q_{\mathbf{x}}} = K\frac{q(\mathbf{r}-\mathbf{x})}{\|\mathbf{r}-\mathbf{x}\|^3}, \tag{5}$$

where $\mathbf{S}(\mathbf{x})$ is the social saliency field evaluating at $\mathbf{x}$, induced by a single social charge, $Q=(q,\mathbf{r})$.

When multiple electric charges exist, the net electric field induced by the charges are the superposition of the electric fields by all charges, i.e., $\mathbf{E}_{\text{net}} = \sum_{i=1}^{I}\mathbf{E}_i$ where $\mathbf{E}_{\text{net}}$ is the net electric field and $\mathbf{E}_i$ is the electric field generated by the i^{th} electric charge. Unlike the electric field, the net social saliency field selectively takes one of the social saliency fields[1], i.e.,

$$\mathbf{S}(\mathbf{x}) = \underset{\{\mathbf{S}_i(\mathbf{x})\}_{i=1}^{I}}{\operatorname{argmax}} \|\mathbf{S}_i(\mathbf{x})\|, \tag{6}$$

where $\mathbf{S}_i(\mathbf{x})$ is the social saliency field induced by the i^{th} social charge, $\mathbf{Q}_i$.

To reflect the selective gaze behavior, we model the underlying probability distribution of a primary gaze direction using a mixture of von-Mises Fisher distributions,

$$p\left(\mathbf{v}|\mathbf{x},\{Q_i\}_{i=1}^{I}\right) = \sum_{i=1}^{I}\pi_i\mathcal{V}\left(\mathbf{v}\,\middle|\,\frac{\mathbf{S}_i(\mathbf{x})}{\|\mathbf{S}_i(\mathbf{x})\|},\kappa\right), \tag{7}$$

where $\mathcal{V}$ is a von-Mises Fisher distribution[2] that accounts for eye-in-head motion and κ is a concentration parameter of the distribution. The mixture coefficients, $\pi_i = \|\mathbf{S}_i(\mathbf{x})\|/\sum_{k=1}^{I}\|\mathbf{S}_k(\mathbf{x})\|$, reflect the inverse squared function prior for the charges. Each von-Mises Fisher distribution measures the distance between the primary gaze direction, $\mathbf{v}$, and a unit vector from each social saliency field, $\mathbf{S}_i/\|\mathbf{S}_i\|$.

Each social charge may move independently depending on the primary gaze behavior of the participating group. A trajectory of a social charge, $Q(t)$, can be written as

$$Q(t) = \begin{cases} \{q(t),\mathbf{r}(t)\} & t_e \le t \le t_d, \\ \text{undefined} & \text{otherwise}, \end{cases} \tag{8}$$

where t_e and t_d are the emergence and dissolution time instances of the social charge. The charge is defined between the emergence and dissolution times, and otherwise the charge does not exist.

Given the saliency field from each charge at each time instant, the net time-varying saliency field can be written as

$$\mathbf{S}(\mathbf{x},t) = \underset{\{\mathbf{S}_i(\mathbf{x},t)\}_{i=1}^{I}}{\operatorname{argmax}} \|\mathbf{S}_i(\mathbf{x},t)\|. \tag{9}$$

5. Social Saliency Field Estimation

In this section, we present a method to estimate the time-varying location and magnitude of the social charges $\{Q_i(t)\}_{i=1}^{I}$, given the primary gaze directions of members, $\{(\mathbf{v}_j(t),\mathbf{p}_j(t))\}_{j=1}^{J}$, in the scene, i.e., maximize Equation (2). The data likelihood of Equation (2) can be rewritten by exploiting Equation (7) as

$$L\left(\{Q_i\}_{i=1}^{I}|\{(\mathbf{p},\mathbf{v})\}_{j=1}^{J}\right) = \prod_{j=1}^{J}\left(\sum_{i=1}^{I}\pi_i\mathcal{V}\left(\mathbf{v}_j\,\middle|\,\frac{\mathbf{S}_i(\mathbf{p}_j)}{\|\mathbf{S}_i(\mathbf{p}_j)\|},\kappa\right)\right). \tag{10}$$

Maximizing Equation (10) finds the optimal estimates of $\{Q_i\}_{i=1}^{I}$ that explain the observed primary gaze directions, $\{(\mathbf{v}_j,\mathbf{p}_j)\}_{j=1}^{J}$, given the number of social charges.

5.1. Expectation Maximization

An Expectation-Maximization (EM) algorithm allows us to solve this optimization problem. In the expectation step, we estimate the membership of each social charge given the social charge locations, i.e.,

$$\gamma_{ij} = \frac{\pi_i\mathcal{V}\left(\mathbf{v}_j\,\middle|\,\frac{\mathbf{S}_i(\mathbf{p}_j)}{\|\mathbf{S}_i(\mathbf{p}_j)\|},\kappa\right)}{\sum_{k=1}^{I}\pi_k\mathcal{V}\left(\mathbf{v}_j\,\middle|\,\frac{\mathbf{S}_k(\mathbf{p}_j)}{\|\mathbf{S}_k(\mathbf{p}_j)\|},\kappa\right)}, \tag{11}$$

where γ_{ij} is the probability that the j^{th} member looks at the i^{th} social charge. This also allows us to compute the social saliency $q_i = \sum_{j=1}^{J}\gamma_{ij}$, i.e., how many members focus on the social charge. In the maximization step, we estimate the social charge locations based on the membership, i.e.,

$$Q_i = \underset{\mathbf{r}_i}{\operatorname{argmin}} \sum_{j=1}^{J}\gamma_{ij}^2 d((\mathbf{v}_j,\mathbf{p}_j),\mathbf{r}_i)^2, \tag{12}$$

where $d(\cdot,\cdot)$ is a distance between a ray and point defined as follows,

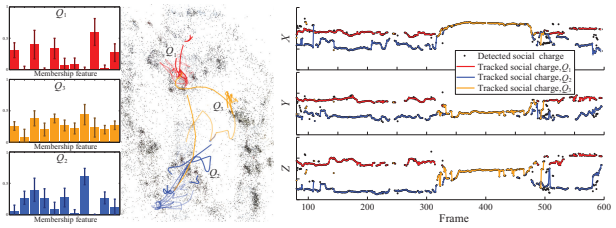
$$d((\mathbf{v},\mathbf{p}),\mathbf{x}) = \begin{cases} \frac{\|\mathbf{x}-\widehat{\mathbf{x}}\|}{\mathbf{v}^{\mathsf{T}}(\mathbf{x}-\mathbf{p})} & \text{for } \mathbf{v}^{\mathsf{T}}(\mathbf{x}-\mathbf{p}) \ge 0 \\ \infty & \text{otherwise}, \end{cases} \tag{13}$$

where $\widehat{\mathbf{x}} = \mathbf{p}+\mathbf{v}^{\mathsf{T}}(\mathbf{x}-\mathbf{p})\mathbf{v}$ is the projection of $\mathbf{x}$ onto the primary gaze direction. Equation (12) estimates the optimal location of Q_i, where the primary gaze directions that belong to Q_i intersect. This is equivalent to the triangulation of a 3D point given 2D projections [17].

For a time-varying social saliency field, we can modify the expectation and maximization steps in Equation (11) and (12) as follows:

[1] A primary gaze direction is not oriented towards an average location between two social charges but towards one of the charges.

[2] The von-Mises Fisher distribution is the nominal equivalent of the normal distribution over $\mathbb{S}^2$.



(a) Membership feature (b) Social charge tracking

Figure 2. (a) The membership feature reflects the participating members in a group. We temporally associate the detected charges based on the membership features. The membership features for Q_1 and Q_2 are complementary because the groups are formed in the same time. (b) The trajectories of the social charges are illustrated. Q_1 and Q_2 dissolve at frame 350 and reappear at frame 500. Our membership based tracking allows us to associate the temporally separated trajectories.

$$\mathrm{E}: \ \gamma_{ij} = \frac{\int_{t_e}^{t_d} \pi_i \mathcal{V}\left(\mathbf{v}_j \left| \frac{\mathbf{S}_i(\mathbf{p}_j)}{\|\mathbf{S}_i(\mathbf{p}_j)\|}, \kappa\right.\right) \mathrm{d}t}{\sum_{k=1}^{I} \int_{t_e}^{t_d} \pi_k \mathcal{V}\left(\mathbf{v}_j \left| \frac{\mathbf{S}_k(\mathbf{p}_j)}{\|\mathbf{S}_k(\mathbf{p}_j)\|}, \kappa\right.\right) \mathrm{d}t}, \tag{14}$$

$$\mathrm{M}: \ Q_i = \underset{\mathbf{r}_i}{\operatorname{argmin}} \int_{t_e}^{t_d} \sum_{j=1}^{J} \left(\gamma_{ij} d\left((\mathbf{v}_j, \mathbf{p}_j), \mathbf{r}_i\right)\right)^2 \mathrm{d}t + \lambda_g \mathbf{G}(\mathbf{r}_i),$$

where $\mathbf{G}(\cdot)$ is a temporal filter[3] that regularizes the temporal coherency of the social charge and λ_g is a weight on the filter term. Note that emergence and dissolution times, t_e and t_d, are the same for all social charges in Equation (14). In practice, we split the time windows such that the number of the social charges remains constant for each time window. This EM method requires prior knowledge of the number social charges and a good initialization of $\{Q_i\}_{i=1}^{I}$. In the following subsection, we will present a method to initialize EM.

5.2. Initialization

Detecting social charges in a static scene has been presented by Fathi et al. [10] and Park et al. [34]. While these two methods are complementary to each other, we make use of the method from Park et al. because it initializes the number of charges at each time instant automatically. They find local maxima of the distribution using a meanshift algorithm. We present a method to track the detected social charges across time based on membership features to initialize the EM algorithm.

Let $\mathbf{M}_i \in \mathbb{R}^J$ be a *membership feature* associated with each social charge. Each element of the membership feature indicates a probability that the j^{th} member belongs to the i^{th} social charge obtained by Equation (11), i.e.,

$$\mathbf{M}_i = \begin{bmatrix} \gamma_{i1} & \cdots & \gamma_{iJ} \end{bmatrix}^{\mathsf{T}}. \tag{15}$$

[3]For example, if discrete time instances are considered, $\mathbf{G}(\mathbf{r})$ can be $\sum_{t=t_e+1}^{t_d} \|\mathbf{r}(t) - \mathbf{r}(t-1)\|^2$ if one considers minimal displacement of the trajectory, or $\sum_{t=t_e+1}^{t_d-1} \|2\mathbf{r}(t) - \mathbf{r}(t-1) - \mathbf{r}(t+1)\|^2$ if one regularizes acceleration [47].

This membership feature enables us to describe a social charge in terms of the participating members.

The membership feature from a social charge remains a similar pattern across time because the same members tend to stay in their social clique as shown in Figure 2(a). We compute the membership features of all the detected social charges and cluster the charges using the classic meanshift algorithm [12] based on the features. The meanshift clustering enables us to label each charge across time instances. A set of the charges clustered by the same label forms a trajectory of a single social charge. When multiple charges at the same time instant are labeled in a single cluster, we choose the charge that is close to the center of the feature cluster.

The social charge representation via a membership feature enables us to track a social charge invariant to locations and time. A charge may move in 3D as long as the participating members remain the same. It can dissolve and re-emerge as the group disperses and re-unites, respectively. This introduces missing data because of temporary dissolution of the social charge as shown in Figure 2(b). Our tracking method can re-associate with the re-emerging charges based on the membership feature clustering because two temporally separated trajectories of the social charge have the same membership feature.

6. Results

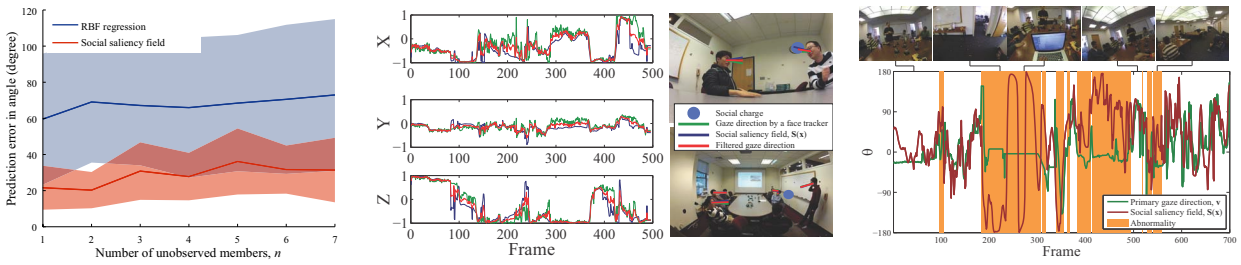
We validate our social saliency field model and evaluate the prediction accuracy, quantitatively and qualitatively via four real world sequences capturing various human interactions from third person and first person cameras[4].

6.1. Quantitative Evaluation

We validate our gaze prediction via a leave-one-out cross validation on the Meeting sequence provided by Park et al. [34]. In the scene, 11 people interact with each other by forming two subgroups. We leave out one of the members and estimate the time-varying social charges from the primary gaze behaviors of the rest of members. Using the estimated social charges, we evaluate the predictive validity of the left-out primary gaze direction. We run this cross validation scheme and measure the angle difference between the predicted gaze direction and the ground truth gaze direction. The mean error is 21.67 degrees with a standard deviation 15.73 degrees. In most cases, our prediction angle error is lower than 30 degrees which is within the range of eye-in-head motion.

We use a leave-k-out cross validation scheme to compare our gaze prediction against a field generated by Radial Basis Function (RBF) regression. This model was used by Kim et al. [26] to predict players' behaviors in soccer, which directly regresses from the observed directions to the predicted one. We randomly choose k number of members

[4]First person cameras refer to head-mounted or wearable cameras that produce video from the point of view of the wearer; third person cameras refer to infrastructure cameras in the scene looking at the social interaction.



(a) Comparison (b) 3D Face tracking (c) Anomaly detection

Figure 4. (a) We evaluate predictive validity using cross validation as the number of members decreases. Our social saliency model produces lower error with less standard deviation. (b) Our gaze prediction method can be used as a filter for a face tracking task. We exploit social charge motion estimated by other members to regulate the noisy face tracking process. (c) We detect anomalies in the scene based on social attention. A member who is not involved in any common social activity is classified as an outlier.

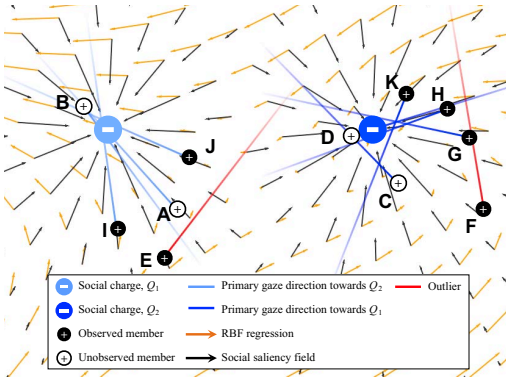


Figure 3. We compare our model against RBF regression [26]. We estimate the social charges from randomly chosen observed members (E to J) and predict the primary gaze directions of the unobserved members (A to D). The social saliency model shows superior predictive precision. For instance, the outlier E or F does not contribute to estimate the field while the RBF regression model produces inaccurate estimation at A. Also our model is insensitive to the spatial distribution of the observed members while the RBF prediction is not reliable at extrapolated points such as B, C, or D.

out of 11 members and predict their primary gaze directions using (11-k) number of the primary gaze directions. The orange vector field and dark gray vector field in Figure 3 are the RBF regression model and a social saliency field, respectively. The social saliency field outperforms over the RBF regression in three aspects: (1) The social saliency field is insensitive to outliers while the RBF regression is often biased by the outliers. For example, prediction at A is highly influenced by the outlier E, which results in inaccurate prediction. (2) The RBF model does not reflect selective gaze behavior. It produces a weighted average vector particularly at extrapolated area (see B, C, and D) that are not necessarily oriented towards a source of attention. (3) The magnitude of the field does not reflect the probability of primary gaze direction. The further from the social charges, the larger vectors that are formed. In Figure 4(a), we evaluate prediction error in angle produced by two methods by increasing k. This illustrates that our model produces approximately three times greater average accurate prediction.

6.2. Qualitative Evaluation

We applied our algorithm on two datasets of Park et al. [34]. Two sequences (Party and Meeting) are used to estimate the social saliency field as shown in Figure 5(c) and 5(d). These results are best seen in the supplementary video. As a proof-of-concept, we also applied our the social field model as prior within a simple filtering framework for tracking and for anomaly detection.

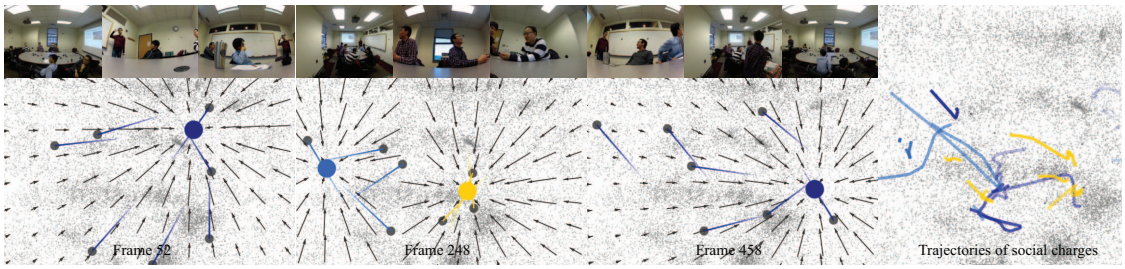
Tracking: We collected data from a meeting scene where 7 people including a presenter were engaged in a discussion. We instrumented 17 cameras in a meeting scene and calibrated the cameras using structure-from-motion. We used an off-the-shelf face detector (PittPatt) to find faces and IntraFace [51][5] to align and track the faces in 3D for each video. We also exploit the multiple cameras to register all the face poses in the same 3D coordinate frame. Face identity correspondences across the cameras were obtained by the PittPatt face recognition with manual refinement. Based on the tracked face poses, i.e., primary gaze directions, we generated a social saliency field as shown in Figure 5(a).

We used the social field as a naive prior for tracking 3D facial pose. We estimated social charge motion from other members and fused gaze prediction by the social saliency field with the face orientation estimate at each frame from the PittPatt system. We average out these two measurements to get the filtered direction. Figure 4(b) shows the noisy face tracking measurements[6] can be regularized by our gaze prediction.

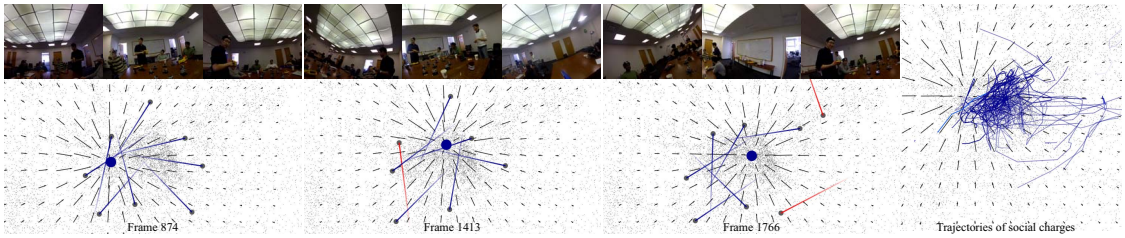
Anomaly Detection: We captured data of 8 people playing a social game called *Mafia*. GoPro Hero3 Black Edition cameras were mounted on each player and calibrated by structure from motion. While they interrogated each other during the game, the social charge stays in the group. Once

[5] http://www.humansensing.cs.cmu.edu/intraface/

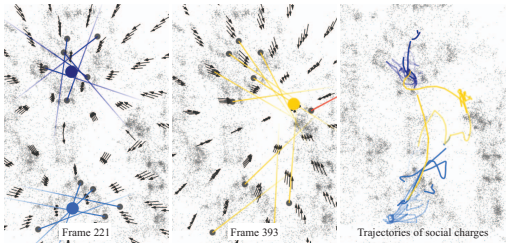
[6] Face pose tracking from a third person camera is noisy when the face is not directly oriented to the camera.



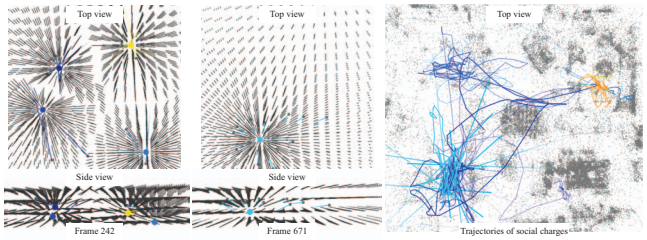
(a) Third person cameras



(b) First person cameras



(c) Meeting scene



(d) Party scene

Figure 5. We estimate a social saliency field from both third person cameras and first person cameras. (a) A social charge is formed at the presenter and splits into two subgroups at frame 248 in the meeting scene. The social saliency field reflects the selective gaze behavior. (b) 8 members in the scene play the social game called mafia with first person cameras. Our method can correctly detect anomalies (the red rays) based on social attention. We also apply our method to estimate a social saliency field on a public dataset provided by Park et al. [34].

a particular player is identified as a mafia, the player no longer stays in the group. In Figure 5(b), we estimate the social charge motion. In most cases, the social charge stays near the player who is investigated. Based on the social saliency field, we show that we can detect the outliers whose primary gaze direction does not behave in accord with social attention. This results in the detection of anomalous behavior, as shown in Figure 4(c). These outliers are the players who were identified as a mafia and were not participating in the game.

7. Discussion

To understand the social context of our everyday environment, humans extract tremendous meaning from subtle cues. Thus, to build perceptual systems that can similarly interpret human social interaction, the systems need to be equipped with internal models of social behavior that they can appeal to, when direct measurements from data is noisy or insufficient. The social saliency field model we present in this paper is a step towards this vision. By describing the activity in the scene in terms of the motion of latent social charges, we move beyond measuring scene activity, and towards understanding the narrative of the events of the scene, as interpreted by the members of the social group itself.

Summary. We present the social saliency field induced by the motion of social charges as a model to predict primary gaze behavior of people in a social scene. The motion of the charges is estimated from the observed primary gaze behavior of members of a social scene. The net social saliency field is created by selecting the maximum of a mixture of von-Mises Fisher distributions, each produced by a different social charge. We evaluate the predictive validity of spatial and temporal forecasting on real sequences and demonstrate that the social saliency field model is supported empirically.

Limitations. The principal assumption in the model is the conditional independence of gaze behavior between two observers given the behavior of the social charges. In practice, the gaze behavior of each observer in the scene is known to have a degree of influence on the gaze behavior of other observers [9, 38].

Future Work. In this paper, we limited our analysis to a

single social signal: primary gaze behavior. In future work, we will to extend this analysis to other social signals, such as facial expressions and body gestures, towards a coherent understanding of human behavior. An exciting future direction is investigating the predictive validity of the "sociodynamic" saliency field, obtained by differentiating the potential field over space *and* time. This field would include the influence of observer prediction of the behavior of social charges.

acknowledgment

This material is based upon work supported by the National Science Foundation under Grants No. 1353120 and 1029679. In addition, this research was supported by the Intel ISTC on Embedded Computing. We also thank Zijun Wei for the facial detection code and Natasha Kholgade for comments and suggestions on this work.

References

[1] R. Adolphs. The social brain: Neural basis of social knowledge. *Annual Review of Psychology*, 2009. 1

[2] L. Bazzani, D. Tosato, M. Cristani, M. Farenzena, G. Pagetti, G. Menegaz, and V. Murino. Social interactions by visual focus of attention in a three-dimensional environment. *Expert Systems*, 2011. 2

[3] M. Bernhard, E. Stavrakis, and M. Wimmer. An empirical pipeline to derive gaze prediction heuristics for 3d action games. *TAP*. 2

[4] E. Birmingham and A. Kingstone. Human social attention. *Brain Research*, 2009. 2

[5] K. Chawarska, S. Macari, and F. Shic. Decreased spontaneous attention to social scenes in 6-month-old infants later diagnosed with autism spectrum disorders. *Biological Psychiatry*, 2013. 1

[6] W. Choi, K. Shahid, and S. Savarese. Learning context for collective activity recognition. In *CVPR*, 2011. 2

[7] M. Cristani, L. Bazzani, G. Paggetti, A. Fossati, D. Tosato, A. Del Bue, G. Menegaz, and V. Murino. Social interaction discovery by statistical analysis of F-formations. In *BMVC*, 2011. 2

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

[9] N. J. Emery. The eyes have it: The neuroethology, function and evolution of social gaze. *Neuroscience and Biobehavioral Reviews*, 2000. 1, 7

[10] A. Fathi, J. K. Hodgins, and J. M. Rehg. Social interaction: A first-person perspective. In *CVPR*, 2012. 2, 3, 5

[11] C. K. Friesen and A. Kingstone. The eyes have it! reflexive orienting is triggered by nonpredictive gaze. *Psychonomic Bulletin and Review*, 1998. 2

[12] K. Fukunaga and L. D. Hostetler. The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Transactions on Information Theory*, 1975. 5

[13] R. Gayle, W. Moss, M. C. Lin, and D. Manocha. Multi-robot coordination using generalized social potential fields. In *ICRA*, 2009. 2

[14] C. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 1969. 2

[15] A. Gupta, P. Srinivasan, J. Shi, and L. S. Davis. Understanding videos, constructing plots learning a visually grounded storyline model from annotated videos. In *CVPR*, 2009. 2

[16] E. Hall. *The hidden dimension*. Doubleday New York, 1966. 2

[17] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2004. 4

[18] D. Helbing and P. Molnár. Social force model for pedestrian dynamics. *Physics Review E*, 1995. 2

[19] J. M. Henderson. Human gaze control during real-world scene perception. *Trends in Cognitive Sciences*. 2

[20] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *TPAMI*, 1998. 2

[21] R. S. Jampel and D. X. Shi. The primary position of the eyes, the resetting saccade, and the transverse visual head plane. head movements around the cervical joints. *Investigative Ophthalmology and Vision Science*, 1992. 3

[22] A. Johansson, D. Helbing, and P. K. Shukla. Specification of a microscopic pedestrian model by evolutionary adjustment to video tracking data. *Advances in Complex Systems*, 2008. 2

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

[24] A. Kendon. Spacing and orientation in co-present interaction. In *Development of Multimodal Interfaces: active Listening and Synchrony*, 2010. 2

[25] O. Khatib. Real-time obstacle avoidance for manipulators and mobile robots. In *ICRA*, 1985. 2

[26] K. Kim, M. Grundmann, A. Shamir, I. Matthews, J. Hodgins, and I. Essa. Motion fields to predict play evolution in dynamic sport scenes. In *CVPR*, 2010. 2, 3, 5, 6

[27] C. Koch and S. Ullman. Shifts in selective visual attention: towards the underlying neural circuitry. *Human Neurobiology*, 1985. 2

[28] O. Komogortsev and J. Khan. Perceptual attention focus prediction for multiple viewers in case of multimedia perceptual compression with feedback delay. In *ETRA*, 2006. 2

[29] M. Marin-Jimenez, A. Zisserman, and V. Ferrari. heres looking at you, kid. detecting people looking at each other in videos. In *BMVC*, 2011. 2

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

[31] R. Mehran, A. Oyama, and M. Shah. Abnormal crowd behavior detection using social force model. In *CVPR*, 2009. 2

[32] H. Murphy and A. T. Duchowski. Gaze-contingent level of detail rendering. In *Eurographics*, 2001. 2

[33] N. Oliver, B. Rosario, and A. Pentland. Graphical models for recognising human interactions. In *NIPS*, 1998. 2

[34] H. S. Park, E. Jain, and Y. Shiekh. 3D social saliency from head-mounted cameras. In *NIPS*, 2012. 2, 5, 6, 7

[35] S. Pellegrini, A. Ess, K. Schindler, and L. van Gool. Youll never walk alone: Modeling social behavior for multi-target tracking. In *ICCV*, 2009. 2

[36] A. S. Pentland. To signal is human. *American Scientist*, 2010. 1

[37] R. J. Peters and L. Itti. Applying computational tools to predict gaze direction in interactive visual environments. *ACM Transactions on Applied Perception*, 2007. 2

[38] M. I. Posner. Orienting of attention. *Quarterly Journal of Experimental Psychology*, 1980. 2, 7

[39] K. Prabhakar, S. Oh, P. Wang, G. D. Abowd, and J. M. Rehg. Temporal causality for the analysis of visual events. In *CVPR*, 2010. 2

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

[41] R. Raghavendra, A. Del Bue, M. Cristani, and V. Murino. Abnormal crowd behavior detection by social force optimization. In *HBU*, 2011. 2

[42] V. Ramanathan, B. Yao, and L. Fei-Fei. Social role discovery in human events. In *CVPR*, 2013. 2

[43] J. H. Reif and H. Wang. Social potential field: a distributed behavioral control for autonomous robots. *Robotics and Autonomous Systems*, 1999. 2

[44] M. S. Ryoo and J. K. Aggarwal. Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities. In *ICCV*, 2009. 2

[45] D. J. Simons and C. F. Chabris. Gorillas in our midst: Sustained inattentional blindness for dynamic events. *Perception*. 2

[46] M. Sophie, G. Nathalie, and P. Denis. Gaze prediction improvement by adding a face feature to a saliency model. *Recent Advances in Signal Processing*, 2009. 2

[47] G. Strang. The discrete cosine transform. *SIAM Reivew*, 1999. 5

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

[49] P. Wang, G. D. Abowd, and J. M. Rehg. Quasi-periodic event analysis for social game retrieval. In *ICCV*, 2009. 2

[50] X. Wang, V. H. Ablavsky, B. Shitrit, H. Beny, and P. Fua. Take your eyes off the ball: Improving ball-tracking by focusing on team play. *CVIU*, 2013. 2

[51] X. Xiong and F. De la Torre Frade. Supervised descent method and its applications to face alignment. In *CVPR*, 2013. 6

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

[53] Y. Zhou, S. Yan, and T. S. Huang. Pair-activity classification by bi-trajectories analysis. In *CVPR*, 2008. 2