Discovering Social Interactions in Real Work Environments

Chih-Wei Chen, Rodrigo Cilla Ugarte, Chen Wu, Hamid Aghajan

***Abstract*— The goal of this work is to detect pairwise primitive interactions in groups for social interaction analysis in real environments. We propose a system that extracts locations and head poses of people from videos captured in an unconstrained environment, namely a research lab. Our system is designed to work with realistic data capturing natural human interactions. An efficient tracking method based on Chamfer matching finds the head and shoulder silhouettes of people in real-time, and a head orientation classifier estimates their head poses. The location, relative distance and head orientation of people capture the use of space by individuals and their interactive behavioral patterns which are inferred with a probabilistic model. We present quantitative evaluation and experimental results of our system, demonstrating the effectiveness of our proposed approach on challenging real-world data.**

I. INTRODUCTION

This paper presents a complete system to study the social interactions among people in a realistic work space. The final goal of the system is to obtain the knowledge of the behavioral patterns of people. We want to answer queries like "How do people utilize the office?" or "How often do they interact?" in a real-world environment. Work space can be optimized by understanding space utilization, and productivity can be increased by exploiting the synergies among people.

To this end, we propose to use a video camera to capture the daily interactions among people in an office environment. In this paper, we focus on pairwise primitive interaction detection in groups of people, i.e., whether two people have direct interaction defined as being close in location and looking at each other. The motivation to detect primitive interaction is that it can be used as the basic element to classify more complex group interactions. For example, by analyzing the pattern of primitive interactions among pairs in a group of people for a period of time, the group interaction type (e.g., a presentation or a discussion) can be inferred. Features extracted from the videos include the locations and head orientations of multiple people in the videos. The intermediate location and head pose information, together with the identity of the people and the spatial layout of the environment, allow high-level reasoning of social behavior to be performed. Our design philosophy is to build a system that runs in real-time, since the rapid accumulation of the day-to-day videos prohibit expensive operations that require off-line processing. We also aim to develop an extendable system that can operate based on a single-view or multi-view system.

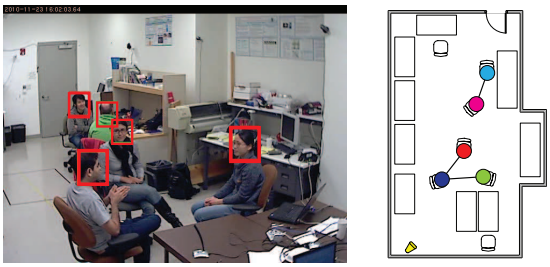

Fig. 1. Discovering Social Interactions: (Left) Head tracking results. The locations of people are tracked in a challenging environment, and their head poses are estimated. (Right) The locations of the people mapped onto the schematic floor plan of the office. A discriminative model is trained to classify whether primitive interactions occur between a pair of people given their locations and relative head orientations. We connect two people, represented by the colored vertices in the figure, if they are directly interacting. Further grouping can be performed by finding connected components in the graph. In this example, there are two groups consisting of two and three people, respectively.

The study of proxemics [6] has shown that the social distance is reliably correlated with physical distance. Body spacing and posture indicate the spaces in which people feel comfortable conducting social interactions. In this paper, we focus on primitive interactions in a work environment, i.e. direct interactions that are considered personal or social. It has also been shown that the orientation of a person's head conveys rich information about the focus of attention [7]. In the social contexts, the direction of a person's head indicates the intended target of interactions. The head pose also reflects the social dynamic. For example, in group meetings people turn their heads to the person who is speaking. Therefore, we propose to use the head poses in the discovery of social patterns. We estimate the head poses by learning a head orientation classifier.

Detecting and tracking multiple people in real-world environments is a challenging problem. We use a tracking-by-detection approach, a paradigm that has been dominant in recent years [1], [2], [12]. To achieve real-time processing, we adopt an efficient Chamfer distance matching [3]. More specifically, following [14] we construct an Ω-shape head-shoulder template, and match it against the edges in the video frames. Because of cluttered backgrounds and heavy occlusion in the video sequences, the detector might fail to localize all heads within the frame. In such cases we turn to low-level tracking [9], which uses the temporal coherence to find the lost person, until next valid detection is available.

We also acquire the identity information by incorporating a radio-frequency identification (RFID) reader in our system.

This work was developed in Stanford AIR (Ambient Intelligence Research) Lab.

C.-W. Chen, C. Wu, and H. Aghajan are with the Department of Electrical Engineering, Stanford University, Stanford, CA 94305, USA {louistw, chenwu, aghajan}@stanford.edu

R.C. Ugarte is with Departamento de Informática, Universidad Carlos III de Madrid, Madrid, Spain rcilla@inf.uc3m.es

Each person in the office is asked to carry a unique RFID tag, which assists the system to associate detection with each person in the videos across different days. The identity information is useful for the study of behavioral patterns of individuals. More precisely, as the appearance of an individual may change from day to day, the RFID system helps to associate the day-to-day correspondences of detected persons.

Given a calibrated camera and the head location in the video frame, the real-world coordinate of the person can be estimated. To do this, we assume people in the scene are either sitting or standing, and ignore the transient transitions since they only occupy short periods of time in our long observation. The state of the target is determined from motion cues and geometric constraints. The estimated location and head pose allow us to focus on the analysis of high-level behavioral patterns. We plot the trajectories of individuals in the video to visualize their patterns. For human interactions, we build a probabilistic model that predicts whether or not two people are interacting based on their relative locations and head orientations.

Our contributions in this work are:

1) We have built a system that is capable of tracking the location and head pose of multiple people in a challenging, realistic, and unconstrained environment in real-time.
2) We address the problem of discovering the space utilization and human interactions. Despite the noise from low-level modules, our proposed approach is still capable of inferring interaction patterns.
3) We introduce a data set with realistic human activities and social interactions. While many previous data sets are captured in controlled environments where actors are instructed to perform in restrictive manners, our videos capture the realistic day-to-day activities and natural social interactions among a group of people working in a research lab. The tracks, the head pose estimations and the individual identities serve as the building foundation for inferring social behavior models to be constructed on.
4) Our method reported in this paper relies on a single-view system; however, the developed algorithms are extendable to multi-view cases where additional geometric information can be fused in to enhance the detection as well as assisting the high-level reasoning.

II. System Overview

Our system consists of three main modules: tracking, head orientation estimation, and behavioral pattern inference. An overview diagram is shown in Figure 2. The tracker keeps tracks of the locations of people. We address the problem of discovering the social dynamics among a group of people. The social role of each individual is ever-changing in different social contexts and settings. To monitor the social dynamics over a period of time, the identity of each individual must be recognized. To this end, we employ a radio-frequency identification (RFID) system to automate the task. Each member in the office is given a unique tag. An RFID reader is set up at the entrance of the office to detect the entrance of each individual member. An auxiliary camera captures the appearance of the person when an RFID tag is detected. The appearance, together with the identification information, not only assists the tracking across different days, but also provides a growing data set for a face recognition algorithm to be trained on.

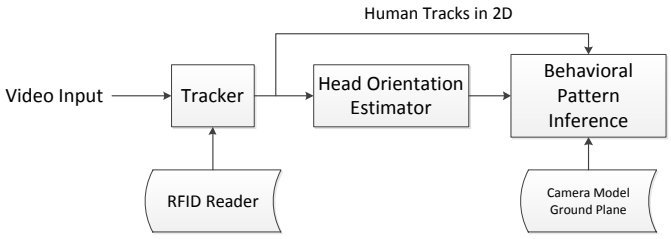

Fig. 2. System overview. The three main modules of our system are tracker, head orientation estimator, and interaction classifier.

III. Tracking Multiple People in Real-Time

We first describe how multiple people are tracked in real-time from a single camera. The tracker we use combines head detection and low-level tracking. A diagram of the tracker is illustrated in Figure 3. The tracking-by-detection approach we employ is similar to that of [9], which incorporates a face detector and is designed to work with archive films. However, in our case, we are dealing with videos where the targets do not necessarily face the camera. Therefore, a different detector must be used. We use an edge-based head-and-shoulder detector.

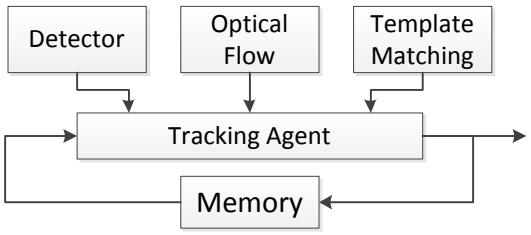

Fig. 3. The block diagram of the tracker. The tracking agent uses a tracking-by-detection approach, where gaps between detections are filled by low-level template matching. The locations of the tracked targets are kept in memory, so stationary or occluded people can be tracked and associated with detections when they become active or visible again. Optical flow is used to estimate the change in posture of the target, i.e. transition between standing and sitting.

A new track is initialized if a new detection meets two criteria. First, the detection is not associated with an existing track. Second, the detection is close to the top of a foreground blob. In our experiments, the second condition significantly suppresses false detections in regions with edges resembling the shape of a person's head, such as the back of an office chair.

A detection d is associated with an existing track r_t at time t if the overlap is greater than a predefined threshold. Formally, we define the overlap between a detection d and a track r as $O(d, r) = (d \cap r)/(d \cup r)$. Each track has its own appearance template, denoted by A. Let I be the frame image and I_d denote the image patch at the detection d.

Once a detection is associated with the track, the appearance template of that track is updated, $A = I_d$.

Due to the extensive human pose articulation, the head and shoulder detector might fail to detect some people. Therefore, we incorporate a low-level tracking function to keep track of the target. While several low-level tracking approaches have shown their capabilities to achieve reliable tracks, we employ low-level tracking only to fill in the gaps before the next candidate is found. As a result, we propose to use a simple and efficient method. We use template matching to guide our tracker. In particular, normalized correlation is used to search for the best match in the incoming frame.

The detector in our system is efficient and runs in real-time. The block diagram of the detector is shown in Figure 4. First of all, a foreground mask is generated by creating an adaptive background mixture model [10]. The foreground mask allows us to focus on the foreground objects and reduce the search space. A head detector then matches an Ω-shape template against the foreground edge map of the incoming frame. Such a template captures the silhouette of a person's head and shoulder, and the matching is invariant to colors and robust to shape variations by using Chamfer distance [3] as the difference metric. Multiple scales of the template are matched against the edge map. Locations with matching distance smaller than a threshold are considered candidate locations, and are passed to the tracker. In our experiments, we use 4 different scales of the template.

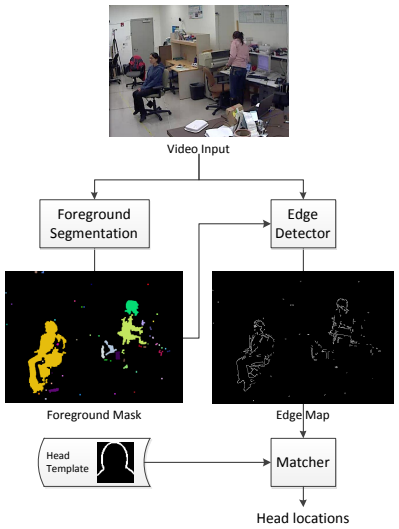

Fig. 4. The block diagram of the head detector. By creating a background model, the detector focuses on foreground objects. The foreground edge map is matched against a head template using Chamfer distance. Chamfer matching provides robust detection so reliable tracking can be achieved.

IV. Head Pose Estimation

Head pose [7] is a useful cue for the detection of interactions between people, as it provides information related to their focus of attention. A head pose is described by three numerical values: *yaw*, *pitch* and *roll*, describing a 3D angle. The main difficulty associated with this representation is to obtain accurate ground truth data to train head pose estimation algorithms. Some experiments, e.g. [7], have used electronic head pose tracking devices attached to the people being imaged to obtain an accurate training corpus. In our system a weaker approach is proposed based on the notion that accurate estimations of head pose are not needed to recognize social interactions. Hence the space of possible head pose configurations is discretized to $K = 8$ different values. Ground truth data is obtained by manually labeling training samples. Figure 5 shows example images from each head pose category.

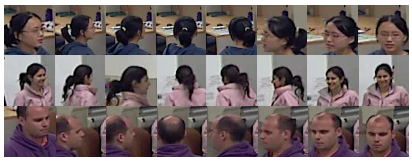

Fig. 5. Examples of the different head pose categories. The training examples are obtained from the tracker, therefore contain large variance in scale and alignment.

The bounding boxes provided by the head tracker proposed in the previous section have a large variance in the head alignment with respect to the bounding box location. Head pose estimation methods based on facial feature segmentation are robust to this issue [7]. However, the input images in our system are acquired from far field views of the scene, containing faces too noisy or small to try any facial feature segmentation methods. That is why a holistic approach to head pose estimation is used. First, input bounding boxes are transformed to gray-scale, equalized and resized to an $X \times Y$ normalized-size box. Image rows are concatenated to obtain a head descriptor of $d = X \times Y$ dimensions. This descriptor is projected into a low-dimensional space using a function learned with the method proposed below. The result is introduced into a classifier to obtain the head pose category.

Given a set of N input points (head pose descriptors) $X = \{x_i\}_{i=1}^{N}$, $x_i \in R^n$, Isometric Projection (IsoP) [4] aims to find a function f that maps these N points to a set of points $Y = \{y_i\}_{i=1}^{N}$ in R^d ($d << n$), such that $y_i = f(x_i)$. The method is of special applicability when the points $X \in M$, where M is a nonlinear manifold embedded in R^n.

Let d_M be the geodesic distance measure on M and d the standard Euclidean distance measure in R_d. IsoP aims to find an embedding function f such that Euclidean distances in R_d can provide a good approximation to the geodesic distances on M. Thus, the function to obtain is the one that minimizes:

$$\sum_{ij} \left(d_M(x_i, x_j) - d(f(x_i), f(x_j))\right)^2 \qquad (1)$$

As the underlying manifold M where real dataset is defined is unknown, the geometric distance measure d_M is also unknown. To discover the intrinsic geometrical structure of M, a neighborhood graph G containing all the points in X is constructed. There are two standard ways to construct this graph:

1) ϵ-graph: A link is established between x_i and x_j if $d(x_i, x_j) < \epsilon$
2) kNN-graph: A link is established between each point x_i and its k nearest neighbors.

Once the neighborhood graph G has been obtained, the geodesic distance on the manifold d_M between two points x_i, x_j is approximated as the distance on the graph between the points, i.e. $d_G(x_i, x_j)$. Thus, the matrix $D_G = d_G(x_i, x_j)$, which contains the distances between all the points on X needs to be computed. The standard procedure to obtain D_G is to use Floyd-Warshall algorithm, with complexity $O(N^3)$. As the graph G is usually sparse, Johnson's algorithm can reduce the problem to $O(NV \log V)$, where V is the number of edges in G.

If the projection function f is restricted to be a linear function of the form $f(x) = A^T x$, the columns of the matrix $A = [a_1 \dots a_d]$ are given by the solutions of the generalized eigenvalue problem:

$$X[\tau(D_G)] X^T a = \lambda X X^T a \tag{2}$$

where the matrix $\tau(D_G) = -HSH/2$ is an inner product matrix, being $H = I - \frac{1}{N} ee^T$, I the identity matrix and e a vector of all ones; and S a matrix such that $S_{ij} = D_{ij}^2$

Readers are referred to the original publication [4] for additional derivation and implementation details.

V. Analysis of Social Behavior

Understanding social behavior has gained more attention in recent years , partly driven by the growing volume of available data. Researchers have also shown that understanding social interactions can be a basis for improving the performance of many existing systems, such as video tracking (by modeling the dynamics of the crowd [8]) and face recognition (by learning the interpersonal relationship of the people in the pictures [11]). In [13], 3D trajectories and individual identities are obtained from multiple cameras with face recognition. In this paper, we focus on interaction discovery based on realistic data.

An interesting question to answer is how people occupy a workspace. In particular, we are interested in how people spend time in different areas in the office according to the functionality. Given the bounding boxes in the image frames, the 3D coordinates of the person can be estimated if the height of the person is known and the camera is calibrated with respect to the room's geometry. In our experiments, we assume that people are either standing or sitting. We use a uniform height assumption for everyone, setting $h = 1.70m$ if the person is standing or $h = 1.25m$ if the person is sitting on a chair. Three cues are useful when determining the human position. The first one is the geometry constraints. The person must be on the ground plane inside the office. Also, the scale of the detection gives us an estimation of the size of the full body. The second cue comes from motion. Strong upward or downward optical flow fields imply changes in the position. Our tracker keeps track of the strongest direction of the motion field, and uses it to resolve ambiguities when both standing and sitting locations are geometrically feasible. The last cue is temporal consistency. If both previous methods fail, the location is set to the one closer to where the person was reported in the previous step.

The relation between the 2D point in an image and the 3D point in the world is governed by the following equation:

$$y \sim Cx \tag{3}$$

where C is the 3 x 4 camera projection matrix, $y \in \Re^3$ and $x \in \Re^4$ are homogeneous representations of the 2D image and 3D world coordinates, respectively, and the notion $\sim$ denotes equality up to scaling by a non-zero factor. The camera projection matrix C is obtained by performing camera calibration. If we assume the height in the 3D coordinates is known, the other two unknowns can be found by solving Eq. 3.

We train a discriminative classifier for pairwise primitive interaction detection on pairs of people based on their distances and relative head poses. The primitive interactions in our focus are direct communications , i.e. two people talking to each other, or two people making eye contacts. In our system, the features we use capture the nature of this type of interactions we focus on.

We now describe how the features are computed. Given the 2D locations of two people y_1, y_2, their 3D locations x_1, x_2 are first estimated. The distance $d = \|x_1 - x_2\|$can be readily calculated. Assume the head orientation is quantized to N directions. Their head orientation estimations o_1, o_2 can be respectively represented by an N-dimensional vector with the sum 1, indicating the probability of the head's orientation being in that direction. The head orientation is estimated with respect to the image plane. This must be taken into account when calculating the relative orientation of the heads. Assume the two people form an angle θ with respect to the camera. Then the relative head orientation of the two is $o_1 - o_2 - \hat{\theta}$, where $\hat{\theta}$ is θ quantized to N directions.

Head orientation estimation is very sensitive to alignment errors. To increase the robustness of the system, we run frontal and profile face detectors in local image patches, and use the output as a coarse 4-way head orientation classification (frontal, profile left, profile right, back). We combine the features in a discriminative model using a support vector machine (SVM). The resulting $N + 5$-dimensional feature vectors are passes to an SVM with a linear kernel, and a classifier is trained using LIBSVM [5].

VI. Experiments

We present results from the full implementation of the proposed system. We set up a fixed video camera in our office, which is a research lab. An additional auxiliary camera is pointed at the entrance of the office to capture the appearance of the person passing the entrance. An RFID reader is set up at the entrance. We attach an RFID tag to the door, so when the door is opened the system is triggered. The RFID tag on the door triggers a surveillance mode that notifies the system to start recording. The system stops recording after a short period of time unless another RFID tag containing a lab member's identity is detected. The surveillance mode handles the case when personnel we are not monitoring, such as a visitor or the janitor, enters the office. The RFID reader logs in a person when his RFID

tag is detected (upon entering), and logs him off when his RFID tag is detected again (upon leaving). The system stops recording when everyone is logged out, or can be forced to stop by presenting a special stop RFID tag to the reader. The schematic floor plan of the setup is shown in Figure 6.

We have deployed the system in our office and performed data collection for more than two weeks[1]. Unlike previous data sets, our videos capture realistic human activities and natural social interactions. Considering the long duration of the collected videos, and that our algorithms have been in the development and testing phase, we found it impractical to evaluate the system on all videos. We therefore test our system on a few short sequences sampled from our recordings to report in this paper while further refinements in the different parts of the methods are being investigated.

We use an Axis IP camera to capture the videos. The resolution of the video is 640 x 480 pixels and the frame rate is 30 fps. While other modules in our tracker do run at 30 fps, the detector only processes 1 out of every 5 frames, hence operating at 6 fps. The rate is more than enough for social analysis to be performed, since social interactions usually last for seconds or minutes. Moreover, we will shortly show that the resulting system generates reliable tracks with this sampling rate.

Tracking: We evaluate the performance of our tracking system on a test video containing multiple subjects. We extract from our daily recording a short sequence for testing. The test sequence contains 5 people chatting and meeting occasionally. The length of the video is about 11 minutes and 30 seconds, with a total of 4150 frames (sampled at 6 fps) and 5 tracks.

Since manual labeling of the ground truth locations of the targets is prohibitive, we report the precision of the tracker. We manually go through the bounding boxes returned by the tracker, and count the number of true positives. We show the results in Table I. Denote the ground truth head location by G. A bounding box B is considered correct if $\frac{(G \cap B)}{B} > 0.5$ and $\frac{(G \cap B)}{G} > 0.5$. As shown in the table, the precision of the tracker is very high.

Another important performance indicator of the tracker is the recall rate. Our tracker can detect people soon after they enter the field of view of the camera, and starts tracking. Due to the nature of the tracker, a target might be lost if the head and shoulder detector cannot find a good candidate and the appearance changes dramatically. This happens when people change their postures, for example by sudden turning or bending. In our experiments, we found that this only happens occasionally, mostly relating to transient moments. Our tracker is able to recover lost or occluded targets once they become visible and detectable.

Space utilization: We use the same video sequence to analyze the social behavioral pattern. We first visualize the trajectories of the 5 individuals in the video. In Figure 6, the location heat map is shown. Due to the limited field of view of the camera, not all people in the office are always observed. As shown in the figure, people spend more time near their own seats (cyan circle and red diamond markers in the figure). The trajectory of the person walking into the office and to his seat (shown in magenta cross marker) is tracked. The heat map also shows that the center of the office is the common place where multiple people pass and stay on.

Track	Number of Frames	True Positives	Precision
1	2836	2681	94.5%
2	2295	2118	92.3%
3	3619	3458	95.6%
4	2208	2017	91.4%
5	1988	1872	94.2%

TABLE I
TRACKING PERFORMANCE.

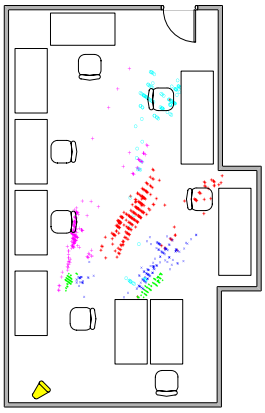

Fig. 6. The heat map of people in the office. We plot the locations of people on the schematic floor plan of the office.

Primitive human interaction: We use the same test video sequence for evaluating our primitive human interaction classifier. In the 11 and a half minutes long test sequence, we have interaction instances performed by different people in diverse ways. The first interaction instance is a conversation between two people sitting in their office chair, followed by a group meeting among the other three. After both meetings were dismissed, two people started working together using the same desk. Finally a group meeting took place, and all five people got involved in the meeting gradually, sitting or standing in different postures. Example frames are shown in Figure 7. We label interaction primitive instances in our data set, and use them to train our interaction classifier. None of the training examples appear in the test sequence. We sample approximately 200 frames from the test sequence and run our interaction classifier. We report the average precision of our classifier, and visualize the results. Example results are shown in Figure 7.

The average precision of the interaction classifier is 0.71. We compare our interaction classifier to a simple baseline. We train a baseline classifier using only the distance but not head orientation. The average precision of the baseline classifier is only 0.56. By incorporating the head orientation, the performance of the interaction classifier is noticeably improved. This is because proximity doesn't always imply two people are interacting, especially in an office envi-

[1]Please see example videos and results at `http://airlab.stanford.edu/videos/social/examplevideo.avi`

ronment where people might sit next to each other. Our classifier is also capable of dealing with occlusion, because our tracker integrates temporal information and stores the location and head orientation in the memory. As a result, we can still correctly infer an interaction's continuation even when the location or head orientation of the persons cannot be accurately estimated in the current frame but is detected in a subsequent frame.

VII. Discussions and Future Work

We have presented a system that tracks human locations and estimates the head poses in a real world environment from videos acquired from a single stationary camera in an unconstrained environment. The real-time system allows high-level social behavior analysis to be performed. Specifically, we study the occupancy of the space by each individual. This provides valuable information for space utilization and management. We also present a probabilistic model for analyzing their interactions using the locations and head poses. The detected primitive interactions can be used not only for the analysis of social interaction, but also for the construction of more complex group interaction models. For example, by looking at the duration and proportion of primitive interactions in a meeting, we can study each participant's engagement in the group interaction. Similarly, the role of each participant can be inferred by incorporating the temporal pattern of primitive interactions among the social group.

In our approach, we keep every component simple so the overall system can run in real-time and be robust. A natural future extension is to employ a more complex tracking method. In addition, both localization and head orientation estimation can be improved by using multiple cameras. With multiple views, ambiguities can be resolved, especially when occlusions occur. By fusing multiple visual cues from different cameras, or even auditory cues using a multimodal approach, the robustness and accuracy of the system can be further improved.

We have provided a computer vision solution to obtain useful information from videos. It is desirable to run the system for an extended period of time, so a richer data set can be constructed. More thorough analysis can be performed and interesting results can be shown if data over a long time is available and processed. For example, we will be able to observe how a person's interaction pattern changes over time.

VIII. Acknowledgments

We gratefully acknowledge the reviewers for their careful and constructive comments. We would like to thank Amir Hossein Khalili and Parinaz Sayyah for useful discussions and helping with the collection of the data.

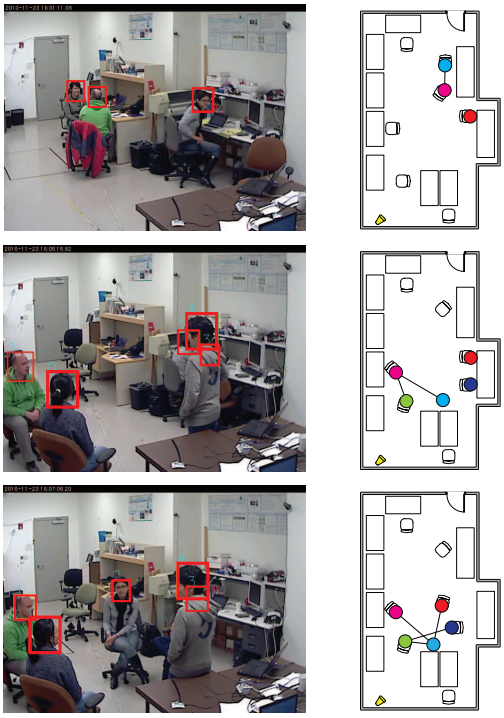

Fig. 7. **Examples of discovered interactions:** We show tracked head bounding boxes in the frames on the left, and the corresponding estimated 3D locations of the people in the office on the right. **(Top)** Three people in the office. The two people talking near the desk is detected by our interaction classifier, and hence connected by a line in the visualization. **(Middle)** Five people in the office as two groups. Our classifier failed to detect the interaction between the two people sitting next to each other near the desk on the right because of heavy occlusion. However our tracker is still capable of correctly keeping the locations of the occluded heads in its memory. **(Bottom)** Five people in the office as a big group. By using temporal reasoning, our system remembers the locations and head orientations of the sitting person, occluded by the standing person, and can correctly detect interaction even if he is not visible in the scene.

References

[1] Mykhaylo Andriluka, Stefan Roth, and Bernt Schiele. People-tracking-by-detection and people-detection-by-tracking. In *IEEE CVPR*, 2008.
[2] Shai Avidan. Ensemble tracking. In *IEEE CVPR*, 2005.
[3] Gunilla Brogefors. Hierarchical chamfer matching: A parametric edge matching algorithm. *IEEE PAMI*, 10:849–865, 1988.
[4] Deng Cai, Xiaofei He, Jiawei Han, Deng Cai, Xiaofei He, and Jiawei Han. Isometric projection. In *AI*, pages 2006–2747, 2007.
[5] Chih-Chung Chang and Chih-Jen Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.
[6] Edward T. Hall. A System for the Notation of Proxemic Behavior. *American Anthropologist*, 65(5):1003–1026, 1963.
[7] Erik Murphy-Chutorian and Mohan Manubhai Trivedi. Head pose estimation in computer vision: A survey. *IEEE PAMI*, 31:607–626, 2009.
[8] Stefano Pellegrini, Andreas Ess, Konrad Schindler, and Luc van Gool. You'll never walk alone: Modeling social behavior for multi-target tracking. In *ICCV*, 2009.
[9] Xiaofeng Ren. Finding people in archive films through tracking. In *IEEE CVPR*, 2008.
[10] Chris Stauffer and W. Eric L. Grimson. Adaptive background mixture models for real-time tracking. In *IEEE CVPR*, 1999.
[11] Gang Wang, Andrew C. Gallagher, Jiebo Luo, and David A. Forsyth. Seeing people in social context: Recognizing people and social relationships. In *ECCV*, pages 169–182, 2010.
[12] Bo Wu and Ramakant Nevatia. Detection and tracking of multiple, partially occluded humans by bayesian combination of edgelet based part detectors. *IJCV*, 75(2):247–266, 2007.
[13] Ting Yu, Ser-Nam Lim, Kedar Patwardhan, and Nils Krahnstoever. Monitoring, recognizing and discovering social networks. In *IEEE CVPR*, 2009.
[14] Tao Zhao, Ram Nevatia, and Bo Wu. Segmentation and tracking of multiple humans in crowded environments. *IEEE PAMI*, 30:1198–1211, 2008.