

ricing social interactions in egocentric video. Our method makes it possible to capture a continuous record of an outing and then distill from it the most salient moments.

First-person video is an obvious choice for capturing personal day-long experiences in an amusement park, or other social events in which thousands of individuals participate. In this context, first-person video provides many advantages in comparison to fixed video recorders: (1) the first-person camera always records where the wearer is attending and provides natural videos of her family and friends, (2) occlusions are less likely in an egocentric setting, because the wearer naturally moves to provide a clear view and (3) it is not practical to simultaneously track all of the individuals in an amusement park and record all of their interactions using static cameras.

Our method uses two sources of information for analyzing the scene in order to detect social interactions: (1) faces and (2) first-person motion. The work-flow of our approach is shown in Fig 1. We transfer detected faces into the 3D scene and estimate their locations and orientations. The location of the faces around the camera wearer provides significant evidence for the type of social interaction. Further, the social interactions are characterized by the patterns of attention shift and turn-taking over time. We therefore estimate these patterns such as who looks at who, whether a group of individuals look at a common location, etc. We analyze these patterns of attention over time to recognize the type of social interaction. For example, when most of the individuals in a group are looking at a single person over a long period of time, our algorithm will label this as a monologue. In addition to patterns of face locations and attention, the head movements of the first-person provide additional useful cues as to their attentional focus.

We believe this is the first work to utilize egocentric video in order to detect and categorize social interactions among groups of individuals. Our focus on real-world social events, such as trips to an amusement park, make the task especially challenging due to the complex visual appearance of natural scenes and the presence of large numbers of individuals in addition to the social group of interest. We hope to encourage other researchers to tackle this challenging new problem domain, and we provide a large, extensively-annotated video dataset to support this goal. We have released our dataset at <http://cpl.cc.gatech.edu/projects/FPSI/>.

This paper makes four contributions: (1) we introduce a method for detection and analysis of social interactions such as monologue, discussion and dialogue, (2) we address this problem from the first-person point of view, which is crucial for capturing individual experience, (3) we present a dataset of 8 subjects wearing head-mounted cameras at a theme park, containing more than 42 hours of real world video and (4) we develop a method which estimates the pat-

terns of attention in video and analyzes these patterns over time to detect the social interactions.

2. Previous Work

We divide the previous work into three sections: (1) first-person wearable sensors, (2) social networks and (3) activity recognition.

2.1. First-Person Wearable Sensors

An early study of wearable cameras is reported in [19]. Recently there has been a growing interest in using wearable cameras, motivated by the advances in hardware technology. In our previous work [8, 9] we recognize daily activities such as meal preparation. Kitani et al. [11] recognize atomic actions such as turn left, turn right, etc. from first-person camera movement. Aghazadeh et al. [1] extract novel scenarios from everyday activities. In comparison, this is the first work that detects and recognizes social interactions in day-long videos recorded from a first-person vantage point.

Additional early work on experience capture using wearable cameras was conducted using SenseCam [10, 2]. For example, Gemmel et al. [10] present a lifetime recording system that takes images based on lighting change. Aris et al. [2] bind GPS information with photos taken over time to provide a search method using time and location.

2.2. Social Networks

There has been a recent interest in building the social network of individuals present in movies or other types of video using computer vision techniques. Choudhury [6] recovers the social network and patterns of influence between individuals. Yu et al. [20] use face recognition and track matching to associate people together in videos using an eigen vector analysis method which they call modularity-cut. Ding and Yilmaz [7] group movie characters into adversarial groups. In contrast to these works, our primary goal is to identify specific categories of social interaction and not estimate the social network structure for a group of individuals.

2.3. Activity Recognition

Human activity and action recognition is a popular topic in computer vision. Previous works have focused on recognizing atomic actions such as running, walking, etc. or more realistic actions performed by one or two individuals like opening the door, smoking and kissing [13]. More relevant are recent works that address the problem of recognizing group activities such as standing in line and crossing the street in images and videos. Lan et al. [12] use a discriminative latent SVM model to recognize group activities in images based on individual actions and pairwise context. Choi

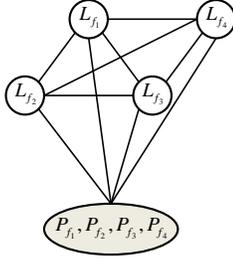


Figure 2: MRF model for inferring where each person is attending. The observations P_{f_i} contain the location and orientation of the face f_i in the scene, and the hidden variables L_{f_i} are the 3D location at which the face f_i is looking.

et al. [5] recognize group activities in videos using features which capture the relative location of pedestrians in space and time. Patron-Perez et al. [17] extract features for human interactions like hand shaking based on face orientation. Ni et al. [16] recognize group activities in surveillance videos from self, pair and group-localized causalities. Morariu and David [15] recognize multi-agent events in scenarios where structure is imposed by rules that agents must follow.

Our method differs in three ways from these works: (1) our videos are recorded from a first-person camera in which the bodies of other individuals are usually off-camera but faces and first-person head movement are easy to detect, (2) our focus is on categorizing extended social interactions such as conversations, and (3) we assign roles to individuals using patterns of attention and first-person movement. There are previous work that estimate where people are looking in the scene [3, 14]. However, our method goes beyond these works by showing that these attention patterns can be used for recognizing social interactions.

3. Faces and Attention

In this section, we describe our method for estimating the location and orientation of faces in space and analyzing the patterns of attention (to whom or where in 3D each face is looking). In Section 4, we analyze the attention patterns over time to detect the types of social interaction in video. We use faces as our main source of information because (1) faces and their attention patterns play a primary role in social interactions and (2) the state of the art computer vision methods for face detection and recognition are more robust in comparison to algorithms for detection of pedestrians or other objects.

Given only one person’s face location and orientation in the scene, we can estimate its line of sight but it is not possible to estimate where in space it is looking at. However, we show that the context provided by other faces can help to estimate where each faces is attending in space.

We start by tracking the faces in video. Then we identify individuals by clustering the face tracks into multiple bins. In addition, we compute the orientation (yaw, pitch,

roll) of every detected face¹. In each frame, we estimate the location of every face in 3D space with respect to the first-person. Since our videos are recorded from a linear scale fish-eye lens, we can estimate a face’s view angle θ from the camera by $\theta = \frac{r}{f}$ where r is the pixel distance of the center of the face from the image center and f is the camera’s focal length. We use the height h of a detected face to approximate its distance d from the camera by $d = \frac{c}{h}$ where c is a constant. We estimate c and f by calibrating our cameras, asking multiple subjects to stand at pre-defined locations and orientations with respect to a camera mounted on a tripod. We estimate the face orientations in 3D using its computed orientation in 2D image. Examples are shown in Fig 4(a-c).

Only a subset of individuals present in the scene are visible in each frame. This issue impacts the effectiveness of our attention estimation method. We solve this problem by building a map of faces around the first-person at local time intervals. Our assumption for making these local maps is that the positions of faces around the first-person does not significantly change locally in time. For each interval, we first pick the frame with the maximum number of faces as the reference frame. We initialize the 3D location of the faces in the reference frame. We set the origin of the world coordinate frame to the camera coordinate in the reference frame. We iteratively add the faces in adjacent frames to the map. For each frame, we match the faces to the ones already added to the map based on their assigned cluster number acquired in the recognition process.

The location and orientation of a face in 3D provides us with an approximate line of sight. We use the context provided by all the faces to convert lines of sight into 3D locations. We make three assumptions to achieve this goal: (1) It is more likely that a person looks at something in the direction of her face’s orientation, (2) a person looks at a person with a higher probability than at other objects, (3) if other people in the scene are looking at a particular location, then it is more probable for a face that is oriented towards that location to be looking at it as well. Next we describe our method for estimating where faces attend.

3.1. Reasoning about People’s Attention

Our goal is to find out where each person is attending in 3D space. We build an MRF (Fig 2) in which the observations P_{f_i} contain the location and orientation of the face f_i in the scene, and the hidden variables L_{f_i} are the 3D location at which the face f_i is looking. To make the inference feasible, we discretize the space into a grid at a resolution of $5cm \times 5cm$. Our goal is to estimate at which grid point each face is looking. The label space for each L_{f_i} is the set of grid locations. We have depicted an example in Fig 3.

¹We use Pittpatt software (<http://www.pittpatt.com>) for face detection and recognition.

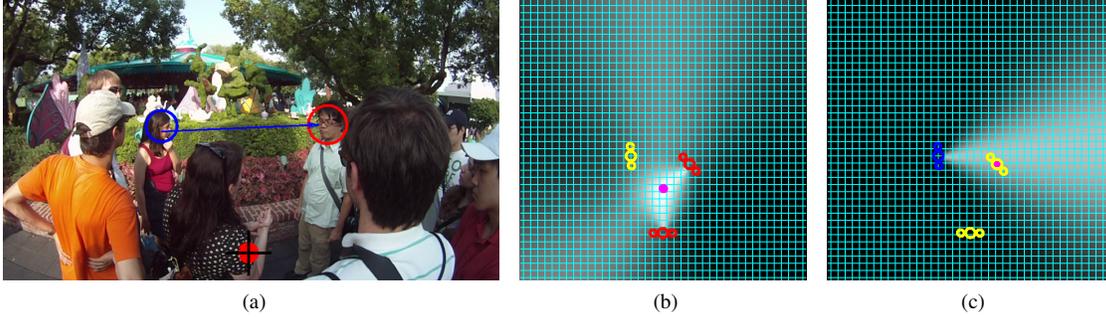


Figure 3: MRF Inference Procedure. Our method groups the faces looking at a common location together. In (a) the color of the circle around the face determines the group it belongs to. The camera wearer and the man on the right are looking at the lady wearing a polkadot shirt. In (b), our algorithm cannot detect lady’s face, but realizes that the first-person and the man are looking at the same location in space. In (c), our algorithm estimates that the lady with the red shirt is looking at the man.

Our MRF model in the case of four faces is shown in Fig 2. The unary potentials capture the likelihood of looking at a grid cell based on the observations, while the pairwise terms model the context between faces in the scene. The pairwise terms model the likelihood of looking at a grid cell given where other faces are looking.

Unary Potentials: Consist of three terms as follows:

$$\begin{aligned} \phi_U(L_{f_i}, P_{f_1}, P_{f_2}, \dots, P_{f_N}) &= \phi_1(L_{f_i}, P_{f_i}) \times \\ &\quad \phi_2(L_{f_i}, P_{f_i}) \times \\ &\quad \phi_3(L_{f_i}, P_{f_1}, \dots, P_{f_N}) \end{aligned}$$

where f_i represents face i in the scene, L_{f_i} is the location at which f_i is looking at in space, and $P_{f_i} = \begin{bmatrix} V_{f_i} \\ T_{f_i} \end{bmatrix}$ contains the orientation unit vector V_{f_i} and location vector T_{f_i} of the face f_i . The first potential ϕ_1 is modeled as a Gaussian function that computes the possibility of f_i looking at a location ℓ based on f_i ’s location and orientation in space:

$$\phi_1(L_{f_i} = \ell, P_{f_i}) = \frac{1}{\sigma_1 \sqrt{2\pi}} \exp \left\{ -\frac{\|V_{f_i} - \overline{(\ell - T_{f_i})}\|^2}{2\sigma_1^2} \right\}$$

where σ_1 is the standard deviation. The second potential ϕ_2 is modeled as a sigmoid function to put a threshold on how close $L_{f_i} = \ell$ can be to the face f_i , added mainly to avoid a face looking at itself:

$$\phi_2(L_{f_i} = \ell, P_{f_i}) = \frac{1}{1 + \exp \{-(c_2 \cdot \|\ell - P_{f_i}\|)\}}$$

where c_2 is a constant. Finally the third term ϕ_3 is meant to bias faces to look at where other faces are in comparison to looking at objects:

$$\phi_3(L_{f_i} = \ell, P_{f_1}, \dots, P_{f_N}) = \begin{cases} c_3 & \ell = P_{f_j} \forall j \neq i \\ 1 & \text{otherwise} \end{cases}$$

where c_3 is a constant increasing the chance of f_i looking at a location ℓ if another face f_j is at that location.

We set the parameters σ_1 , c_2 and c_3 using the training data. We manually annotate faces looking at each other in a set of frames and learn the parameters from these examples.

Pairwise Potentials: The binary potentials capture the interaction between people. They bias the faces towards looking at the same location in the scene. Basically, if others are looking at something in the scene, the probability that another person is looking at the same thing is higher. We define the following function for the binary potentials:

$$\phi_B(L_{f_i} = \ell_1, L_{f_j} = \ell_2) = \begin{cases} c_B & \text{if } (\ell_1 = \ell_2) \\ 1 - c_B & \text{if } (\ell_1 \neq \ell_2) \end{cases}$$

where c_B is a constant greater than $\frac{1}{2}$ and smaller than 1. We set c_B by cross validation on the annotated examples.

Optimizing the MRF: We need to optimize the MRF to infer the locations $L_{f_i} = \ell$ where each face f_i is attending. There are a large number of possible locations (cells in the grid) and there can be up to 10 faces in a frame in some cases. Because the location at which a face is looking at is dependent on that of other faces, exact inference is intractable. We propose an approximate algorithm to solve this problem which is inspired by the α -expansion method. Our algorithm iteratively groups or separates faces based on whether they are looking at a common location or not.

Our algorithm starts by assigning each face’s attention to a location by only optimizing its unary terms. Thus, faces are first assigned to different groups. In the next stage, it considers both unary and pairwise terms and iteratively merges or splits the groups. At each step, it considers a pair of groups and measures if the total MRF energy increases as a result of merging them. If it does, the two groups are merged. Similarly, in each group, it measures whether removing a face increases the total energy. The procedure iterates until convergence. An illustration of this procedure is depicted in Fig 3. Qualitative results are shown in Fig 4.



Figure 4: Faces attending to a common location are shown with the same color. The bird’s eye view of the location and orientation of faces in 3D space is shown. The first person is shown by a circle at the bottom center of the images. Note that our method can estimate the common attention even if the faces are not looking at a person (c).

4. Method

In this section we describe our approach for detecting and recognizing types of social interactions in day-long first-person videos. We introduce three categories of features and provide an analysis of their capability to describe social interactions: (1) location of faces around the first-person, (2) patterns of attention and roles taken by individuals and (3) patterns of first-person head movement. We use these features in a framework that explores the temporal dependency over time to detect the types of social interactions.

4.1. Location of Faces around First-Person

Important evidence for the detection of social interactions is provided by the location of faces in the 3D space around the first-person. This is very similar in nature to the approach of [5], where they use the relative location of pedestrians to categorize group activities. For example, one can imagine that in a monologue, faces tend to appear in a circle around the person who is talking to the rest. In a dialogue a face tends to appear in front of the camera, looking at the first-person. To build location-based features, we divide the area in front of the first-person into 5 angular bins (from -75 to 75 degrees) and 4 distance bins (from 0 to

$5m$). Our method counts the number of faces in each bin, and returns a 20 dimensional histogram as a feature.

4.2. Attention and Roles

Social interactions are characterized by patterns of attention between individuals over time. When a person speaks, she attracts the attention of others. Once another individual takes the floor, the attention shifts to the new person.

Our idea is that during a social interaction, each individual present in the scene adopts a specific role. For example, in a monologue, there is a particular role that can be assigned to the person who is speaking, and another role played by the individuals listening to the speaker. Analyzing the change in roles over time can describe the patterns of turn taking and attention shift that are crucial elements of social interactions.

We assign roles to individuals based on four features that capture the patterns of attention for each individual x :

- Number of faces looking at x
- Whether first-person looks at x
- If there is mutual attention between x and first-person (both are looking at each other)
- Number of faces looking at where x is attending

We assign a 4 dimensional feature vector to each individual and then cluster all the examples in training sequences to a few bins using k-means. Each bin represents a role. We represent each frame by building the histogram of roles involved in a short interval around that frame.

4.3. First-Person Head Movement

A further cue for the categorization of social interactions is provided by the first-person head movement. The movement patterns complement the coarse attention estimation with transition information. In addition, in cases where two individuals are speaking while walking, and faces are absent from the video, the first-person head movement provides significant information.

We propose an additional feature to capture first-person head motion patterns. We extract features from dense optical flow [4] at each frame. We split each frame horizontally and vertically into a few sub-windows. We split the flow vector field in each sub-window into horizontal and vertical components, V_x and V_y , each of which is then half-wave rectified into four non-negative channels V_{x+} , V_{x-} , V_{y+} and V_{y-} . We represent each sub-windows with a vector containing the mean value of its motion channels. In our experiments, we split each frame into nine (3×3) sub-windows.

4.4. Temporal Model

The features described in previous sections encode a local snapshot in time. However, the temporal change in these features is crucial for detection and understanding of social interactions. The intuition behind our solution is that each frame is assigned to a state based on its features, and then an interaction type is assigned to the whole sequence based on the state labels and their dependencies. We model our problem with Hidden Conditional Random Field (HCRF) [18] for this purpose. In our model (Fig 5), frames are assigned hidden state labels and these states are connected by a chain over time. In HCRF, state labels are latent variables and are learned by the algorithm.

The HCRF model is learned over the following potential function Ψ :

$$\Psi(y, \mathbf{h}, \mathbf{x}; w) = \sum_{i=1}^n w_{h_i} \cdot \varphi_{x_i} + \sum_{i=1}^n w_{y, h_i} + \sum_{(k,l) \in E} w_{y, h_k, h_l}$$

where the graph E is a chain with nodes corresponding to hidden state variables \mathbf{h} , φ_{x_i} contains the feature vector from the small sub-window around frame i , and w contains the parameters of the model, which are learned during training using BFGS optimization. The label assigned to the whole sequence y , takes binary values in case of detection and takes multiple values (dialogue, discussion, monologue,

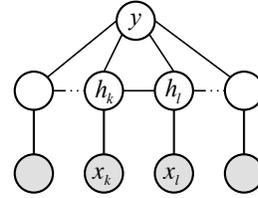


Figure 5: Our model. y is the social interaction label, h_l is the hidden state label assigned to frame l and x_l contains the features extracted from a local window around frame l .

walk dialogue, walk monologue) when trained for recognition. During testing, the label y for which the potential Ψ is maximum is assigned to the sequence.

5. Experiments

We present our social interaction detection and recognition results on a dataset collected at theme parks.

Dataset: To collect our dataset, we sent a group of more than 25 individuals to theme parks for three days. Each day a subset of the individuals used a head-mounted GoPro camera to record throughout the day. Our dataset contains more than 42 hours of video recorded by 8 subjects. The group usually broke into smaller groups during the day. As a result, each video contains a significant amount of experiences that are not present in the other videos. The cameras were fixed on caps. The GoPro cameras capture and store a high definition 1280×720 , 30 fps video. We extract images at 15 fps, resulting in over two million images in total.

We manually labeled the start and end time of intervals corresponding to types of social interactions throughout the videos. We have six labels: dialogue, discussion, monologue, walk dialogue, walk discussion and background. Each of these interactions can take place at a dinner table with group of friends, while walking, or while standing in a line, etc. We train our social interaction detectors on videos from five subjects and test on videos from the remaining three subjects.

Attention Estimation Results: Example results for face localization and attention estimation are shown in Fig 4. Our method both estimates who is looking at who, and in addition uses the context from the rest of the faces to estimate where in space an individual is attending. For example in Fig 4(c), the group of individuals with red circles around their faces are looking at the lady wearing a white shirt whose face was not detected. Our method realizes that these four individuals are looking at the same location and estimates this location in space. We quantitatively measure the performance of our method. We manually label who each person is looking at in a subset of the frames (about 1000 frames). For each frame, we connect each detected face to the one it is looking at. We split the ground-truth into two sets and use the first set to train the parameters of our model. In 71.4% of the cases our method correctly es-

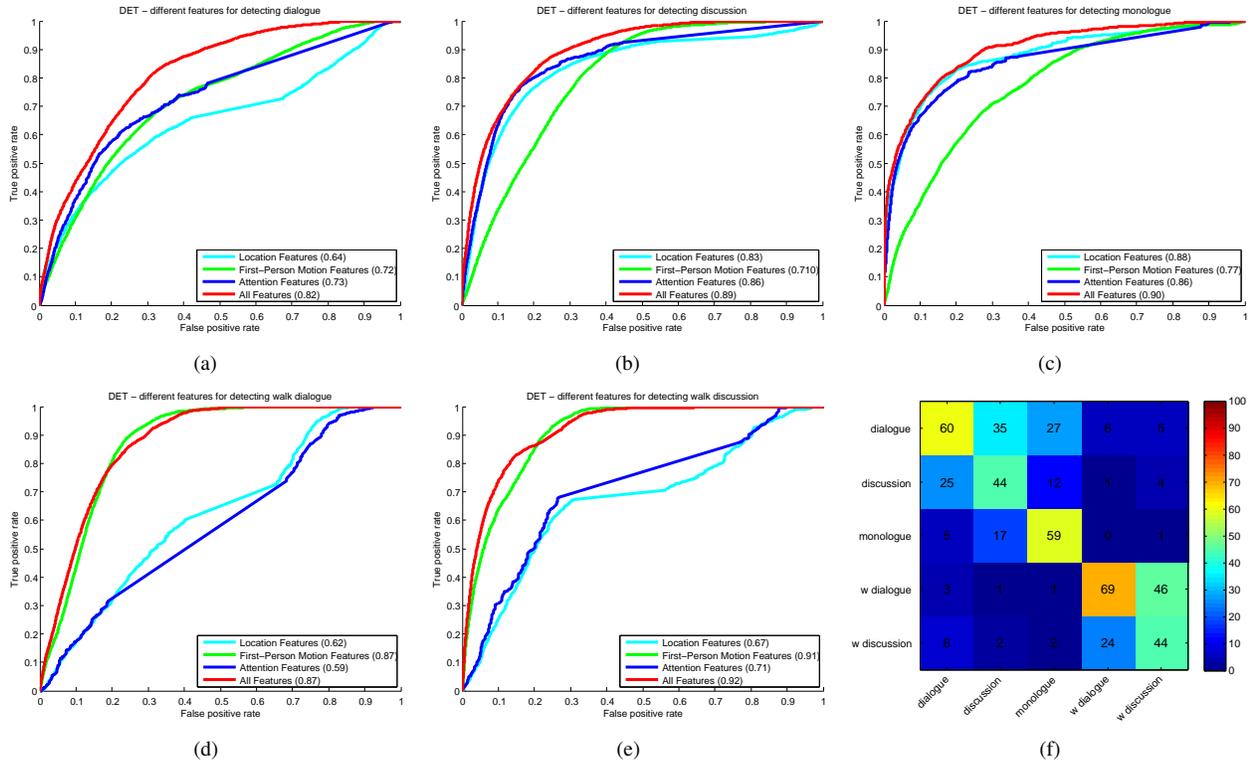


Figure 6: ROC curves of detecting types of social interactions are shown in (a-e). The area under each curve is provided in the figure. In case of dialogue and discussion, the attention features outperform flow and location features. In case of monologue, location features perform the best. First-person motion features significantly outperform the rest in detecting walk dialogue and walk discussion. In addition, we show our recognition results using all features in (f).

timates who is looking at who.

Detection and Recognition of Social Interactions:

During training, for each type of social interaction, we randomly select 100 intervals (of 200 frames each) from each subject’s video and 300 intervals from the background. As a result, the total number of intervals used for training are 4000. To learn a detector for a particular type of social interaction, we set the label of intervals corresponding to that type to positive and the rest to negative. During the test, we perform the detection on a 200 frame long interval around every frame of the test video. We set the number of hidden states of HCRF to 5 for the detection task. In Fig 6(a-e), we show the performance of our method on detecting different types of social interactions. For each type, we compare the performance of different features. Attention and location based features perform better at detecting dialogue, discussion and monologue, while first-person motion features perform better on walk dialogue and walk discussion. We show that the combination of these features together significantly improves the results for every type of social interaction.

We train a multi-label HCRF model for the recognition of social interactions. We set the number of hidden states to 10 for the recognition task. In Fig 6(d), we show the confusion matrix for recognizing social interactions. Walk

dialogue and walk discussion contain very similar motion patterns and there is a significant confusion between them.

Social Networks: Our focus in this paper is not analyzing or recovering social network of individuals, however, here we show the great potential for such task in first-person videos. We cluster the faces into multiple bins. We manually assign each bin to one of the individuals by looking at the faces it contains. We weigh the connection of a subject (person wearing the camera) to other people based on the number of faces in the cluster corresponding to that individual. The resulting network is illustrated in Fig 7.

6. Conclusion and Statistics

We describe a novel approach for detection and recognition of social interactions such as dialogue, discussion, and monologue, in day-long first-person videos. Our method constructs a description of the scene by transferring faces to 3D space and uses the context provided by all the faces to estimate where each person is attending. The patterns of attention are used to assign roles to individuals in the scene. The roles and locations of the individuals are analyzed over time to recognize the social interactions. We believe this is the first work to present a comprehensive framework for analyzing social interactions based on the patterns of attention

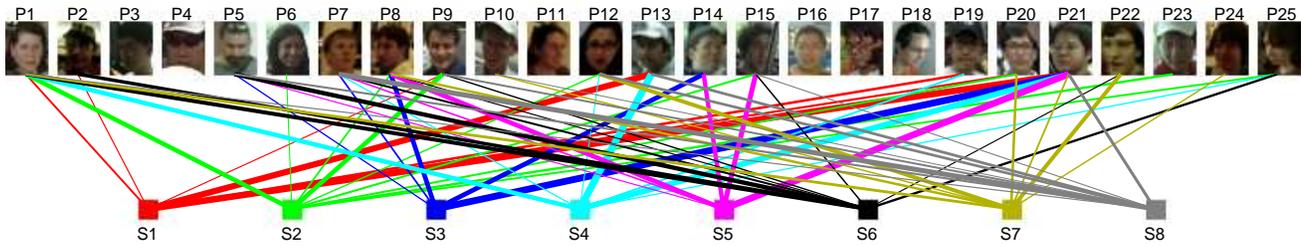


Figure 7: The social network built using our method. The representative faces of persons in the group $P_1 \dots P_{25}$ are shown. Subjects wearing the cameras $S_1 \dots S_8$ are shown by squares. We weigh the connections based on how frequently a person's face appears in the video captured by a subject. It is possible to notice some individuals like P_1 who was the tour guide are popular among the subjects. In addition, one can notice the similar connection patterns between S_1 and S_4 who were spending a significant time together throughout the day.

which are visible in first-person video. We present encouraging results on a challenging new dataset consisting of 42 hours of video captured at a popular amusement park.

7. Acknowledgment

Portions of this work were supported in part by ARO MURI award number W911NF-11-1-0046, National Science Foundation award IIS-1029679, and a gift from the Intel Corporation.

References

- [1] O. Aghazadeh, J. Sullivan, and S. Carlsson. Novelty detection from an ego-centric perspective. In *CVPR*, 2011. 2
- [2] A. Aris, J. Gemmell, and R. Lueder. Exploiting location and time for photo search and storytelling in mylifebits. In *Technical Report, MSR-TR-2004-102*, 2004. 2
- [3] B. Benfold and I. Reid. Guiding visual surveillance by tracking human attention. In *BMVC*, 2009. 3
- [4] T. Brox, A. Bruhn, N. Papenbergh, and J. Weickert. High accuracy optical flow estimation based on a theory for warping. In *ECCV*, 2004. 6
- [5] W. Choi, K. Shahid, and S. Savarese. Learning context for collective activity recognition. In *CVPR*, 2011. 3, 5
- [6] T. Choudhury. Sensing and modeling human networks. In *Doctoral Thesis, MIT*, 2004. 2
- [7] L. Ding and A. Yilmaz. Learning relations among movie characters: a social network perspective. In *ECCV*, 2010. 2
- [8] A. Fathi, A. Farhadi, and J. M. Rehg. Understanding egocentric activities. In *ICCV*, 2011. 2
- [9] A. Fathi, X. Ren, and J. M. Rehg. Learning to recognize objects in egocentric activities. In *CVPR*, 2011. 2
- [10] J. Gemmell, L. Williams, K. Wood, R. Lueder, and G. Bell. Passive capture and ensuing issues for a personal lifetime store. In *ACM Workshop on Continuous Archival and Retrieval of Personal Experiences*, 2004. 1, 2
- [11] K. M. Kitani, T. Okabe, Y. Sato, and A. Sugimoto. Fast unsupervised ego-action learning for first-person sports videos. In *CVPR*, 2011. 2
- [12] T. Lan, Y. Wang, W. Yang, and G. Mori. Beyond actions: discriminative models for contextual group activities. In *NIPS*, 2010. 2
- [13] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *CVPR*, 2008. 2
- [14] M. J. Marin-Jimenez, A. Zisserman, and V. Ferrari. "here's looking at you, kid." detecting people looking at each other in videos. In *BMVC*, 2011. 3
- [15] V. I. Morariu and L. S. Davis. Multi-agent event recognition in structured scenarios. In *CVPR*, 2011. 3
- [16] B. Ni, S. Yan, and A. Kassim. Recognizing human group activities with localized causalities. In *CVPR*, 2009. 3
- [17] A. Patron-Perez, M. Marszalek, A. Zisserman, and I. D. Reid. High five: recognizing human interactions in tv shows. In *BMVC*, 2010. 3
- [18] A. Quattoni, S. Wang, L-P. Morency, M. Collins, and T. Darrell. Hidden-state conditional random fields. In *PAMI*, 2007. 6
- [19] B. Schiele, N. Oliver, T. Jebara, and A. Pentland. An interactive computer vision system - dypers: dynamic personal enhanced reality system. In *ICVS*, 1999. 2
- [20] T. Yu, S-N Lim, K. Patwardhan, and N. Krahnstoeber. Monitoring, recognizing and discovering social networks. In *CVPR*, 2009. 2