SOCIAL INTERACTION INFERENCE AND GROUP EMERGENT LEADERSHIP DETECTION USING HEAD POSE

A Thesis

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Master

by

Shenghao Liu

December 2020

ABSTRACT

Understanding the interaction between people from images is yet a challenging task in recent years. Several approaches attempted to address this challenge by different proposed models and compact descriptors encoding the consistency between people's spatial-temporal body features or their overall activities. Among all human body features, the head pose provides a distinct description of an individual's attention and is considered a key feature to interpret social interactions. In this thesis, I present a novel approach to infer interaction among a group of people in a group conversation scenario using the Markov Random Field model. A novel interaction feature is proposed to represent the transactional segment using the head pose. Furthermore, I extend the approach to infer the hierarchical structure of a group with the contextual information, which improves the leading method in the analysis of interactions among people and detect the emergent leader of the group. The qualitative result shows the effectiveness of interaction inference using a state-of-art dataset.

BIOGRAPHICAL SKETCH

Shenghao Liu is a Master of Science graduate student in the Laboratory for Intelligent Systems and Controls supervised by Professor Silvia Ferrari at Cornell from 2018 to 2020. He was an undergraduate researcher in the Mechanical Systems Control Lab supervised by Professor Masayoshi Tomizuka at University of California, Berkeley. His research interest focused on robotics and computer vision includes human keypoint detection and human interaction modeling.

This document is dedicated to all Cornell graduate students.

ACKNOWLEDGEMENTS

I would like to express my sincere appreciation to Prof. Silvia Ferrari and Prof.Bharath Hariharan for their precious guidance and advice throughout my research at Cornell University. I want to give a special thanks to Graduate Affair Director Marcia J. Sawyer for her kindness and patience, which effectively relieved my anxiety. I wouldn't make this far without the help of Cornell's talented staff and faculty. I also want to thank Junyi Dong in our lab for her suggestions and help on the research reports and the thesis. Finally, I would like to thank all my friends, Xinyu, Wenbo, Yifan, Yifeng, Gulai, Dongheng and Zhihao, for their generous help in the academy over the two years at Cornell.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Understanding the interaction between people from images is yet a challenging task in recent years. The application of inferring interaction between individuals is diverse from video surveillance systems to human-robot interaction, even in understanding human behavior and personality in cognitive science. The problem addressed in this thesis is how to identify the interaction between people from their head pose as an interaction cue. This chapter discusses the challenge of estimating head pose and interaction inference and provides an overview of the thesis contributions.

1.1 Motivations

In many daily-live environments, people usually form groups to collaborate for a certain goal to improve society's efficiency. Interaction between people is an important social contextual cue that contains rich cognitive information of each participant. Modeling and tracking the interactions is useful for many applications: such as video surveillance [21], group activity recognition [16, 17, 22], target tracking [15, 19, 52], and participant personality classification [54, 65]. Recently, sociologic reasoning has been incorporated into video surveillance algorithms aiming to interpret social interaction rather than directly infer the interaction result by deep learning. In cognitive science, social interaction can be characterized by verbal cues, non-verbal cues, or both [18]. Concretely, recent research shows interest in non-verbal cues such as vocal cues and visual cues. Thanks to the development of convolutional neural network algorithms,

non-verbal cues such as human gesture [48], human body orientation [12, 59], especially head pose [49], even gaze [62], can be directly estimated from images and be further used to infer interaction.

In this thesis, I will be focused on investigating the interaction between people based on the head pose, concretely, the location and orientation of their head with respect to a certain global coordinate system. The head pose contains rich, interpersonal information in different forms. For example, a person indicates the object of interest by pointing his head toward the target and maintaining fixation for seconds. Similarly, during a conversation, the head pose of participants implies the current speaker and turn-taking of leadership which deduces the emergent leader of the group [7]. Meanwhile, some head gestures convey additional semantic information during a conversation. The nodding gesture implies the consent to the current speaker's opinion, and a repetitive horizontal head movement indicates disagreement. The interpersonal interaction can be observed by establishing the visual focus of attention (*VFOA*) from head pose estimation. Research [64] proposed that in a meeting scenario, the head pose alone tracks the VFOA in 89% of the time. If two people exchange visual attention on each other within a certain distance in an empty space, they are likely to engage in a discussion. A person's head pose also provides his perception of the environment. When entering a new environment, people usually scan the environment and memorize some space characteristics by shifting their head towards some random direction or some specific objects of interest.

Conversational groups can be formalized in the form of *F-formation*, that is, as defined by Kendon [37], a spatial and directional arrangement of people gathered for conversation where each person has direct and unhindered contact with

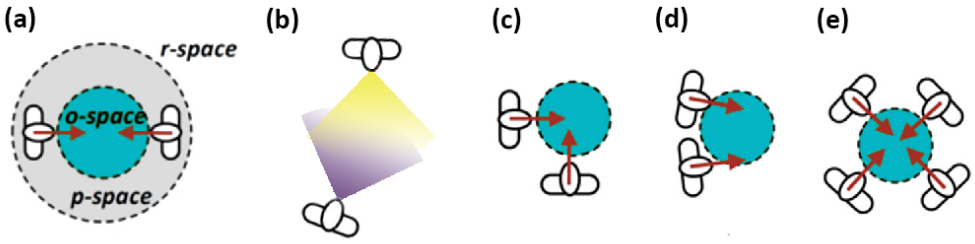


Figure 1.1: Configuration Space of F-formation (a), transactional segment of two people (b), and three cases of F-formation (c-e)

other people, as shown in Fig 1.1a. The configuration space of F-formation is split into three mutually exclusive sub-spaces, namely, o-space, p-space, and r-space. The o-space is an empty and convex space where the interaction exhibits and is surrounded by the narrow p-space that includes all the participants' positions. The remaining space that lies outward p-space is called r-space. The composition of configuration space is displayed in Fig 1.1c-e. Within the o-space, a person's *transactional segment* denotes the space in front of the person that overlaps with other transactional segments when interaction occurs, which is the overlapping area in Fig 1.1b. For the case of interaction between two people, their F-formation can be categorized into three types of F-formation: L-shape F-formation, side-by-side F-formation, and circular F-formation. Various non-verbal cues that utilize the theory of F-formation are proposed in [79, 57, 34, 70, 58, 20, 38] to facilitate the sociologic reasoning of social interaction. Non-verbal cues are categorized as follows: the low-level features which describe the spatial configurations of each person in a scene such as the person's position and head/body orientation, and the high-level features that integrate the low-level features to implement sociological and biological definitions such as 3D subjective view frustum (SVF) [24] or transactional segment-based frustum (TSbF)[70]. The interaction feature proposed in this thesis inherits the idea of TSbF which will be introduced in chapter 5.

1.2 Challenges

Previous research had attempted to cluster people by directly learning their interactions in image by machine learning approaches. For example, [71] obtained convincing result on the classification of group activities based on the consensus of individual activity learnt by a fine-tuned CNN. However, analogous to other deep-learning-based algorithms, the model suffer from lack of interpretability. Meanwhile, feature-based models rely on the reliability of graphical models and the representability of interaction features, and thus often only valid under certain conditions. In this thesis, I attempted to resolve the limitation of feature-based models by a novel interaction feature based on head pose.

Meanwhile, in some cases, some social interaction cues are intractable because of noisy observation or unobserved communication between people. As proposed by [71], social interaction cues take various forms: vocal behavior, forward posture, mutual gaze, gesture, and height. However, not all cues are tractable in most cases. For example, vocal behavior in video surveillance are difficult to collect and associate to each person unless the experiment is conducted thoughtfully as [3] where the participants' speaking status are collected by a microphones attached in front of them. Therefore, inferring interaction with limited information is still an active area of research. In this thesis, only image-based information can be obtained and is assumed to be tractable and credible.

1.3 Contributions

To address the above challenges, a novel baseline approach, i.e. *interaction graph*, is proposed that inherits the benefit of graphical models uses Markov Random Fields (MRFs) to represent social interaction by undirected probabilistic arcs. A representative spatial-temporal feature describing pairwise interaction is then proposed and contributes to an energy function which models the strength of interaction associated with the undirected arcs. Based on the construction of MRF model and the energy function, an optimal interaction configuration is obtained by Maximum a Posteriori (MAP) algorithm. An experiment is performed a benchmark dataset with conversational groups; the qualitative result consistently shows the reasoning of inference result and the robustness to noisy environment.

Meanwhile, the shortcoming and limitation of the baseline approach are analyzed based on sociological reasoning. A revised approach, named *attention graph*, relaxes the constraints enforced on the baseline approach and uses directed probabilistic arcs and a discriminative feature that better describes each individual's transactional segment. The energy function is modified accordingly. The qualitative result performed on the same dataset consistently shows the superiority of the revised approach over the baseline approach.

1.4 Thesis Overview

The thesis is organized in the following order: in chapter 2, an overview of state-of-art achievement on social interaction inference is presented. In chap-

ter 3, the assumption and formulation of the head pose estimation problem are introduced as a classical computer vision problem. The interaction graph and attention graph are formulated using the graph theory of Markov Random field, the observations based on the head pose estimation are collected under certain assumptions. Chapter 4 defines the interaction feature and interaction energy function as proposed, followed by the inference algorithm that searching for the optimal interaction configuration. The selected state-of-art head pose estimation algorithms are compared in chapter 5, and the corresponding numerical results demonstrate the accuracy and robustness of the algorithms. Finally, the qualitative inference result of the interaction graph and attention graph are compared under the same scenarios, demonstrates the characteristics of the two graphical models.

CHAPTER 2

BACKGROUND

Previous works has attempted to estimate head pose from single images or consecutive image sequences. Several literature proposed different approaches to deduce the interaction from the detected head pose. This chapter will review the state-of-art literature of head pose estimation followed by interaction modeling.

2.1 Head Pose Estimation

The human pose estimation from images has been a challenging task, among which the head pose estimation is one of the most important area of research. The high variance of human facial appearance (expression, race, and gender) and environmental factors (occlusion and illumination) makes the task more challenging. Early methods had attempted to solve the task in various approaches as organized in [49].

Recently, the *nonlinear regression method* captured the researcher's attention as the computation power explodes with the release of GPU. Various deep learning models model a nonlinear functional mapping from the image space to the pose directions. Methods proposed in previous works estimate head pose either from the original image [72, 46, 2, 10, 60, 51], the depth image [8], the RGB-D image[77], or the optical flow of the image sequences [78]. Borghi [8] concatenated a recurrent layer at the output of convolutional layers in a regression manner to capture the continuous 3D head pose angle values (roll, pitch and yaw). Ruiz [60] attempted to jointly estimate the roll, pitch and yaw by

training three separate layers after the output of ResNet [31], and obtained an accurate estimation.

In this thesis, the head pose is reconstructed by the facial keypoints detection. Various network structures [11, 63, 66] attempted to extracted human body key points by fine-tuned neural networks. Among those detection algorithms, OpenPose [11] captures the position of eyes, nose, and ears in the images in real-time. In this thesis, the position of facial keypoints in 2D images are computed from OpenPose, and the head pose estimation problem is transformed into a *perspective-n-points problem*.

2.2 Perspective-n-points Problem

The perspective-n-points (PnP) problem was first brought up by Fischler and Bolles [28], it is to determine the position and orientation of a calibrated camera with known camera intrinsic (e.g focal length, distortion factor) from n known 3D reference points with respect to a fixed frame and their corresponding 2D image projections. In this thesis, the relative locations of facial points in a proposed head frame are fixed and are known as a priori. To solve the six unknowns (the 3D position and orientation of the camera), at least six equations (three correspondences) are required to obtain a closed-form solution. However, Fischler and Bolles [28] observed that a maximum of four solutions is possible given the three correspondences. They proposed a unique closed-form solution when four points are given and not co-planar. Given the biological fact that the facial keypoints are not co-planer, a minimum of four points is required to obtain a closed-form solution to head pose estimation problem. Meanwhile, a set of con-

straints is commonly adopted to formulate PnP solutions to find a closed-form solution following reality conditions.

There are many existing solutions to solve the PnP problem considering robustness and time efficiency. The family of PnP solutions splits into two categories: iterative solutions and non-iterative solutions.

Iterative solutions find the solution by iteratively minimizing certain objective functions. Fischler and Bollesv [28] proposed the RANSAC algorithm which employs outlier rejection schemes that eliminate the effect of less-accurate points. In the Procrustes PnP (PPnP) [29], the solution is obtained by iteratively adjusting the depth and the estimated camera parameters until convergence by minimizing the error of reference 3D points and the back-projection points computed from 2D points. However, iterative methods usually find local optima that may actually differ from the ground truth.

The early non-iterative PnP solutions proposed by Ansar, Quan, and Fiore are computationally ineffective, all of which take at least quadratic computational complexity, $O(n^2)$, given n correspondences. The first non-iterative solution with linear computational complexity is EPnP [44], which reduces the complexity by expressing the position of 3D points as a weighted sum of the position of four virtual control points, and uses their coordinates to construct quadratic equations in four cases to select the weights, and keeps the solution that yields the smallest reprojection error.

Other non-iterative state-of-the-art solutions decouple PnP problems as a set of low-dimensional polynomials and find the solution with minimal objectives: Robust PnP (RPnP)[45] decomposes the PnP problem into a set of P3P prob-

lems, and retrieved the solution by solving the fourth-order polynomial system using Singular Value Decomposition (SVD). Direct-Least-Squares (DLS)[32] relaxes the solution by Cayley-Gibbs-Rodriguez (CGR) parametrization and find the optimal solution among 27 sub-optimal solutions by minimizing the Mean Squared Error (MSE) of reference images points and the projection of 3D points. Recent literature Accurate and Scalable PnP (ASPnP)[81] and the Optimal PnP (OPnP)[80] uses quaternion instead of rotation matrix to represent rotations of the solution and uses Grobner basis technique to solve the polynomial system to find the global optimal solution by minimizing projected error expressed in forms of the quaternion. Robust Efficient Procrustes PnP (REPPnP)[26] inherits the idea of virtual control points proposed by EPnP [44], and incorporates an outlier rejection scheme to remove the contribution of outliers to the virtual control points. Recently, the Covariant Efficient Procrustes PnP (CEPPnP)[27] and Maximum-likelihood PnP (MLPnP) [69] are proposed that inherently incorporate observation uncertainty into the framework.

The various PnP algorithms are compared on simulated data and real data in chapter 5, and the optimal algorithm is selected to estimate head pose, which will be converted to interaction features.

2.3 Interaction Modeling

The interaction modeling task is generally analogous to group detection and clustering, where usually the position and orientation of people in the image or the real world are known as a priori. Most recent literature obtained fair and accurate results on people detection and tracking [9, 6, 75, 39, 36]. Qin

[56] proposed a probabilistic model that simultaneously performs multi-target tracking, head pose/direction estimation, and social grouping in surveillance video. The model jointly optimizes social grouping and multi-target tracking as a constrained nonlinear optimization framework given the conditional independence assumption on each other, and the head pose is estimated by maximizing a posteriori likelihood of graph labeling modeled by conditional random field. The group interaction inference can be partitioned in three categories:

1. The dynamic information based inference is where individuals are clustered by individuals' trajectories, including position, speed, the direction of motion, destination, or tracklets.
2. The static information based inference, where individuals' velocity is irrelevant to grouping within a time interval of interest, and the head/body orientation and audio information dominate individuals' interaction. The visual attention can also be characterized by a frustum used to model the transactional segment of individuals in a F-formation model.
3. Combined information based inference, where both kinetic information, head/body orientation are considered simultaneously, and contextual information such as activity classification of individuals.

Dynamic information based inference In a crowded environment where most people are moving individually, pedestrians' head and body orientation are less important as they can be represented by the direction of traveling most of the time, as suggested by [5]. In this case, the term interaction is considered the consensus of motion, and the pairwise interaction between individuals within the group is ignored. Existing methods attempted to track social grouping using either proximity [53, 43], position, speed. and direction of motion

[13, 30, 74, 42, 23], and the linkage of tracklets [55].

Static information based inference In a narrow space where people tend to conduct a face-to-face conversation, the head/body orientation and audio information are the deterministic cues that contribute to establishing an interaction. Odobez et al. [50], Hung et al. [33], and Ba et al. [3] investigated the VFOA in group meetings with multimodal features including head pose, position, the participant speaking activity, and the slide activity. They proved that head pose could be used to recognize the VFOA of meeting participants with sufficient confidence [50], while audio and other contextual information improved the result of VFOA recognition performance by nearly 20% [33, 3].

Combined information based inference Leash [41] further extend his previous work [42] with additional visual attention. A visual interest feature is introduced, which promotes the pairs of individuals closer in proximity and exchange visual sights but punishes both of the pairs who are looking in the direction of travel. Besides, other contextual information is incorporated as cues to infer interactions. A group context activity descriptor proposed by Tran [68] incorporates each individual's activity and the relationship with ac- activities of his neighbors to represent the similarity of neighboring activities, and a dominant set based clustering algorithm is used to discover interacting groups based on the proposed descriptor. Choi [14] discovered the structure of groups by minimizing an energy function composed of potentials that encode the compatibility between pairwise interaction patterns and the interaction labels learned from training data, the individual appearance of being a singleton, the consensus of interaction labels and individual patterns within the same group, and the repulsion of interaction labels of being different groups. The framework pro-

posed by Lan [40] jointly captures the connections between individual action labels, the similarity of action among the group, and the group activity labels. The group structure is learned from mixed-integer linear programming (MILP) to find the optimal graph structure based on the weighted average of four potential functions. The potential functions models the likelihood of the individual actions and its image feature, the compatibility between the group activity and individuals' action, the action compatibility between a pair of individuals, and the likelihood of the group activity and root feature vector whole image, respectively. Yoo [76] further explored the psychological cues to infer group interaction by considering the spatial information of individuals and the emotional correlation between individuals predicted by Social Relation Network.

In this thesis, the model I presented lies in the second category, which is implemented in a static conversation group scenario using head pose and body position.

2.4 Attention Modeling

Previous works attempted to simplify the redundant configuration features by sociological theories and derived a high-level feature to model attention's visual focus. The simplest feature models the affinity between a pair of individuals as a multivariate normal distribution with their relative distance and orientation as variables [79]. Tran [67] represents the pairwise interaction by summarizing two distance social force functions introduced by sociologist Was et al. [73]. The distance social force can be modeled using a linear, step, power, or polygonal function. Farenzena [24] proposed 3D Subjective View Frustum, namely SVF,

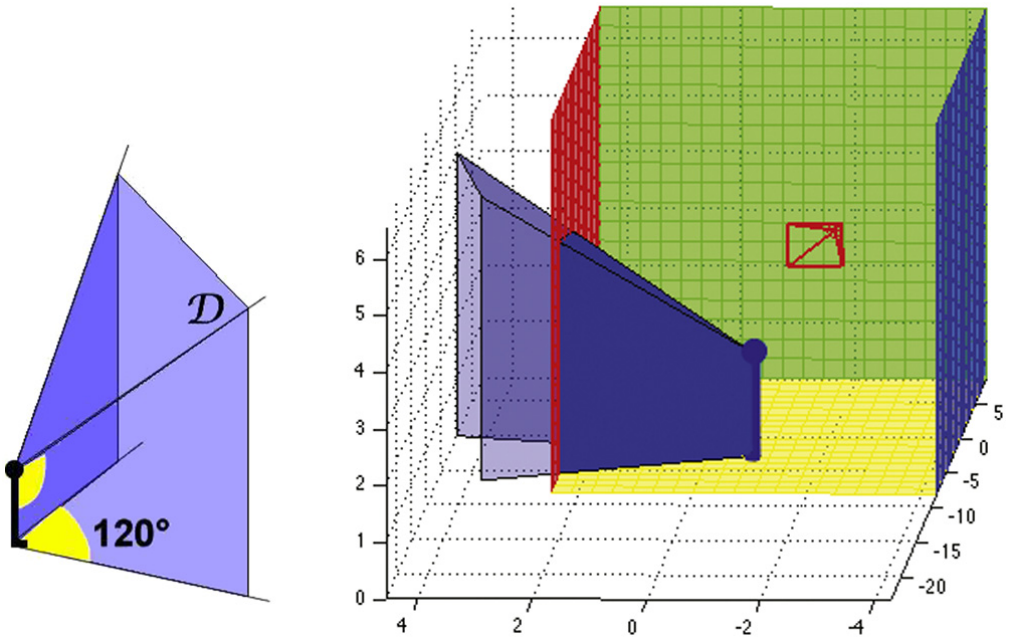


Figure 2.1: (Left) The SVF model. (Right) An example of SVF delimited by the scene constraints (in solid blue), the SVF orientation is estimated with respect to the camera frame.

to represent the focus of attention. The Subjective View Frustum is viewed as a space enclosed by three planes indicating the boundary of the view angles on the left, right, and top sides illustrated by the pyramid in 2.1.

The intersection of the planes corresponds to the 3D position of the head and feet, while the head pose determines the orientation of SVF. A later work of Bazzani [4] suggests that SVF can be employed to discover the dynamic interaction of multiple people by the overlapped SVF over a specific time interval. A similar definition of the interaction area called Transactional Segment-based Frustum (TSbF) is inspired by the definition of the transactional segment [1]. The TSbF model proposed by Vascon [70] not only constraints the area where interaction occurs but also represents the angular and longitudinal likelihood of interaction as shown in Fig 2.2. For each person, a set of particles are sampled from the TSbF model based on the head pose and position, and the particles are

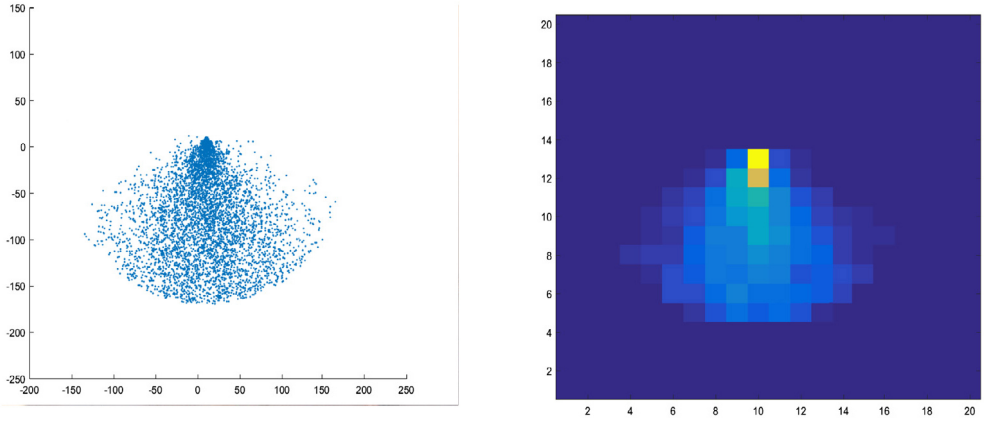


Figure 2.2: (Left) The TSbF based on sampling from two distributions proposed in [70]. (Right) The TSbF descriptor, the intensity reflects the number of particles that fall in each bin of the 2D histogram; the denser of the particles, the higher the probability of interaction.

organized into a 2D histogram. The likelihood between the histograms of two people quantifies the strength of interaction.

2.5 Emergent Leadership

Given the clustering of people, an *emergent leader* is determined as the person with the maximum interaction with others that has a predominant effect on the group. Several social signal processing studies investigated the detection of an emergent leader in terms of the nonverbal features were extracted from audio [25], video [7] and the fusion of audio and video [33, 61]. In this thesis, the emergent leader is determined as the person who received the most attention from others, referred to as the Attention Received (AR).

CHAPTER 3

PROBLEM FORMULATION

This thesis attempted to develop an algorithm to infer interactions between detected people from image sequences based on the 3D reconstruction of their head pose defined as the 3D position and orientation relative to the camera. The estimated head pose is further used to infer people's gaze in a designated scenario and determine the emergent leader of each group. This thesis assumes that a human's gaze is only affected by the head pose rather than eye movement variation. In practice, the eye movement is difficult to capture as the head pose conveys enough information on the focus of interest. Meanwhile, this thesis investigates a conversational-group scenario where the signs of communication are predominant and exclusive.

3.1 Head Pose Estimation

In this thesis, the head pose estimation problem is modeled as a *Perspective-n-point* (PnP) problem, originally used to estimate the position and orientation of a pin-hole camera from known correspondences of 3D reference points and their projected positions in the image. Since the camera's position and orientation are assumed to be fixed over the experiment, the position and orientation of the 3D reference points can be inferred by the homogeneous transformation.

In a Perspective-n-point problem, a Cartesian camera frame C is attached to an ideal pin-hole camera as shown in Fig. 3.1(Left), and a body frame B is attached to the nose tip of the person of interest as shown in Fig. 3.1(Right). The

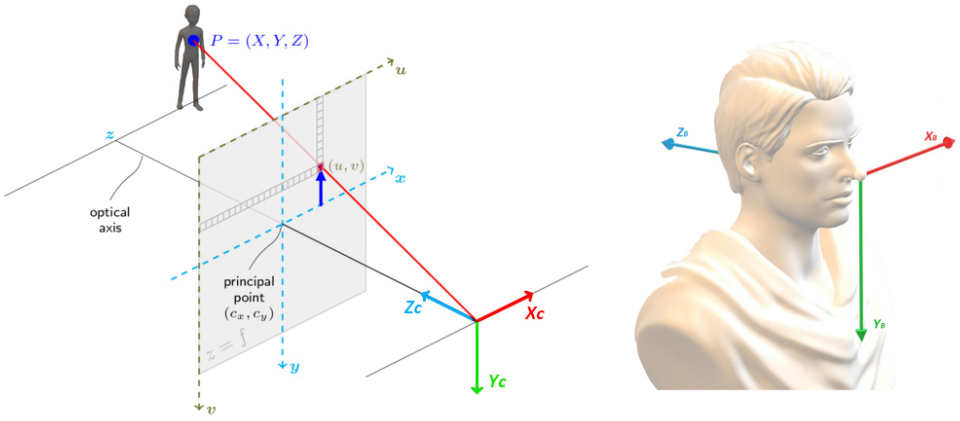


Figure 3.1: (Left) Definition of pin-hole camera and camera frame C. (Right) Definition of body frame B of a person

camera's focal length and optical center are known as a priori, and the pin-hole camera model does not account for pixel skew or lens distortion. The camera parameters are given by the matrix product of the intrinsic and extrinsic matrix. The intrinsic matrix K is characterized by the optical center (c_x, c_y) and focal length f of the camera given by equation 3.1.

$$K = \begin{bmatrix} f & 0 & c_x \\ 0 & f & c_y \\ 0 & 0 & 1 \end{bmatrix} \tag{3.1}$$

The extrinsic matrix is the concatenation of the orientation R_B^C and position t_B^C of the body frame B with respect to the camera frame C, and is expected to be recovered from the PnP algorithms.

In a PnP problem, it is assumed that we know n 3D reference points $\mathbf{q}_i = \begin{bmatrix} X_i & Y_i & Z_i \end{bmatrix}^T \in \mathbb{R}^{3\times1}$, $i = 1, 2, ..., n$, in the body frame B, and their corresponding projections $\mathbf{p}_i = \begin{bmatrix} u_i & v_i \end{bmatrix}^T \in \mathbb{R}^{2\times1}$, such that the perspective projection equation

$$s_i \mathbf{p}_i = K(R_B^C \mathbf{q}_i + t_B^C) \quad i = 1, 2, .., n \tag{3.2}$$

holds for all points, in which s_i is a scaling factor to ensure that the projected image point lies on the image plane.

The 3D reference points and the projected points are organized into the input matrix, $\mathbf{Q}$ and $\mathbf{P}$ respectively, where $\mathbf{Q} \in \mathbb{R}^{3\times n}$ and $\mathbf{P} \in \mathbb{R}^{3\times n}$ is given by equation 3.3

$$\begin{aligned} \mathbf{Q} &= \begin{bmatrix} \mathbf{q}_1 & \mathbf{q}_2 & \dots & \mathbf{q}_n \end{bmatrix} \\ \mathbf{P} &= \begin{bmatrix} \mathbf{p}_1 & \mathbf{p}_2 & \dots & \mathbf{p}_n \end{bmatrix} \end{aligned} \tag{3.3}$$

The 3D reference points are predefined facial keypoints (i.e., left eye, right eye, nose, left ear, and right ear) relative to the body frame B whose relative location is known as prior knowledge from biological literature. The variation of facial keypoints among individuals is ignored in this thesis. The location of facial keypoints in images are detected by OpenPose [11] 2D pose estimation algorithm. The person with less than four detected facial keypoints is ignored because a minimum of four points is required to obtain a closed-form solution to the estimation of head pose. Finally, the 3D reference points and the corresponding image points are fed into the Perspective-n-Point algorithms to compute each person's head pose.

3.2 Graph-based Social Interaction Modeling

3.2.1 Interaction Graph Modeling

The Interaction Graph is modeled as a Markov random field represented by an undirected graph $\mathcal{G} = (\mathcal{N}, \mathcal{A})$ where $\mathcal{N}$ is the set of nodes and $\mathcal{A}$ is a set

of undirected arcs. A set of random variables $\mathbf{X} = \{X_{i,j} | i, j \in \mathcal{N}\}$ is associated with the arcs where $X_{i,j}$ is associated with arc $(i, j) \in \mathcal{A}$ between node i and j. Each node represents a person with index of $i \in \{1, 2, ..., N\}$ separately where N is the size of $\mathcal{N}$. The value of random variable $X_{i,j}$, denoted as $x_{i,j}$, is binary, which indicates the occurrence of interaction between person i and person j such that $x_{i,j}$ equals one when the interaction exists between person i and j, and equals zero when the interaction does not exist, i.e., $x_{i,j} \in \mathcal{L}$, where $\mathcal{L} = \{0, 1\}$. The interaction configuration across all participants in a scene is denoted as $\mathbf{x} = \{x_{i,j} | i, j \in \mathcal{N}\}$, with domain $\mathcal{X} = \mathcal{L}^{N^2}$.

For each person i, the head pose is transformed into a 3D directional vector $\mathbf{v}_i \in \mathbb{R}^{3\times1}$, and 3D position $\mathbf{p}_i \in \mathbb{R}^{3\times1}$ relative to the camera frame. The 3D directional vector $\mathbf{v}_i$ is perpendicular to the facial plane and points forward from the tip of the nose, in the opposite direction of the z-axis in Fig 3.1 (Right). $\mathbf{p}_i$ is the position of the nose tip relative to the camera frame, which is equal to $t^C_{B_i}$. $\mathbf{v}_i$ is computed by Eq.3.4 where $R^C_{B_i}$ and $t^C_{B_i}$ is the calibrated camera rotation and position with respective to the body frame B_i of person i.

$$\mathbf{v}_i = R^C_{B_i} \begin{bmatrix} 0 & 0 & -1 \end{bmatrix}^T \tag{3.4}$$

The two measurements are organized into an observation vector $\mathbf{z}_i \in \mathbb{R}^{6\times1}$ for person i,

$$\mathbf{z}_i = \begin{bmatrix} \mathbf{v}_i^T & \mathbf{p}_i^T \end{bmatrix}^T \tag{3.5}$$

Then the observation vector of all people is concatenated to form an observation matrix $\mathbf{Z} \in \mathbb{R}^{6\times N}$,

$$\mathbf{Z} = \begin{bmatrix} \mathbf{z}_1 & ... & \mathbf{z}_N \end{bmatrix} \tag{3.6}$$

Given the observation of all people in the frame, an energy function is defined to model the interaction strength attached to each edge, and a constrainted optimization algorithm is applied to infer the realization of $\mathbf{x}$ with the highest confidence.

3.2.2 Attention Graph Modeling

The undirected interaction graph described in the previous section is unable to describe people's focus in a conversational group, yet it is not a good representation for monologue scenarios. To address these two problems, the attention graph is proposed, where a node still represents a participant, but the outgoing arcs are directed and point towards the focus of attention of the corresponding node. As a result, the attention graph shows all participants' attention, and the node that receives the most attention is perceived as the emergent leader.

The attention model is represented by a directed graph $\mathcal{H} = (\mathcal{N}, \mathcal{E})$. The node set is the same as in the interaction graph but the arc set $\mathcal{E}$ comprises of directed arcs. A set of random variables describes attention is denoted as $\mathbf{Y} = \{Y_{i,j} | i, j \in \mathcal{N}\}$ where $y_{i,j}$ denotes the binary random variable associated with the directed arc starting from node i and ending at node j. The value of the binary random variable $Y_{i,j}$ indicates the attention from person i to person j such that $y_{i,j}$ equals to one when person i exhibits attention to person j, and equals zero when the attention does not exist, i.e., $y_{i,j} \in \mathcal{L}$, where $\mathcal{L} = \{0, 1\}$.

In this model, the observation matrix for person i is measured in its body frame attached to the nose tip, as is shown in Fig.3.1 (Right). Let $\mathbf{p}_{ij} \in \mathbb{R}^3 = \begin{bmatrix} \mathrm{x}_{ij} & \mathrm{y}_{ij} & \mathrm{z}_{ij} \end{bmatrix}$ denote the relative position of person j with respect to person i, $\mathbf{p}_{ij}$

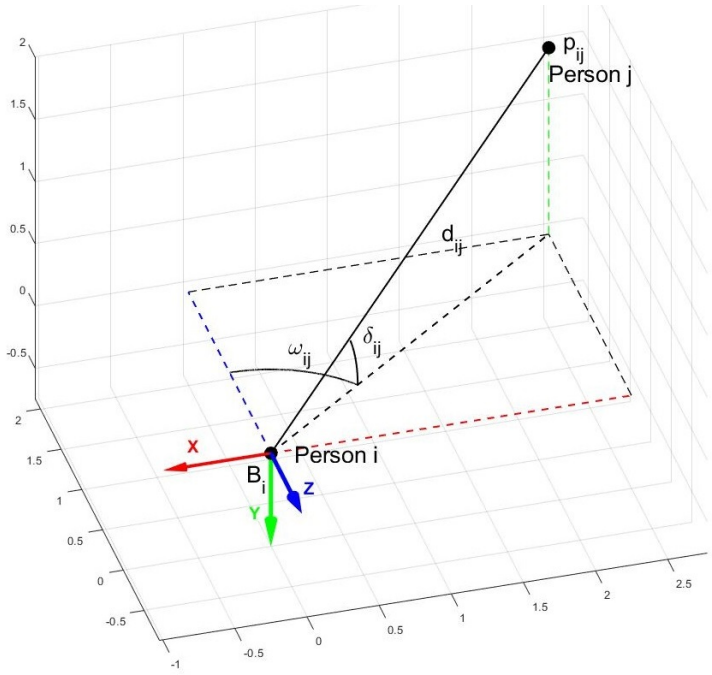


Figure 3.2: Relative position and angle of person j with respect to person i

is transformed to a pan angle $\omega_{ij} \in [-\pi, \pi]$, a tilt angle $\xi_{ij} \in [-\pi/2, \pi/2]$, and the distance $d_{ij} \in \mathbb{R}^+$ as shown in Fig.3.2.

The pan angles, tile angles and distances are computed by the equation below,

$$\begin{aligned} \omega_{ij} &= arctan\left[\frac{x_{ij}}{z_{ij}}\right] \\ \xi_{ij} &= arctan\left[\frac{y_{ij}}{\sqrt{x_{ij}^2 + z_{ij}^2}}\right] \\ d_{ij} &= \sqrt{x_{ij}^2 + y_{ij}^2 + z_{ij}^2} \end{aligned} \tag{3.7}$$

The pan angles, tile angles and distances of all people with respect to person i are organized into an observation vector $\hat{\mathbf{z}}_i \in \mathbb{R}^{3N}$,

$$\hat{\mathbf{z}}_i = \begin{bmatrix} \omega_{i1} & \xi_{i1} & d_{i1} & \dots & \omega_{iN} & \xi_{iN} & d_{iN} \end{bmatrix}^T \tag{3.8}$$

Note that the hat mark is to distinguish from the observation $\mathbf{z}_i$ of the interaction graph. Then the observations of all people are concatenated to form an

observation matrix $\hat{\mathbf{Z}} \in \mathbb{R}^{3N \times N}$,

$$\hat{\mathbf{Z}} = \begin{bmatrix} \hat{\mathbf{z}}_1 & \dots & \hat{\mathbf{z}}_N \end{bmatrix} \tag{3.9}$$

Similarly, an energy function that models the attention strength is computed using the proposed attention feature, and the connectivity of the graph is optimized accordingly. The Attention Received (AR) factor for each person is calculated by the number of connections to each node to determine the emergent leadership of the group.

CHAPTER 4

METHODOLOGY

4.1 Interaction Graph Modeling and Inference

4.1.1 Interaction Energy Function

An energy function is defined to model the interaction strength based on two interaction features. The interaction feature used in this thesis is a interaction potential characterized by the relative head pose θ_{ij} and the relative distance d_{ij} for any node pair (i, j). The relative head pose θ_{ij} computed by Eq.4.1 measures the angle (in radians) between the directional vectors of person i and person j,

$$\theta_{ij} = cos^{-1}\left(\frac{\mathbf{v}_i \cdot \mathbf{v}_j}{\|\mathbf{v}_i\|\|\mathbf{v}_j\|}\right) \tag{4.1}$$

The relative distance d_{ij} is computed in Eq.4.2 as the Euclidean distance between the two people,

$$d_{ij} = \|\mathbf{p}_i - \mathbf{p}_j\| \tag{4.2}$$

Let $\phi_h(\mathbf{z}_i, \mathbf{z}_j)$ denote the head pose potential, which is a measure of interaction strength with respect to θ_{ij}. The head pose potential is computed by a Gaussian-shaped potential function with parameter π, which indicates that people having interaction are more likely to oriented face-to-face ($\theta_{ij} = \pi$) for the ease of communication.

The value of $\phi_h(\mathbf{z}_i, \mathbf{z}_j)$ computed by,

$$\phi_h(\mathbf{z}_i, \mathbf{z}_j) = \frac{1}{\sigma_h\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\theta_{ij}-\pi}{\sigma_h}\right)^2} \tag{4.3}$$

where the variance σ_h^2 is a hyperparameter preset to $\pi^2/36$.

The proximity potential $\phi_d(\mathbf{z}_i, \mathbf{z}_j)$ measures the interaction strength in terms of the normalized distance d_{ij}/L. L is the threshold, beyond which the distance is too large that interactions will not be considered. The proximity potential is computed by the following equation,

$$\phi_d(\mathbf{z}_i, \mathbf{z}_j) = \begin{cases} \frac{1}{\mathrm{B}(\alpha,\beta)}\left(\frac{d_{ij}}{L}\right)^{\alpha-1}\left(1-\frac{d_{ij}}{L}\right)^{\beta-1} & d_{ij} \in (0, L] \\ 0 & d_{ij} \in (L, +\infty) \end{cases} \tag{4.4}$$

where the hyperparameters α and β determine the shape of the potential function, and $\mathrm{B}(\alpha, \beta)$, is a scaling constant determined by α and β.

The interaction potential $\phi(\mathbf{z}_i, \mathbf{z}_j)$ is determined by the product of two interaction potentials given by Eq.4.5.

$$\phi(\mathbf{z}_i, \mathbf{z}_j) = \phi_h(\mathbf{z}_i, \mathbf{z}_j) * \phi_d(\mathbf{z}_i, \mathbf{z}_j) \tag{4.5}$$

The energy function is constructed as the linear combination of the realizations of the random variables $\mathbf{X}$ and the interaction potentials for all people in the scene,

$$E(\mathbf{Z}, \mathbf{x}) \triangleq \sum_{\forall (i,j) \in \mathcal{A}} \phi(\mathbf{z}_i, \mathbf{z}_j) x_{i,j} \tag{4.6}$$

4.1.2 Interaction Inference

The Markov Random Field model is referred as a Gibbs random field under the condition that the joint probability of the random variables is strictly positive. According to the Hammersley–Clifford theorem, the joint posterior probability $P(\mathbf{x}|\mathbf{Z})$ of a MRF model can be factorized over the cliques of the graph as the

product of locally defined clique potentials $\phi_c(\cdot)$ if the MRF model is a Gibbs random field. The clique potentials are then parameterized as a log-linear function so that $P(\mathbf{x}|\mathbf{Z})$ can be written as the weighted sum of exponential family in canonical form with feature function $\mathbf{f_c}$. The derivation is described in Eq.4.7.

$$P(\mathbf{x}|\mathbf{Z}) = \frac{1}{C(\mathbf{x})} \prod_{\mathbf{x} \in \mathcal{X}} \phi_c(\mathbf{x}, \mathbf{Z}) = \frac{1}{C(\mathbf{x})} exp\Big(\sum_{\mathbf{x} \in \mathcal{X}} \mathbf{w_c} \mathbf{f_c}(\mathbf{x}, \mathbf{Z}) \Big) \tag{4.7}$$

The partition function $C(\mathbf{x})$ is a normalizing factor given by Eq.17.

$$C(\mathbf{x}) = \sum_{\tilde{\mathbf{Z}}} exp\Big(\sum_{\mathbf{x} \in \mathcal{X}} \mathbf{w_c} \mathbf{f_c}(\mathbf{x}, \mathbf{Z}) \Big) \tag{4.8}$$

The objective of interaction inference problem is to find a realization of $\mathbf{X}$ denoted as $\mathbf{x}^*$, given the observation matrix $\mathbf{Z}$, that maximize the joint posterior probability $P(x|\mathbf{Z})$. The characteristic of the energy function proposed in the previous section suggests that, the energy function defined in Eq.4.6 is minimized when the interaction is likely to occur. Therefore, the weighted sum of feature functions in Eq.4.7 is replaced by negative of the energy function so that the Maximum A Posteriori (MAP) inference is transformed into the energy function minimization.

$$P(\mathbf{x}|\mathbf{Z}) = \frac{1}{C(\mathbf{x})} exp\Big(- E(\mathbf{Z}, \mathbf{x}) \Big) \tag{4.9}$$

Then the optimal realization $\mathbf{x}^*$ that best describes the interaction relationship is the realization that maximizes the posteriori probability or minimizes the energy function.

$$\mathbf{x}^* = \arg\max_{\mathbf{x} \in \mathcal{X}} P(\mathbf{x}|\mathbf{Z}) = \arg\min_{\mathbf{x} \in \mathcal{X}} E(\mathbf{Z}, \mathbf{x}) = \arg\min_{\mathbf{x} \in \mathcal{X}} \sum_{\forall (i,j) \in \mathcal{A}} \phi(\mathbf{z}_i, \mathbf{z}_j) x_{i,j} \tag{4.10}$$

Meanwhile, to enforce mutuality of interaction and restrict on maximum number of interaction for each person, d, the following structure constraint is applied during the selection of realizations,

$$x_{i,j} \in \{0, 1\} \tag{4.11}$$

$$x_{i,j} = x_{j,i} \tag{4.12}$$

$$1 \leq \sum_{\forall i \in \mathcal{N}} x_{i,j} \leq d \tag{4.13}$$

There is one trick for the structure constraint in equation 4.13. Since the objective of linear programming is to minimize the objectives given by equation 4.10, and the interaction features are strictly positive, the optimization solver tends to minimize the connection between people such that the total number of connections for each person will be limited to one. To enforce more interactions, d fake nodes are added to the interaction graph which act like an interaction threshold. By forcing the number of interactions for each person to d, the interaction will occur between two people or between one person and a fake node. The interaction graph is pruned to remove the fake nodes after interaction inference, and in such case the structure constraint is met.

The solution of Eq.4.10 is the realization of $\mathbf{X}$ with the maximum interaction strength and being mutually correlated to one another, whether the interaction exists or not. By flatting the set of random variable $\mathbf{X}$ and the corresponding interaction potential ϕ into $1D$ vectors, Eq.4.10 is reformulated to a mixed-integer linear programming problem, which is mathematically equivalent to the original problem. The flattened 1D vectors of interaction potentials ϕ_{flat} and the flattened variables to be optimized x_{flat} is given by Eq.4.14

$$\phi_{flat} = \begin{bmatrix} \phi(\mathbf{z}_1, \mathbf{z}_1) & \phi(\mathbf{z}_1, \mathbf{z}_2) & \dots & \phi(\mathbf{z}_1, \mathbf{z}_{N+d}) & \phi(\mathbf{z}_2, \mathbf{z}_1) & \dots & \phi(\mathbf{z}_{N+d}, \mathbf{z}_{N+d}) \end{bmatrix}^T$$
$$x_{flat} = \begin{bmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,N+d} & x_{2,1} & \dots & x_{N+d,N+d} \end{bmatrix}^T \tag{4.14}$$

The original optimization problem converted to a mixed-integer linear programming problem given by Eq.4.15. The constraints given by 4.16 needs to be converted according to the new shape of the optimizing variables x_{flat} and is not expanded further here.

$$\mathbf{x}^* = \min_x \phi_{flat}^T x_{flat} \tag{4.15}$$

subject to

$$\begin{aligned} x_{i,j} &\in \{0, 1\} \\ x_{i,j} &= x_{j,i} \\ \sum_{\forall i \in \mathcal{N}+\lceil} x_{i,j} &= d \end{aligned} \tag{4.16}$$

Finding the optimal solution of Eq.4.15 is an NP-hard problem, and easily becomes intractable to find a closed-form solution. Thus, Eq.4.15 is solved by the MATLAB mixed-integer linear programming solver intlinprog with constraints specified by Eq.4.16.

4.2 Attention Graph Modeling and Inference

4.2.1 Attention Energy Function

The attention energy function is described by three attention potentials, the horizontal attention potential $\psi_h(\hat{\mathbf{z}}_i, \hat{\mathbf{z}}_j)$, the vertical attention potential $\psi_v(\hat{\mathbf{z}}_i, \hat{\mathbf{z}}_j)$ and the proximity potential $\psi_d(\hat{\mathbf{z}}_i, \hat{\mathbf{z}}_j)$.

$\psi_h(\hat{\mathbf{z}}_i, \hat{\mathbf{z}}_j)$ measures the attention strength from i to j in terms of the relative pan angle ω_{ij}. $\psi_h(\hat{\mathbf{z}}_i, \hat{\mathbf{z}}_j)$ is assumed to take the value of a zero-centered Gaussian-shaped potential-function evaluated at ω_{ij}, indicating that the attention strength from i to j is stronger if the person j stands right in front of person i.

$$\psi_h(\hat{\mathbf{z}}_i, \hat{\mathbf{z}}_j) = \frac{1}{\sigma_h \sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\omega_{ij}-0}{\sigma_h}\right)^2} \tag{4.17}$$

Similarly, $\psi_v(\hat{\mathbf{z}}_i, \hat{\mathbf{z}}_j)$ measures the attention strength from i to j regarding the relative tilt angle ξ_{ij}, and is defined as,

$$\psi_v(\hat{\mathbf{z}}_i, \hat{\mathbf{z}}_j) = \frac{1}{\sigma_v \sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\xi_{ij}-0}{\sigma_v}\right)^2} \tag{4.18}$$

which implies that the attention strength from i to j is stronger if ξ_{ij} is closer to 0.

The proximity potential $\psi_d(\hat{\mathbf{z}}_i, \hat{\mathbf{z}}_j)$ follows the same potential function as defined in Eq. 4.4.

$$\psi_d(\hat{\mathbf{z}}_i, \hat{\mathbf{z}}_j) = \begin{cases} \frac{1}{\mathrm{B}(\alpha,\beta)}\left(\frac{d_{ij}}{L}\right)^{\alpha-1}\left(1-\frac{d_{ij}}{L}\right)^{\beta-1} & d_{ij} \in (0, L] \\ 0 & d_{ij} \in (L, +\infty) \end{cases} \tag{4.19}$$

Finally, the attention potential $\psi(\hat{\mathbf{z}}_i, \hat{\mathbf{z}}_j)$, reflecting the attention strength, is the product of the three attention potentials,

$$\psi(\hat{\mathbf{z}}_i, \hat{\mathbf{z}}_j) = \psi_h(\hat{\mathbf{z}}_i, \hat{\mathbf{z}}_j) * \psi_v(\hat{\mathbf{z}}_i, \hat{\mathbf{z}}_j) * \psi_d(\hat{\mathbf{z}}_i, \hat{\mathbf{z}}_j) \tag{4.20}$$

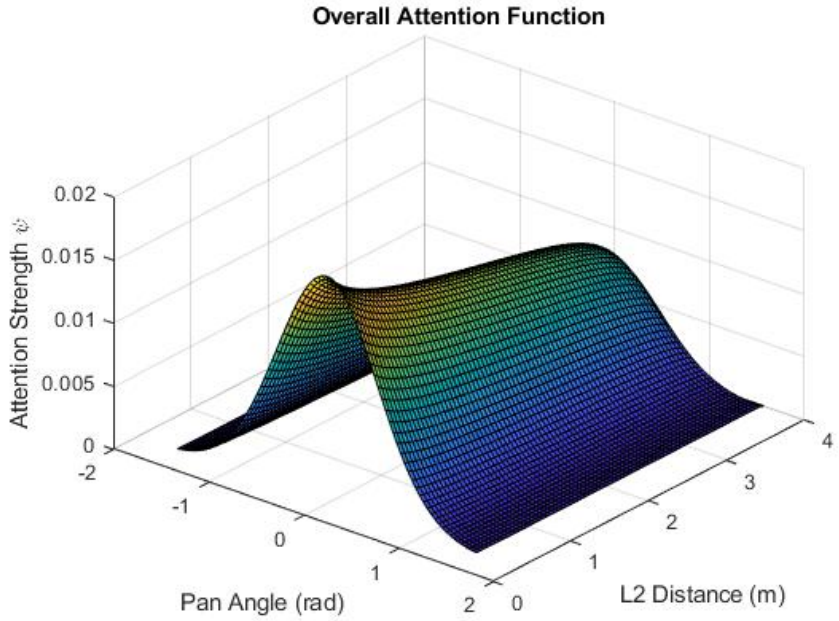


Figure 4.1: Attention Strength: horizontal attention potential is Gaussian PDF, proximity potential follows Beta distribution PDF (vertical attention is also Gaussian and is omitted for clarity)

The pattern of $\psi(\hat{\mathbf{z}}_i, \hat{\mathbf{z}}_j)$ is shown in Fig.4.1 (noticed that the vertical attention potential is set as a constant for the clarity of display).

The attention energy function is constructed as a linear combination of the realizations of the random variables $\mathbf{Y}$ and the attention features for all people in the scene,

$$\hat{E}(\hat{\mathbf{Z}}, \mathbf{y}) \triangleq \sum_{\forall (i,j) \in \mathcal{E}} \psi(\hat{\mathbf{z}}_i, \hat{\mathbf{z}}_j) y_{i,j} \tag{4.21}$$

4.2.2 Attention Inference

The objective of attention inference is to find a realization of $\mathbf{Y}$, denoted as $\mathbf{y}^* = \{y_{i,j} | i, j \in \mathcal{N}\}$ given the observations $\hat{\mathbf{Z}}$ that minimize the energy function, given by $E(\hat{\mathbf{Z}}, \mathbf{y})$, using the same derivation as Eq.4.7-4.10:

$$\mathbf{y}^* = \arg\min_{y \in Y} \hat{E}(\hat{\mathbf{Z}}, \mathbf{y}) = \arg\min_{y \in Y} \sum_{\forall (i,j) \in \mathcal{E}} \psi(\hat{\mathbf{z}}_i, \hat{\mathbf{z}}_j) y_{i,j} \tag{4.22}$$

Because attentions are directed, the inference does not require the symmetric constraints on arcs as that in the interaction graph. The constraints applied on the attention model are as follows,

$$\begin{aligned} y_{i,j} &\in \{0, 1\} \\ 1 \leq \sum_{\forall j \in \mathcal{N}} y_{i,j} &\leq d \end{aligned} \tag{4.23}$$

where the first constraint specifies that $y_{i,j}$ is a binary-valued random variable and the second constraint specifies the maximum number of attention received by each person.

The optimal configuration $\mathbf{y}^*$ is obtained by Mixed-integer linear programming with attention potential ψ by similar approaches as the interaction inference under the different constraints given by equation 4.23.

4.2.3 Emergent Leadership

Given the optimal realization of the attention model $\mathbf{y}^*$, the *emergent leader* (EL) can be readily obtained by finding the person that receives the maximum attention. Let AR_j denote the total amount of attention received by a person j,

then,

$$AR_j = \sum_{i \in N, j \neq i} y_{i,j}$$

$$EL = \arg\max_{j \in N} AR_j$$

CHAPTER 5

EXPERIMENTS AND RESULT

5.1 Accuracy of Head Pose Estimation

In this section, four state-of-art PnP algorithms, **LHM** [47], **PPnP** [29], **RPnP**[45] and **ASPnP**[81], are evaluated in an Unreal Engine Synthetic dataset. The objective of this experiment is to select the PnP algorithms with minimal error to estimated head pose from images for interaction inference.

The Unreal Engine Synthetic dataset is generated by Unreal Engine to capture the human head motions using a fixed camera from different perspective views. The videos are collected by a set of six monocular cameras surrounding the subject, as shown in Fig. 5.1 , similar to the structural configuration of the CMU panoptic studio [35].

The Unreal Engine Synthetic dataset consists of two data collections. The first data collection contains the videos of two different human models performing the head movements, as shown in Fig 5.2(a) and 5.2(b). The monocular camera system show in Fig.5.1 captures the head movement of the human subject from six different views. In total, 12 testing cases are generated to test the PnP algorithms, and the algorithm with the highest accuracy is selected to compute the 3D head pose for interaction inference.

Additionally, three testing cases are organized in the second data collection to perform a numerical analysis for each algorithm separately. Figure 5.3 shows the testing cases with rotation about each axis of the body frame: in test case *YawHead* shown in Fig. 5.3a, the human subject is turning head from left to

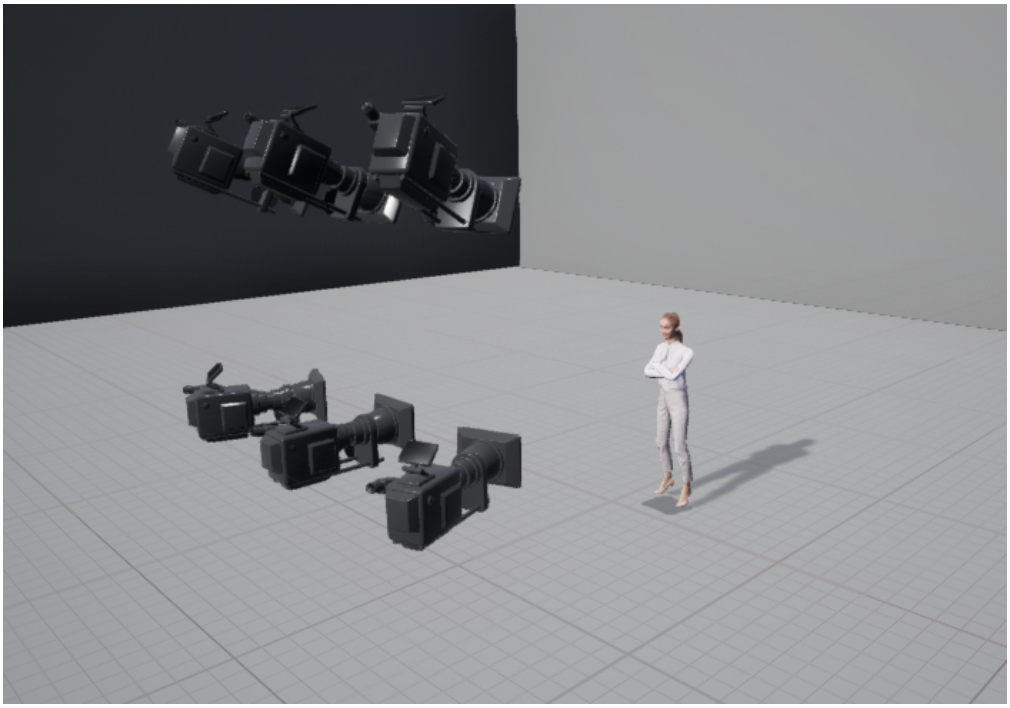


Figure 5.1: Unreal Engine Synthetic dataset data collection

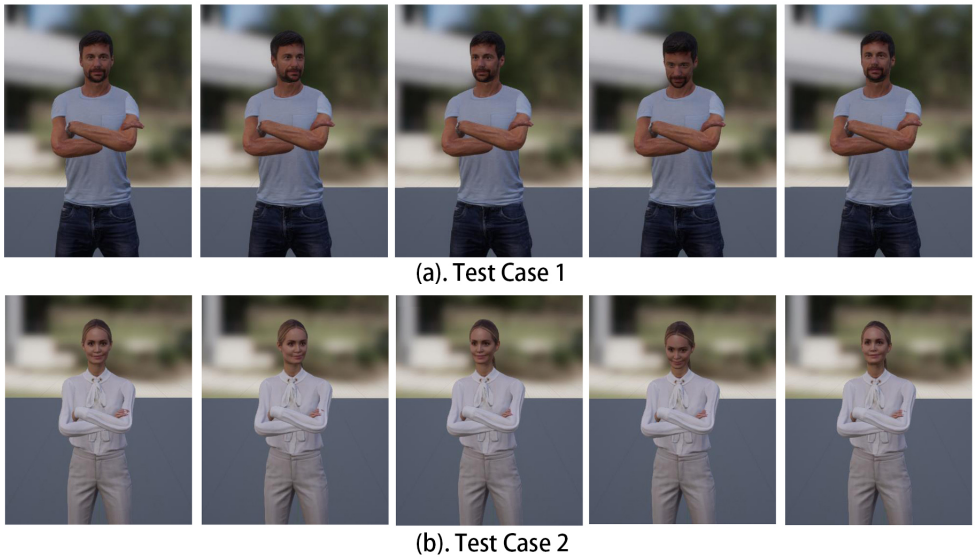


Figure 5.2: Data Collection 1: Head Movement of Two Different Human Models

(a) *YawHead* Scenario

(b) *PitchHead* Scenario

(c) *RollHead* Scenario

Figure 5.3: Data Collection 2: Three Different Head Movement Scenarios of A Human Subject: (a). *YawHead*, (b). *PitchHead* and (c). *RollHead*

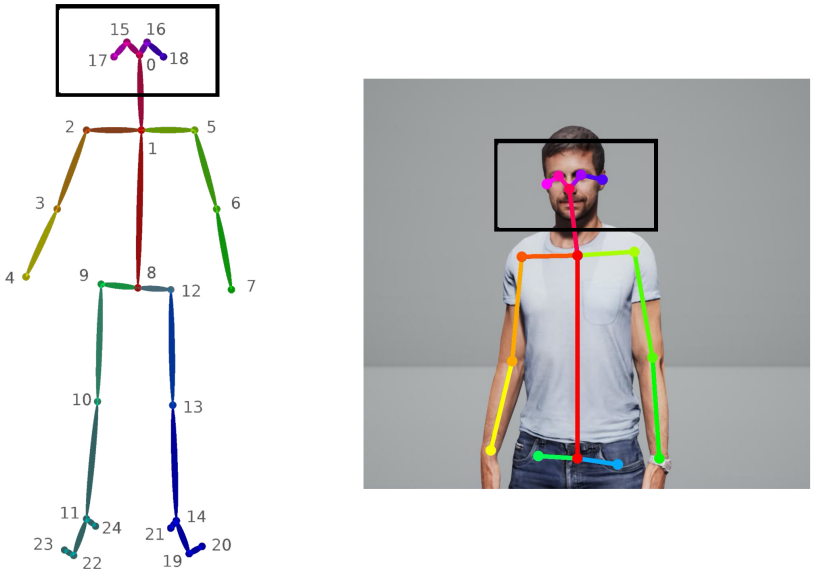


Figure 5.4: (Left) Sample result for OpenPose body keypoint detection. (Right) An example of the detection result from the video, the facial keypoints lies in the black box

right, in test case *PitchHead* shown in Fig. 5.3b, the human subject is raising head upward and downward, and in test case *RollHead* shown in Fig. 5.3c, the human subject is rotating head from left to right. This experiment is conducted to evaluate the accuracy of the PnP algorithms when the camera location and orientation varies along different axes of the human body frame.

The key points of human body in images are obtained by the estimation of OpenPose [11] keypoint detection algorithm from the rendered video taken by the camera in Unreal Engine. An example of OpenPose body keypoint detection is shown in Fig 5.4. The facial keypoints of interest are listed as follows: nose (node 0), right eye (node 15), left eye (node 16), right ear (node 17), and left ear (node 18), and are enclosed by the black boxes in the figure.

The relative 3D location of facial keypoints in the body frame is extracted from the human model's corresponding relative location in the Unreal Engine.

Table 5.1 shows the relative 3D location of the facial keypoints where the nose tip is the center of the body frame. These 3D locations as well as the location of the key points in images detected from OpenPose, are collected as the input to the PnP algorithms.

Facial Keypoints	X	Y	Z
Nose	0	0	0
Right Eye	-3	-4.1	6.2
Left Eye	3	-4.1	6.2
Right Ear	-7.5	-2.8	13.5
Left Ear	7.5	-2.8	13.5

Table 5.1: Location of Facial Keypoints (unit: cm)

For each algorithm, Mean Squared Error (MSE^R) of Euler angles and Mean Percentage Error (MPE^t) of translation are considered as metrics to evaluate the performance of each algorithm. The Mean Squared Error (MSE^R) of Euler angles measures the difference between the Euler angles in degrees correspond to the estimated rotation matrix $\mathbf{R}$ and the ground-truth rotation matrix $\mathbf{R}_{gt}$, and Mean Percentage Error (MPE^t) measures the percentage difference between the estimated translation $\mathbf{t}$ and the ground-truth translation $\mathbf{t}_{gt}$. The ground truth rotation and translation of the facial keypoints relative to the camera are collected from the Unreal Engine internal functionalities.

The MSE^R and MPE^t are defined as:

$$MSE^R = \sum_{k=1}^{3} \left[arccos(\mathbf{r}_{gt}^{\mathbf{k}} \cdot \mathbf{r}^{\mathbf{k}}) \right]^2 \tag{5.1}$$

$$MPE^t = \left\| \mathbf{t}_{gt} - \mathbf{t} \right\|_2 / \left\| \mathbf{t}_{gt} \right\|_2 \times 100\% \tag{5.2}$$

Specifically, in equation (5.1), $\mathbf{r}_{gt}^{\mathbf{k}}$ and $\mathbf{r}^{\mathbf{k}}$ are the k-th column of $\mathbf{R_{gt}}$ and $\mathbf{R}$, and $arccos(\cdot)$ represents the arc-cosine operation. In equation (5.2), $\|\cdot\|_2$ denotes the $L-2$ norm.

	LHM	RPnP	PPnP	ASPnP
MSE^R (degree^2)	25.38 ± 7.59	28.77 ± 7.96	25.23 ± 7.50	**25.11 ± 7.45**
MPE^t (%)	**2.92 ± 0.46**	3.60 ± 0.51	2.98 ± 0.48	2.96 ± 0.48

Table 5.2: First Data Collection. Comparison of PnP Algorithms: ASPnP algorithm performs the best in estimating rotations, LHM algorithm estimated the translation closest to the ground truth

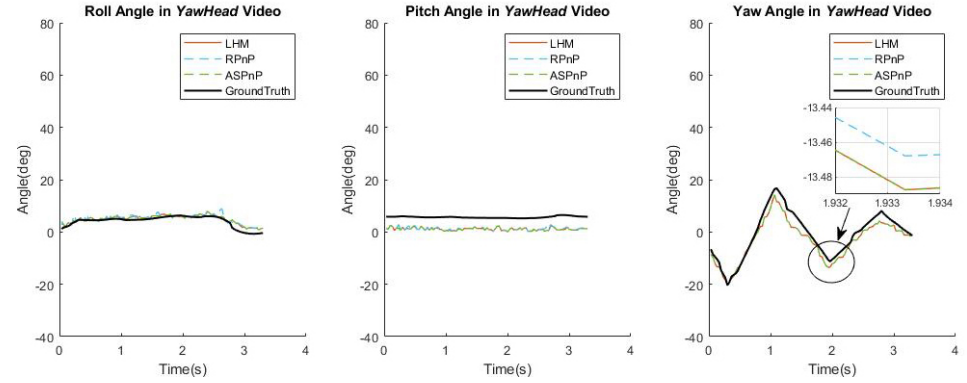
Table 5.2 shows the numeric results of the PnP algorithms. The result is averaged over the 2 human models with 6 cameras view for each human model. For each evaluation metric, a 95% confidence interval is associated with the average error. For the estimation of rotation, ASPnP algorithm outperforms other algorithms with minimal error and variance. For translation estimation, LHM algorithm best tracks the position of the human subject by 2.92% error.

The test above generally sorted the PnP algorithm by their performance in general. In the second data collection, the testing cases decoupled the relative position and orientation between the camera and human subject into three axes of the body frame. While fixing minor changes in position and orientation along two axes, the rotation along one axis is varied so that the performance of the PnP algorithms is evaluated by the estimation of rotation along each axis. For each testing case, the same evaluation metrics are applied to each algorithm, and the estimation of the best two algorithms is compared with the ground truth in a separate plot.

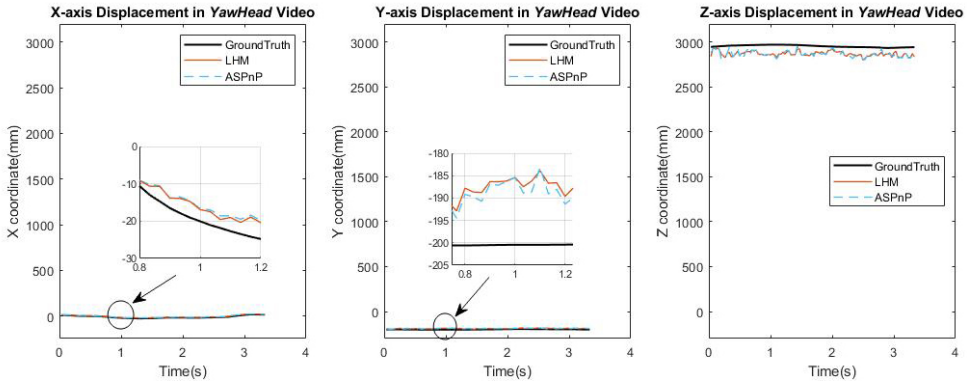
	LHM	RPnP	PPnP	ASPnP
MSE^R (degree^2)	9.93	**9.70**	9.93	9.93
MPE^t (%)	**3.02**	3.13	3.04	**3.02**

Table 5.3: Comparison of PnP Algorithms in *YawHead* scenario

Table 5.3 shows the numeric results of the PnP algorithms for the *YawHead* test case. RPnP algorithm performs the best in the estimation of rotation against



(a)



(b)

Figure 5.5: *YawHead* scenario: (a). Euler angles corresponding to the rotation estimation, the roll and pitch angle hardly changed; (b). Translation Estimation, the camera is placed 3000 mm in front of the human subject

other algorithms. The estimation error of translation for LHM and ASPnP algorithms are equally minimal. It is worth noticing that all four algorithms capture the rotation of the person with an average angle error of about 3 degrees, and the location of the person is estimated within 3% confidence.

The estimated rotation of the best three algorithms, LHM, RPnP and ASPnP, are plotted in Fig 5.5a along with the ground truth over the time window. The rotation matrix is decoupled into Euler Angles for the clarity of the demonstration. The difference between the rotation estimation of LHM (red solid line), RPnP (blue dashed line) and ASPnP(green dashed line) are minimal and are all

close to the ground truth.

Fig 5.5b shows the translation estimation for LHM and ASPnP. In the Unreal Engine Synthetic Dataset, the camera is placing 3 meters in the front of the person and 0.2 meters below the center of the person's body frame. Fig 5.5b shows the estimation given by LHM(red solid line) and ASPnP(blue dashed line); both algorithms recover the location of the human subject accurately. As a brief summary, all four algorithms locate the orientation and position of the human subject fairly when the human subject is glancing horizontally.

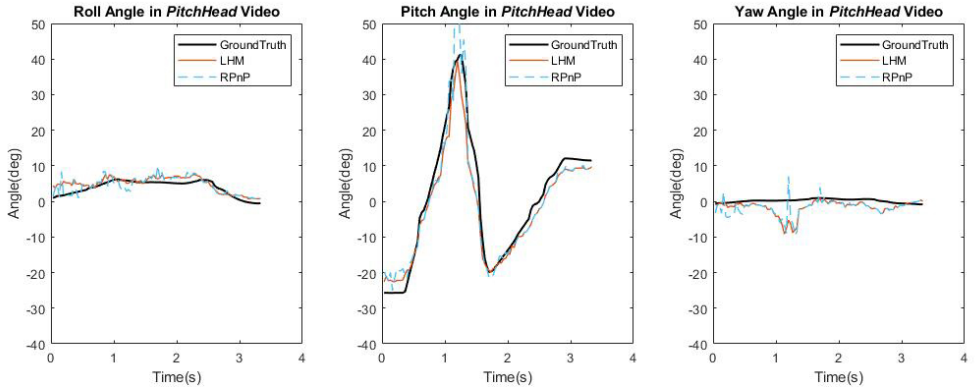
	LHM	**RPnP**	**PPnP**	**ASPnP**
MSE^R (degree^2)	**11.95**	28.48	11.99	12.44
MPE^t (%)	**4.29**	**4.29**	4.32	4.37

Table 5.4: Comparison of PnP Algorithms in *PitchHead* scenario

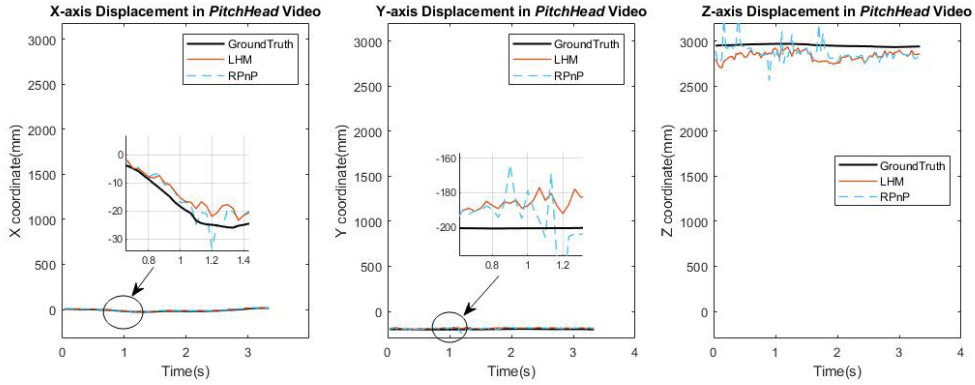
Table 5.4 shows the numeric results of the PnP algorithms for the *PitchHead* scenario. In *PitchHead* scenario, LHM algorithm performs the best in two categories against the other algorithms, the translation estimation from RPnP algorithm is also the closest to the ground truth.

In *PitchHead* scenario, the person is looking upward then downward which is demonstrated by the large angle variation in the middle plot in Fig 5.6a. Among the best two solutions, LHM (red solid line) performs better than RPnP (blue dashed line) with less noisy rotation estimation. Both algorithms perform worse when the pitch angle approaches 40 degrees as the estimation of yaw angle differs from the ground truth. By making a proper assumption on the relative pitch angle between the testing subject and the camera (e.g. placing the camera at around the same height as the subject), the error can be minimized.

Fig 5.6b shows the translation estimation result of LHM (red solid line) and



(a)



(b)

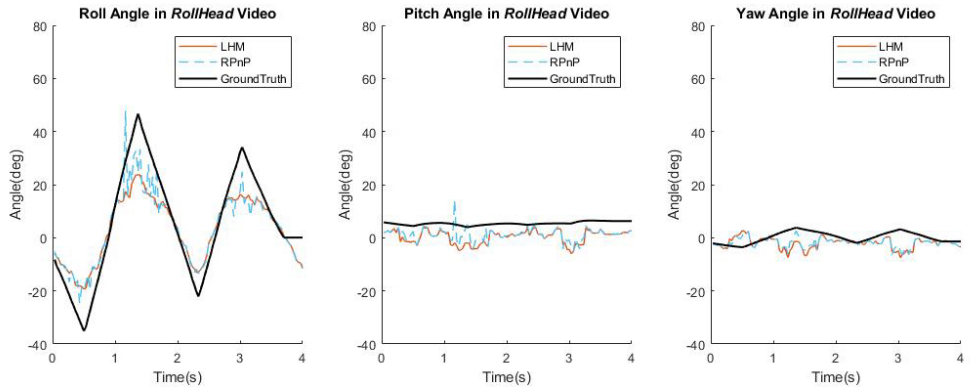
Figure 5.6: *PitchHead* scenario: (a). Euler angles corresponding to the rotation estimation, the roll and yaw angle hardly changed; (b). Translation Estimation, the camera is placed 3000 mm in front of the human subject

RPnP (blue dashed line). LHM outperforms RPnP by the less variance in estimation. The poor performance of RPnP reveals that it is highly unstable when it attempts to decouple the correspondences and solve the sub-problems separately.

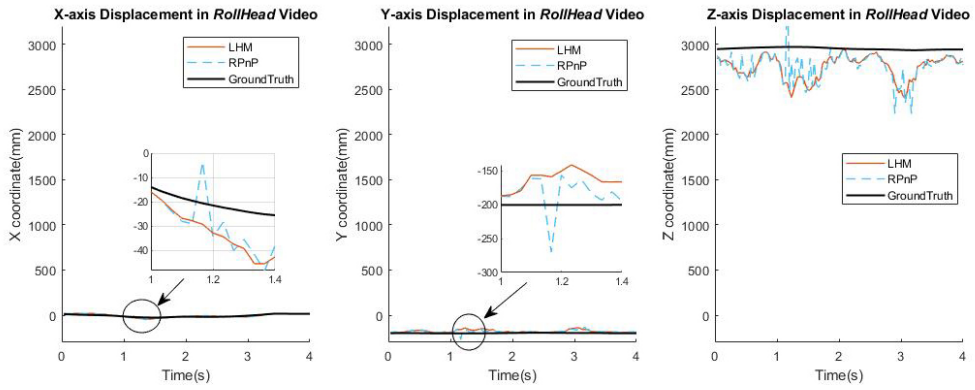
	LHM	**RPnP**	**PPnP**	**ASPnP**
MSE^R (degree^2)	11.94	28.30	11.97	12.42
MPE^t (%)	4.28	4.28	4.31	4.36

Table 5.5: Comparison of PnP Algorithms in *RollHead* scenario

Table 5.5 shows the numeric results of the PnP algorithms for the *RollHead* scenario. In the *RollHead* scenario, LHM algorithm outperforms other algorithms



(a)



(b)

Figure 5.7: *RollHead* scenario: (a). Euler angles corresponding to the rotation estimation, the pitch and yaw angle hardly changed; (b). Translation Estimation, the camera is placed 3000 mm in front of the human subject

in the estimation of translation and RPnP best estimated the rotation.

Among the four algorithms in Table 5.5, the top 2 accurate estimation of Euler angles are plotted in Fig 5.7a, i.e., the estimation of LHM and RPnP (the results of the other two methods are omitted for clarity purpose).

In the *RollHead* scenario, only the roll angle (Fig 5.7a) varies prominently over the sampling period. LHM (red solid line) estimated the roll angle smoother than RPnP (blue dashed line) where the later suffers some sudden changes in the estimation. Both algorithms perform poorly at large roll angles, but under the assumption that people usually will not roll their head by more than 20 de-

grees in daily conversations, the estimation is accurate. As for the estimation of translation as shown in Fig 5.7b, both algorithms suffer from the uncertainty of 2D keypoint detection which causes a drift in the translation estimation. RPnP algorithm (blue dashed line) is less robust and suffer from the large variance of estimation.

As a summary, the relative position and orientation can be estimated accurately by LHM, which exhibits its robustness and accuracy in estimating a person's head location and orientation. This information is further used to infer people's gaze in a group-meeting scenario and determine the emergent leader of a conversational group.

However, as the number of people in each image increases, extracting enough facial keypoints for all people from a single image gets challenging and may need more than one camera. Therefore, the data association problem across different cameras raises, which will not be explored in this thesis.

5.2 Interaction Inference

The dataset used in this thesis to perform interaction and attention inference is provided by the CMU Panoptic Dataset [35]. The CMU Panoptic Dataset provided a massive multi-view camera system from 480 VGA cameras and 31 HD cameras, capturing full-body motion. The dataset contains multiple conversation scenarios where the participants acted naturally with no behavioral restriction instructed. Meanwhile, the position movements of the participants are minimal, that the dataset concentrates on the interaction within static conversation groups. The dataset provides an accurate estimation of the 3D location

of the facial keypoints, which can be easily converted to the 3D head pose by geometric transformations.

In this section, three representative scenarios are selected to illustrate the effectiveness and shortcoming of interaction inference. In scenario 1, 100 consecutive frames are extracted from the surveillance video; in scenario 2, 50 consecutive frames are extracted, and in scenario 3, the number of frames is 30. The inference performance is evaluated over all test frames based on specific metrics compared to the hand-labeled ground truth interaction label. The yellow lines between two participants in the left figures indicate that the interaction is inferred as existing when the green line in the right figures indicate the ground truth label of interaction.

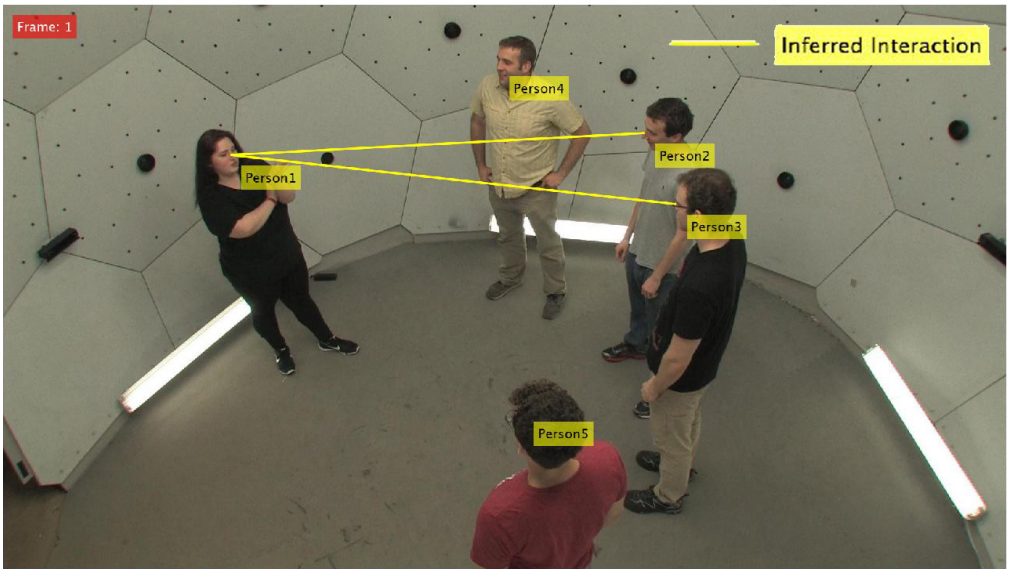


Figure 5.8: Interaction Inference Result, Scenario 1: simple conversation with one leader

In scenario 1 (Fig.5.8), four of the five participants face person 1 while the attention of person 1 lies on person 2 and 3 where the model correctly infers the interaction between them and interprets the other two people as singletons. The result indicates the constraints applied to the inference of the interaction model

enforced the reciprocity of interaction.

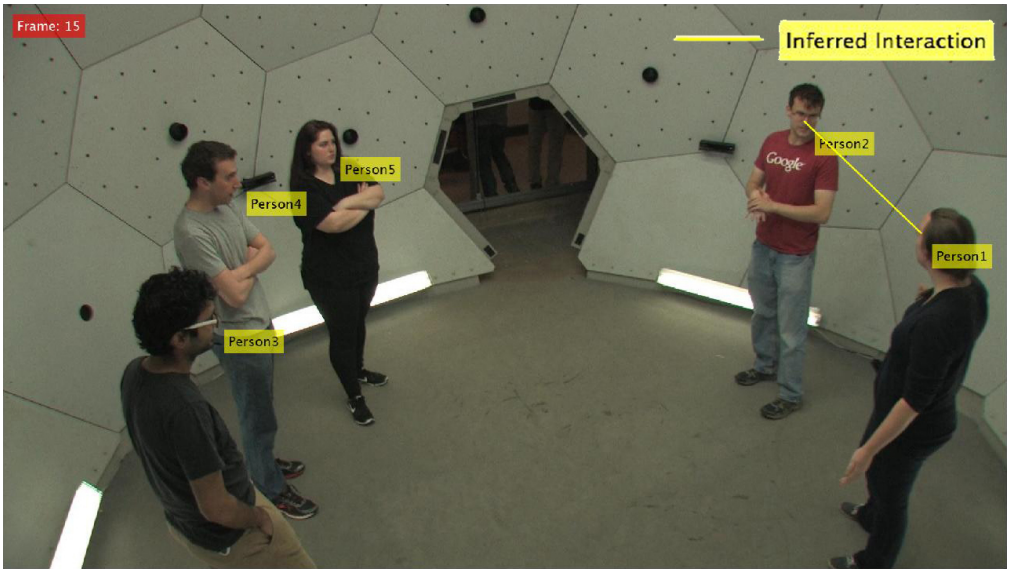


Figure 5.9: Interaction Inference Result, Scenario 2: simple conversation with bystander

As for scenario 2 (Fig.5.9), person 1 and 2 are correctly inferred to have interaction and others are singletons. Obviously, all three singletons participate in the conversation as bystanders, yet the model does not capture this information.

The circumstance in scenario 3 (Fig.5.10) is complicated where six participants forms two groups playing rock-scissors-paper while one participant is considered as the bystander. The four people in the right side of figure 5.10 forms one group where the interactions are inferred by the model, person 2 and person 4 belong to the other group. It is worth noticing that person 1 should be classified as a singleton, yet the gaze of person 4 lies ambiguously between person 1 and 2. The misconnection between person 1 and person 4 reveals a shortcoming of the interaction inference model that the symmetry of the interaction feature in interaction graph sometimes blurs the distinctiveness of individual gaze differences. The attention graph refines this shortcoming.

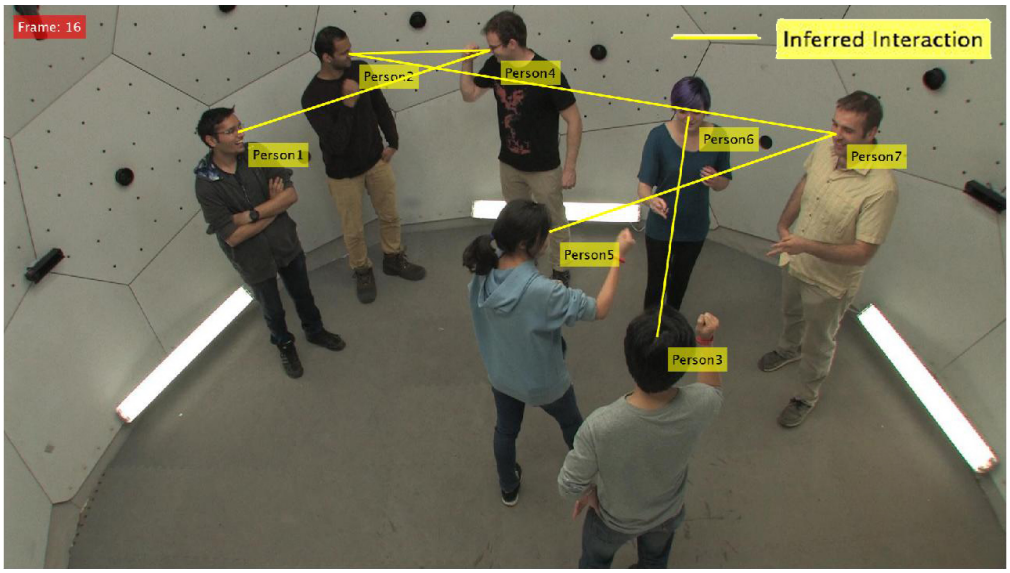


Figure 5.10: Interaction Inference Result, Scenario 3: two conversational group with one bystander

	Inference Result	
Ground Truth	Exists	Does Not Exist
Exists	TP = 514	FN = 124
Does Not Exist	FP = 76	TN = 2350

Table 5.6: Confusion Matrix of The Inference Result

The inference result for all three test scenarios are summarized in Table 5.6 as a confusion matrix. Most interaction does not exist because one human usually cannot communicate with more than 2 individuals concurrently.

The terminologies are abbreviated as: TP: True Positive (The interaction exists and is inferred correctly by the inference algorithm). FP: False Positive (The interaction does not exist but is inferred as existing by the inference algorithm). TN: True Negative (The interaction does not exist and is inferred correctly by the inference algorithm). FN: False Negative (The interaction exists but is not inferred by the inference algorithm)

	Precision	Recall
Inference Result	0.8741	0.8056

Table 5.7: Inference Precision and Recall for the Testing Scenarios

$$Recall = \frac{TP}{TP + FN} \tag{5.3}$$

$$Precision = \frac{TP}{TP + FP} \tag{5.4}$$

Table 5.7 shows the performance of the inference algorithm in terms of precision and recall. The inference precision and recall for all testing frames are averaged. The precision of inference algorithm is high, indicating that the algorithm intended to correctly infer all correct interactions but may miss some interactions undetected.

The result demonstrated above shows the effectiveness of using the relative head pose and proximity to infer interaction. The interacting pairs are the dominant participants in the group, and singletons are side participants. However, the results do not distinguish between singletons and may lose key information of the group. For example, the result of scenario 1 ignores the fact that person 4 and person 5 focus on person 1, whereas in scenario 2, the three singletons have their attention on person 1 and 2. The proposed attention graph resolves these problems by providing a comprehensive understanding of all participants' attention, which is presented in the next section.

5.3 Attention Inference

As mentioned in the previous section, the undirected interaction graph cannot describe the focus of singletons in a conversational group, yet it is not a good

representation for monologue scenarios. To address these two problems, this section proposes the result of attention graph, where a node still represents a participant, but the outgoing arcs are directed and point towards the focus of attention of the corresponding node. As a result, the attention graph shows all participants' attention, and the node that receives the most attention is perceived as the emergent leader.

5.3.1 Inference Result

The arrows indicate the direction of attention received by the person of interest inferred by the attention inference model. The inward arrows are colored in the same color as the attention receiver. Interaction can also be inferred by the attention graph, which is defined as mutual attention and is indicated by the pair of opposite directional arrows next to each other.

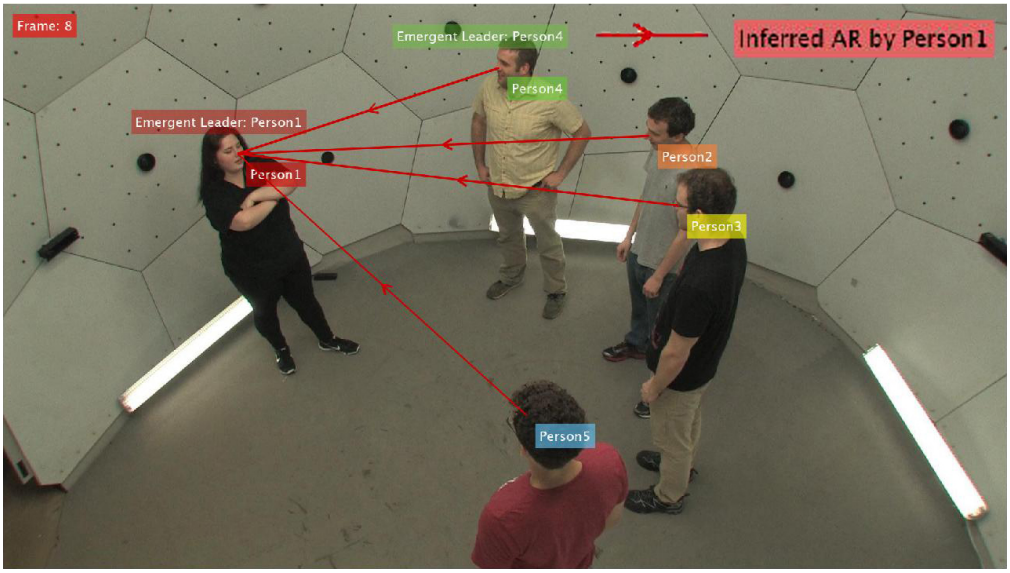


Figure 5.11: Attention Inference Result, Scenario 1: simple conversation. The emergent leader is highlighted

In scenario 1 (Fig.5.11), the mutual attention between person 2, person 3, and

person 1 inferred by the model remains indicated by the opposite arrows. Moreover, singletons' attention is also clearly indicated by directional arrows from person 4 and 5. The directional arrows form a hierarchical structure of the group where arrows point from the participants to the leader.

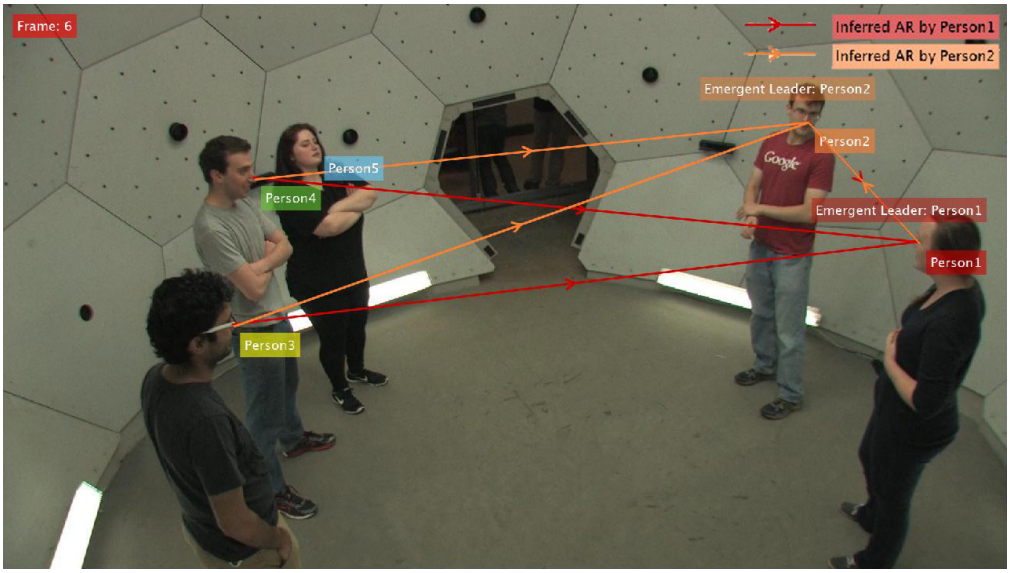


Figure 5.12: Attention Inference Result, Scenario 2: simple conversation with bystander. The two emergent leader is highlighted

Similarly, as for scenario 2 (Fig.5.12), the attention of all three singletons are involved so that they may consider as bystanders rather than participants who are completely irrelevant to the conversation.

As for a complicated case, the attention graph is more robust when inferring interactions by raising distinct attention feature for each person. The attention model disconnects the interaction between person 1 and person 4 as well as person 1 and person 7, represented by the uni-directional arrow shown in figure 5.13. Some predominant interactions inferred by the interaction graph remains in the inference result of the attention model. In all, the attention model captures the structure of the group and distinguishes between bystanders and singletons while maintaining the ability to infer interaction.

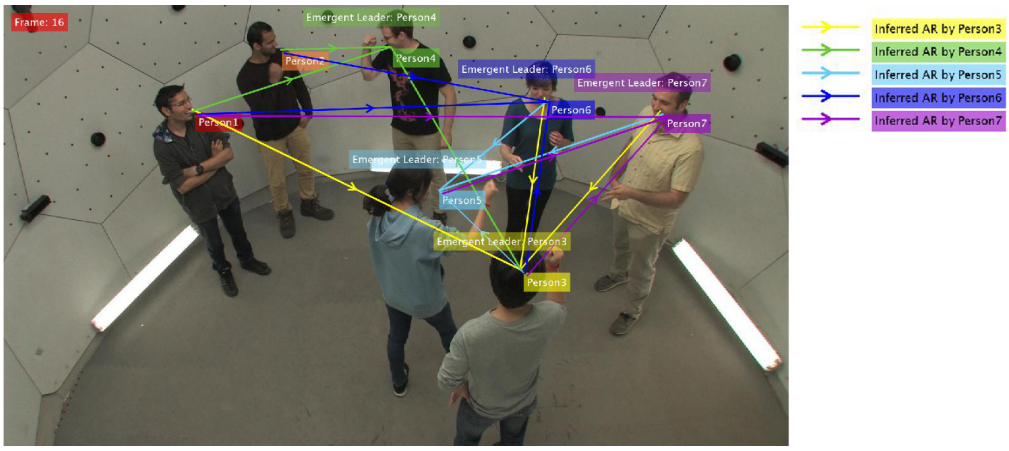


Figure 5.13: Attention Inference Result, Scenario 3: two conversational group with bystander. Multiple emergent leaders exist

5.3.2 Emergent Leader

The emergent leader of a group is the person who receives the most attention from either other leaders or bystanders. As shown in the attention inference result of scenario 1 and 2 (Fig. 5.11, Fig. 5.12), the emergent leader is the person with the maximum number of inward arrows. The emergent leadership detection is accurate in a small group as scenario 1 and scenario 2 where target of communications are obvious. The emergent leadership is less significant in understanding the dominance of a person in a crowded scenes such as scenario 3 (Fig. 5.13), where most people have a similar number of attentions from each other in a large group, yet the emergent leader of a smaller group receives less attention and is less prominent. Nevertheless, emergent leadership is only one metric evaluating the structure of the group; more information can be inferred directly or indirectly from the attention graph.

	Precision	Recall
Inference Result	1.0	0.983

Table 5.8: Emergent Leadership Inference Precision and Recall

As shown in Table 5.8, the attention graph's emergent leadership detection is accurate because the attention frustum successfully models each person's attention in all circumstances, either in a simple group or in a complex scenario.

CHAPTER 6

CONCLUSION

In this thesis, a novel baseline model using the Markov Random Field model to represent social interaction is introduced based on head pose features. The state-of-art algorithms which estimate head pose is compared by a simulated dataset and real-world dataset, the algorithm with the highest accuracy and robustness are selected to estimate head pose from detected facial key points in the image. A spatial-temporal feature describing social interaction is then proposed and con-tributes to an energy function that models the interaction strength. The optimal configuration of interaction is obtained by solving a Mixed-Integer Linear Programming optimization problem that minimizes the sum of energy function across the graph. Meanwhile, the shortcoming and limitations of the baseline model are analyzed based on sociological reasoning. A revised attention model relaxes the constraints enforced on the baseline model, which uses directed arcs to represent social attention. A discriminative attention feature that better describes the social gaze of each individual is introduced accordingly. The qualitative result of the experiment conducted on a benchmark dataset consistently shows inference results for both models. Comparing inference results for both models, concretely shows the superiority of the revised attention model over the baseline model to understand the group structure.

This thesis's possible extension is to assign weights to head pose feature and proximity feature and used machine learning approach to train the weight with the ground truth hand-labeled the inference result. Additionally, since the experiment is conducted in a static environment, the inference of group tructure under a dynamic movement is another extension of this thesis. Other kinematic

features may be involved in interpreting the scenes from a novel perspective.

BIBLIOGRAPHY

[1] Kendon Adam. *Conducting interaction: patterns of behavior in focused encounters*. Cambridge University Press, 1990.

[2] Byungtae Ahn, Dong-Geol Choi, Jaesik Park, and In So Kweon. Real-time head pose estimation using multi-task deep neural network. *Robotics and Autonomous Systems*, 103:1–12, 2018.

[3] S. O. Ba and J. Odobez. Multiperson visual focus of attention from head pose and meeting contextual cues. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(1):101–116, 2011.

[4] L. Bazzani, M. Cristani, D. Tosato, M. Farenzena, G. Paggetti, G. Menegaz, and V. Murino. Social interactions by visual focus of attention in a three-dimensional environment. *Expert Systems*, 30(2):115–127, 2013.

[5] B. Benfold and I. Reid. Stable multi-target tracking in real-time surveillance video. In *CVPR 2011*, pages 3457–3464, 2011.

[6] Philipp Bergmann, Tim Meinhardt, and Laura Leal-Taixe. Tracking without bells and whistles. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[7] Cigdem Beyan, Nicolò Carissimi, Francesca Capozzi, Sebastiano Vascon, Matteo Bustreo, Antonio Pierro, Cristina Becchio, and Vittorio Murino. Detecting emergent leader in a meeting environment using nonverbal visual features only. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, ICMI '16, page 317–324, New York, NY, USA, 2016.

[8] Guido Borghi, Riccardo Gasparini, Roberto Vezzani, and Rita Cucchiara. Embedded recurrent network for head pose estimation in car. In *2017 IEEE Intelligent Vehicles Symposium (IV)*, pages 1503–1508. IEEE, 2017.

[9] Guillem Braso and Laura Leal-Taixe. Learning a neural solver for multiple object tracking. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[10] Ying Cai, Meng-long Yang, and Jun Li. Multiclass classification based on a deep convolutional network for head pose estimation. *Frontiers of Information Technology & Electronic Engineering*, 16(11):930–939, 2015.

[11] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[12] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7291–7299, 2017.

[13] M. Chang, N. Krahnstoever, and W. Ge. Probabilistic group-level motion analysis and scenario recognition. In *2011 International Conference on Computer Vision*, pages 747–754, 2011.

[14] Wongun Choi, Yu-Wei Chao, Caroline Pantofaru, and Silvio Savarese. Discovering groups of people in images. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 417–433, Cham, 2014.

[15] Wongun Choi and Silvio Savarese. A unified framework for multi-target tracking and collective activity recognition. In *European Conference on Computer Vision*, pages 215–230. Springer, 2012.

[16] Wongun Choi, Khuram Shahid, and Silvio Savarese. What are they doing?: Collective activity classification using spatio-temporal relationship among people. In *2009 IEEE 12th international conference on computer vision workshops, ICCV Workshops*, pages 1282–1289. IEEE, 2009.

[17] Wongun Choi, Khuram Shahid, and Silvio Savarese. Learning context for collective activity recognition. In *CVPR 2011*, pages 3273–3280. IEEE, 2011.

[18] Aaron V Cicourel. Cognitive sociology: Language and meaning in social interaction. *American Journal of Sociology*, 1974.

[19] Claudio Coppola, Serhan Cosar, Diego R Faria, and Nicola Bellotto. Automatic detection of human interactions from rgb-d data for social activity classification. In *2017 26th IEEE international symposium on robot and human interactive communication (RO-MAN)*, pages 871–876. IEEE, 2017.

[20] Marco Cristani, Loris Bazzani, Giulia Paggetti, Andrea Fossati, Diego Tosato, Alessio Del Bue, Gloria Menegaz, and Vittorio Murino. Social interaction discovery by statistical analysis of f-formations. In *BMVC*, volume 2, 2011.

[21] Marco Cristani, Ramachandra Raghavendra, Alessio Del Bue, and Vittorio Murino. Human behavior analysis in video surveillance: A social signal processing perspective. *Neurocomputing*, 100:86–97, 2013.

[22] Zhiwei Deng, Mengyao Zhai, Lei Chen, Yuhao Liu, Srikanth Muralidharan, Mehrsan Javan Roshtkhari, and Greg Mori. Deep structured models for group activity recognition. *British Machine Vision Conference (BMVC)*, 2015.

[23] Junyi Dong, Pingping Zhu, and Silvia Ferrari. Oriented pedestrian social interaction modeling and inference. In *2020 American Control Conference (ACC)*, pages 1373–1370. IEEE, 2020.

[24] Michela Farenzena, Loris Bazzani, Vittorio Murino, and Marco Cristani. Towards a subject-centered analysis for automated video surveillance. In *International Conference on Image Analysis and Processing*, pages 481–489. Springer, 2009.

[25] S. Feese, A. Muaremi, B. Arnrich, G. Troster, B. Meyer, and K. Jonas. Discriminating individually considerate and authoritarian leaders by speech activity cues. In *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*, pages 1460–1465, 2011.

[26] Luis Ferraz, Xavier Binefa, and Francesc Moreno-Noguer. Very fast solution to the pnp problem with algebraic outlier rejection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 501–508, 2014.

[27] Luis Ferraz Colomina, Xavier Binefa, and Francesc Moreno-Noguer. Leveraging feature uncertainty in the pnp problem. In *Proceedings of the BMVC 2014 British Machine Vision Conference*, pages 1–13, 2014.

[28] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981.

[29] Valeria Garro, Fabio Crosilla, and Andrea Fusiello. Solving the pnp problem with anisotropic orthogonal procrustes analysis. In *2012 Second International Conference on 3D Imaging, Modeling, Processing, Visualization & Transmission*, pages 262–269. IEEE, 2012.

[30] W. Ge, R. T. Collins, and B. Ruback. Automatically detecting the small

group structure of a crowd. In *2009 Workshop on Applications of Computer Vision (WACV)*, pages 1–8, 2009.

[31] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[32] Joel A Hesch and Stergios I Roumeliotis. A direct least-squares (dls) method for pnp. In *2011 International Conference on Computer Vision*, pages 383–390. IEEE, 2011.

[33] Hayley Hung, Dinesh Babu Jayagopi, Sileye Ba, Jean-Marc Odobez, and Daniel Gatica-Perez. Investigating automatic dominance estimation in groups from visual attention and speaking activity. In *Proceedings of the 10th International Conference on Multimodal Interfaces*, ICMI '08, page 233–236, New York, NY, USA, 2008. Association for Computing Machinery.

[34] Shoichi Inaba and Yoshimitsu Aoki. Conversational group detection based on social context using graph clustering algorithm. In *2016 12th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)*, pages 526–531. IEEE, 2016.

[35] Hanbyul Joo, Tomas Simon, Xulong Li, Hao Liu, Lei Tan, Lin Gui, Sean Banerjee, Timothy Scott Godisart, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic studio: A massively multiview system for social interaction capture. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[36] Shyamgopal Karthik, Ameya Prabhu, and Vineet Gandhi. Simple unsupervised multi-object tracking. *ArXiv*, abs/2006.02609, 2020.

[37] Adam Kendon. *Conducting interaction: Patterns of behavior in focused encounters*, volume 7. CUP Archive, 1990.

[38] Yuki Kizumi, Koh Kakusho, Takeshi Okadome, Takuya Funatomi, and Masaaki Iiyama. Detection of social interaction from observation of daily living environments. In *The First International Conference on Future Generation Communication Technologies*, pages 162–167. IEEE, 2012.

[39] T. Klinger, F. Rottensteiner, and C. Heipke. Probabilistic multi-person localisation and tracking in image sequences. *ISPRS Journal of Photogrammetry and Remote Sensing*, 127:73 – 88, 2017. Geospatial Week 2015.

[40] Tian Lan, Yang Wang, Weilong Yang, and Greg Mori. Beyond actions: Discriminative models for contextual group activities. In J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 1216–1224. Curran Associates, Inc., 2010.

[41] Michael J. V. Leach, Rolf Baxter, Neil M. Robertson, and Ed P. Sparks. Detecting social groups in crowded surveillance videos using visual attention. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2014.

[42] Michael J.V. Leach, Ed.P. Sparks, and Neil M. Robertson. Contextual anomaly detection in crowded surveillance scenes. *Pattern Recognition Letters*, 44:71 – 79, 2013. Pattern Recognition and Crowd Analysis.

[43] L. Leal-Taixé, G. Pons-Moll, and B. Rosenhahn. Everybody needs somebody: Modeling social and grouping behavior on a linear programming multiple people tracker. In *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, pages 120–127, 2011.

[44] Vincent Lepetit, Francesc Moreno-Noguer, and Pascal Fua. Epnp: An accurate o (n) solution to the pnp problem. *International journal of computer vision*, 81(2):155, 2009.

[45] Shiqi Li, Chi Xu, and Ming Xie. A robust o (n) solution to the perspective-n-point problem. *IEEE transactions on pattern analysis and machine intelligence*, 34(7):1444–1450, 2012.

[46] Xiabing Liu, Wei Liang, Yumeng Wang, Shuyang Li, and Mingtao Pei. 3d head pose estimation with convolutional neural network trained on synthetic images. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 1289–1293. IEEE, 2016.

[47] C-P Lu, Gregory D Hager, and Eric Mjolsness. Fast and globally convergent pose estimation from video images. *IEEE transactions on pattern analysis and machine intelligence*, 22(6):610–622, 2000.

[48] Kouichi Murakami and Hitomi Taguchi. Gesture recognition using recurrent neural networks. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 237–242, 1991.

[49] Erik Murphy-Chutorian and Mohan Manubhai Trivedi. Head pose estima-

tion in computer vision: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 31(4):607–626, 2008.

[50] J. Odobez and S. Ba. A cognitive and unsupervised map adaptation approach to the recognition of the focus of attention from head pose. In *2007 IEEE International Conference on Multimedia and Expo*, pages 1379–1382, 2007.

[51] Massimiliano Patacchiola and Angelo Cangelosi. Head pose estimation in the wild using convolutional neural networks and adaptive gradient methods. *Pattern Recognition*, 71:132–143, 2017.

[52] Stefano Pellegrini, Andreas Ess, Konrad Schindler, and Luc Van Gool. You'll never walk alone: Modeling social behavior for multi-target tracking. In *2009 IEEE 12th International Conference on Computer Vision*, pages 261–268. IEEE, 2009.

[53] Stefano Pellegrini, Andreas Ess, and Luc Van Gool. Improving data association by joint modeling of pedestrian trajectories and groupings. In Kostas Daniilidis, Petros Maragos, and Nikos Paragios, editors, *Computer Vision – ECCV 2010*, pages 452–465, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg.

[54] Fabio Pianesi, Nadia Mana, Alessandro Cappelletti, Bruno Lepri, and Massimo Zancanaro. Multimodal recognition of personality traits in social interactions. In *Proceedings of the 10th international conference on Multimodal interfaces*, pages 53–60, 2008.

[55] Z. Qin and C. R. Shelton. Improving multi-target tracking via social grouping. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1972–1978, 2012.

[56] Z. Qin and C. R. Shelton. Social grouping for multi-target tracking and head pose estimation in video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(10):2082–2095, 2016.

[57] Chirag Raman and Hayley Hung. Towards automatic estimation of conversation floors within f-formations. In *2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, pages 175–181. IEEE, 2019.

[58] Omar Adair Islas Ramírez, Giovanna Varni, Mihai Andries, Mohamed

Chetouani, and Raja Chatila. Modeling the dynamics of individual behaviors for group detection in crowds using low-level features. In *2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 1104–1111. IEEE, 2016.

[59] Romer Rosales and Stan Sclaroff. Inferring body pose without tracking body parts. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2000 (Cat. No. PR00662)*, volume 2, pages 721–727. IEEE, 2000.

[60] Nataniel Ruiz, Eunji Chong, and James M Rehg. Fine-grained head pose estimation without keypoints. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 2074–2083, 2018.

[61] Dairazalia Sanchez-Cortes, Oya Aran, Marianne Schmid Mast, and Daniel Gatica-Perez. Identifying emergent leadership in small groups using nonverbal communicative cues. In *International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction*, ICMI-MLMI '10, New York, NY, USA, 2010.

[62] Weston Sewell and Oleg Komogortsev. Real-time eye gaze tracking with an unmodified commodity webcam employing a neural network. In *CHI '10 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '10, page 3739–3744, New York, NY, USA, 2010. Association for Computing Machinery.

[63] Amarjot Singh, Devendra Patil, Meghana Reddy, and SN Omkar. Disguised face identification (dfi) with facial keypoints using spatial fusion convolutional network. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 1648–1655, 2017.

[64] R. Stiefelhagen. Tracking focus of attention in meetings. In *Proceedings. Fourth IEEE International Conference on Multimodal Interfaces*, pages 273–280, 2002.

[65] Ramanathan Subramanian, Yan Yan, Jacopo Staiano, Oswald Lanz, and Nicu Sebe. On the relationship between head pose, social attention and personality prediction for unstructured and dynamic group interactions. In *Proceedings of the 15th ACM on International conference on multimodal interaction*, pages 3–10, 2013.

[66] Yi Sun, Xiaogang Wang, and Xiaoou Tang. Deep convolutional network

cascade for facial point detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3476–3483, 2013.

[67] Khai Tran, Apurva Gala, Ioannis Kakadiaris, and S. Shah. Activity analysis in crowded environments using social cues for group discovery and human interaction modeling. *Pattern Recognition Letters*, 44:49–57, 07 2014.

[68] Khai N Tran, Xu Yan, Ioannis A Kakadiaris, and Shishir K Shah. A group contextual model for activity recognition in crowded scenes. In *VISAPP (2)*, pages 5–12, 2015.

[69] Steffen Urban, Jens Leitloff, and Stefan Hinz. Mlpnp-a real-time maximum likelihood solution to the perspective-n-point problem. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Science*, 2016.

[70] Sebastiano Vascon, Eyasu Z Mequanint, Marco Cristani, Hayley Hung, Marcello Pelillo, and Vittorio Murino. Detecting conversational groups in images and sequences: A robust game-theoretic approach. *Computer Vision and Image Understanding*, 143:11–24, 2016.

[71] Alessandro Vinciarelli, Maja Pantic, Dirk Heylen, Catherine Pelachaud, Isabella Poggi, Francesca D'Errico, and Marc Schroeder. Bridging the gap between social animal and unsocial machine: A survey of social signal processing. *IEEE Transactions on Affective Computing*, 3(1):69–87, 2011.

[72] Yujia Wang, Wei Liang, Jianbing Shen, Yunde Jia, and Lap-Fai Yu. A deep coarse-to-fine network for head pose estimation from synthetic data. *Pattern Recognition*, 94:196–206, 2019.

[73] Jarosław Was, Bartłomiej Gudowski, and Paweł J. Matuszyk. Social distances model of pedestrian dynamics. In Samira El Yacoubi, Bastien Chopard, and Stefania Bandini, editors, *Cellular Automata*, pages 492–501, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg.

[74] K. Yamaguchi, A. C. Berg, L. E. Ortiz, and T. L. Berg. Who are you with and where are you going? In *CVPR 2011*, pages 1345–1352, 2011.

[75] F. Yang, F. Li, Y. Wu, S. Sakti, and S. Nakamura. Using panoramic videos for multi-person localization and tracking in a 3d panoramic coordinate. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1863–1867, 2020.

[76] Haanju Yoo, Taekyu Eom, Jeongmin Seo, and Sang-Il Choi. Detection of interacting groups based on geometric and social relations between individuals in an image. *Pattern Recognition*, 93:498 – 506, 2019.

[77] Yixiao Yun, Mohamed H Changrampadi, and Irene YH Gu. Head pose classification by multi-class adaboost with fusion of rgb and depth images. In *2014 International Conference on Signal Processing and Integrated Networks (SPIN)*, pages 174–177. IEEE, 2014.

[78] Guyue Zhang, Jun Liu, Hengduo Li, Yan Qiu Chen, and Larry S Davis. Joint human detection and head pose estimation via multistream networks for rgb-d videos. *IEEE Signal Processing Letters*, 24(11):1666–1670, 2017.

[79] Lu Zhang and Hayley Hung. Beyond f-formations: Determining social involvement in free standing conversing groups from static images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1086–1095, 2016.

[80] Yinqiang Zheng, Yubin Kuang, Shigeki Sugimoto, Kalle Astrom, and Masatoshi Okutomi. Revisiting the pnp problem: A fast, general and optimal solution. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2344–2351, 2013.

[81] Yinqiang Zheng, Shigeki Sugimoto, and Masatoshi Okutomi. Aspnp: An accurate and scalable solution to the perspective-n-point problem. *IEICE TRANSACTIONS on Information and Systems*, 96(7):1525–1535, 2013.