Panoptic Studio: A Massively Multiview System for Social Motion Capture *

Hanbyul Joo Hao Liu† Lei Tan‡ Lin Gui† Bart Nabbe
Iain Matthews§ Takeo Kanade Shohei Nobuhara‖ Yaser Sheikh
The Robotics Institute, Carnegie Mellon University
{hanbyulj,bana,tk,yaser}@cs.cmu.edu, {liu.hao,gui.lin}@ouc.edu.cn, tanlei@hnu.edu.cn
iainm@disneyresearch.com, nob@i.kyoto-u.ac.jp

Abstract

We present an approach to capture the 3D structure and motion of a group of people engaged in a social interaction. The core challenges in capturing social interactions are: (1) occlusion is functional and frequent; (2) subtle motion needs to be measured over a space large enough to host a social group; and (3) human appearance and configuration variation is immense. The Panoptic Studio is a system organized around the thesis that social interactions should be measured through the perceptual integration of a large variety of view points. We present a modularized system designed around this principle, consisting of integrated structural, hardware, and software innovations. The system takes, as input, 480 synchronized video streams of multiple people engaged in social activities, and produces, as output, the labeled time-varying 3D structure of anatomical landmarks on individuals in the space. The algorithmic contributions include a hierarchical approach for generating skeletal trajectory proposals, and an optimization framework for skeletal reconstruction with trajectory reassociation.

1. Introduction

There is a prevailing scientific consensus that nearly two-thirds of interpersonal communication is transmitted via nonverbal cues [6, 33]. Yet, despite the fundamental role these cues play in enabling social function, the protocol underlying this communication is poorly understood—Sapir [35] called it "an elaborate code that is written nowhere, known to no one, and understood by all". Some structures of this code have been identified through observational study, such as reciprocity [7] or synchrony [10]. However, systematic studies of such phenomena have remained almost entirely focused on the analysis of facial expressions, despite emerging evidence [28, 3] that facial expressions provide a fundamentally *incomplete* characterization of nonverbal communication. One proximal cause for this singular focus on the face is that capturing natural social interaction presents challenges that current state-of-the-art motion capture systems simply cannot address.

There are three principal challenges in capturing social signaling between individuals in a group: (1) subtle motion has to be measured over a volume sufficient to house a dynamic social group; (2) strong occlusions functionally emerge in natural social interactions (e.g., people systematically face each other while interacting, bodies are occluded by gesticulating limbs); (3) social signaling is sensitive to interference. For instance, attaching markers to the face or body, a pre-capture model building stage, or even instructing each individual to assume a canonical body pose during an interaction, primes the nature of subsequent interactions.

In this paper, we present a system designed to address these issues, with integrated innovations in hardware design, motion representation, and motion reconstruction. The organizing principle is that social motion capture should be performed by the consolidation of a large number of "weak" perceptual processes rather than the analysis of a few sophisticated sensors. The large number of views provide robustness to occlusions, provide precision over the capture space, and facilitate the boosting of weak 2D human pose detectors into a strong 3D skeletal tracker. In particular, our contributions include:

1. **Modularized Hardware**: We present the modular design of a massively multiview capture consisting of 480 simultaneously triggered VGA cameras, distributed over the surface of 5.49m geodesic sphere (sufficient to house social groups).

2. **Skeletal Representation**: We present a new representation for social motion capture labeling and embedding a dense 3D trajectory stream within a moving

*http://www.cs.cmu.edu/~hanbyulj/panoptic-studio
†Ocean University of China
‡Hunan University
§Disney Research Pittsburgh
‖Kyoto University

skeletal frame for each individual.

3. **3D Motion Reconstruction Algorithm**: To the best of our knowledge, our method is the first to fully automatically capture the subtle interactions of multiple people in a social group (more than 6 people) without requiring any individual body calibration or markers. To be scalable to a large number of participants, our method avoids subject-specific templates such as body shape, color, and bone length, while providing high accuracy without jitter.

4. **Social Games Dataset**: We collect a novel dataset consisting of 5 vignettes, where multiple people are engaged in social games (*Ultimatum*, *Mafia*, *Haggle*, *007-bang game*). The data are captured by our hardware system with synchronized and calibrated 480 cameras. We also provide multiple Kinect data for a subset of the vignettes calibrated and time-aligned with the 480 cameras for comparison. All the data and results are publicly shared in our website*.

The system described in this paper provides empirical data of unprecedented resolution with the promise of facilitating data-driven exploration of scientific conjectures about the communication code of social behavior.

2. Related Work

Almost as soon as they were invented, cameras have been used to study social interaction. Darwin, in his foundational treatise on the expression of emotion, used photographs to prompt participant response to expressions [11]. Since then, photographs have been–and continue to be–a fundamental tool in studying social behavior [19, 21, 13, 41, 34, 9, 1]. When the video camera was invented, it too became an integral tool to study the dynamics of social interaction [29, 43]. Most recently, with the rapid proliferation of smart phone cameras, crowd capture is an emerging medium for analyzing social behavior as it measures both the attentive behavior of social groups, as well as their interactive dynamics [2, 15, 30, 31].

Multi-camera systems have been used to measure the 3D structure and motion of human motion. Kanade et al. [23] pioneered the use of multi-view sensing systems to "virtualize" reality, using 51 cameras mounted on geodesic dome that was 5 meters in diameter. A number of systems were subsequently proposed to produce realtime virtualizations [26, 25, 18, 32]. To obtain greater detail in the 3D reconstruction, de Aguiar et al. [12], Vlasic et al. [39], and Furukawa and Ponce [16] deformed pre-defined templates of fixed topology to recover details that were subsampled or occluded in the set of views at a time instant.

Markerless motion capture methods have focused on tracking human body motion in multi-camera systems. One direction of approaches pursue high-quality motion capture in well-controlled studio setups [37, 17, 14, 38, 24]. In these approaches, articulated 3D models are often used to utilize subject-specific information such as shape, color, and bone-length. These methods usually require a template generation process, and initial alignment at the beginning of each capture. Much of this work assumes individual activity only, but few exceptions consider occlusions caused by other objects or individuals [14, 38]. In the work of [24], Liu et al. tackle motion tracking of three interacting people, where individual specific color and appearance information are used to resolve the occlusions. Recently, other approaches consider markerless motion capture in more general setups with simpler model assumptions [20, 8, 5].

Besides RGB cameras, marker-based motion capture methods provide precise dynamics measurements and have also been used to study social behavior [27], despite the interference caused by markers on social signaling. Depth sensors such as the Kinect [36, 4] are also emerging as a promising sensing modality, yet they suffer from the frequent occlusions of social interaction.

3. Modular Massively Multiview Capture

We present a massively multiview system designed to reconstruct the labelled time-varying 3D structure and motion of multiple people engaged in a social interaction. The complexity of human motion and frequent occlusions within social groups cause failures in estimating their structure and motion. To handle these challenges, our system uses 480 synchronized cameras mounted over the surface of a geodesic dome, providing redundancy for weak perceptual processes (such as pose detection and tracking) and robustness to occlusion. The large number of cameras placed at unique viewpoints also provides a working volume sufficient for multiple interacting people. The cameras are arranged uniformly to observe the scene from all directions, so that the subjects' motion is not restricted by a predefined dominant system direction. The system produces 29.4 Gbps, and to handle this we present a modularized architecture for parallel and distributed capture and processing. In this section, we describe the modular design of the studio structure and architecture consisting of the acquisition, communication, and synchronization, as shown in Figure 1 and Figure 2.

3.1. Structural Design

The physical frame of the studio is a face-transitive solid called a truncated pentagonal hexecontahedron. This particular structure was selected because it has among the largest number of transitive faces of any geodesic dome [40]. The transitivity of the faces enables the modular architecture, and ensures that the structure remains easy to upgrade and customize with different panels of the same configuration.

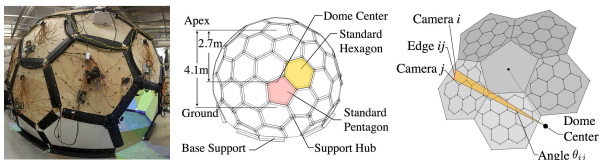


Figure 1. The studio structure. (Left) The exterior of the dome with the equipment mounted on the surface; (Center) The panels are face-transitive to ensure interchangeability across panels; (Right) An optimization was performed to ensure uniform angles with respect to the center between each camera and all its neighbors (e.g., Camera i is a neighbor of Camera j).

The structure has a radius of 5.49m and a total height of 4.15m. The center of the dome is at a height of 1.40m, and it was raised above a hemisphere to allow increased access to the edges of the dome as shown in Figure 1. In all, the structure consists of 6 pentagonal panels, 40 hexagonal panels, and 10 trimmed base panels.

Our design was modularized so that each hexagonal panel houses a set of 24 VGA cameras. To determine the placement of the cameras, we initialized their positions by tessellating the hexagon face into 24 triangles and using this initialization to define a 3-neighborhood structure shown in the right-most panel of Figure 1(c). Using this neighborhood structure and the initialization we determine the placement of the cameras over the geodesic dome by minimizing the difference in angles between all neighbors of every camera,

$$\{\theta_{ij}\}^* = \arg\min_{\{\theta_{ij}\}} \sum_{p=1}^{P} \sum_{i=1}^{N} \sum_{j \in \mathcal{N}(i)} \sum_{k \in \mathcal{N}(i) \neq j} (r(\theta_{ij}|p) - r(\theta_{ik}|p))^2,$$

where $P = 20$ is the number of panels, $N = 24$ is the number of cameras in each panel, $\mathcal{N}(\cdot)$ is the neighborhood of a camera, $r(\cdot|p)$ is a function transforming the angle on a reference panel to the p-th panel. The cameras sample the span of the vertical axis of the space and sample 48.71° of the horizontal axis. With this distribution, the minimum baseline between any camera and its nearest three neighbors is 21.05cm.

3.2. System Architecture

Figure 2 shows the architecture of our system which consists of 480 cameras. The 480 cameras are arranged modularly with 24 cameras in each of 20 standard hexagonal panels on the dome. Each module in each panel is managed by a Distributed Module Controller (DMC) that triggers all cameras in the module, receives data from them, and consolidates the video for transmission to the local machine. Each individual camera is a global shutter CMOS sensor, with a fixed focal length of 4.5mm, that captures VGA (640×480) resolution images at 25Hz.

Cameras of each panel produce an uncompressed video stream at 1.47 Gbps, and, thus, for the entire set of 480

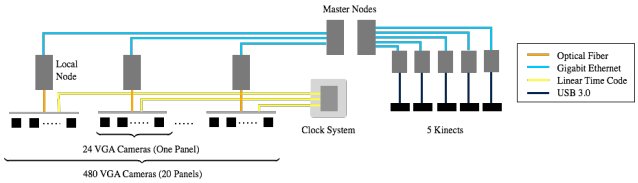


Figure 2. Modularized system architecture. The studio houses 480 cameras synchronized to a central clock system and controlled by a master node. 5 kinects are also located in the studio calibrated in the same coordinate with cameras.

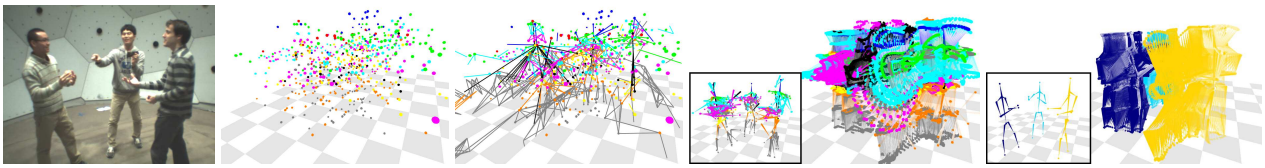
cameras the data-rate is approximately 29.4 Gbps. To handle this stream, the system pipeline has been designed with a modularized communication and control structure. For each subsystem, the clock generator sends a frame counter, trigger signal, and the pixel clock signal to each DMC associated with a panel. The DMC uses this timing information to initiate and synchronize capture of all cameras within the module. Upon trigger and exposure, each of the 24 camera heads transfers back image data via the camera interconnect to the DMC, which consolidates the image data and timing from all cameras. This composite data is then transferred via optical interconnect to the module node, where it is stored locally. Each module node has dual purpose: it serves as a distributed RAID storage unit [1] and participates as a multi-core computational node in a cluster. All the local nodes of our system are on a local network on a gigabit switcher. The acquisition is controlled via a master node that the system operator can use to control all functions of the studio.

4. Notation and Overview

Our algorithm takes, as input, 480 videos of a social interaction (with calibration and time-stamps) and, as output, produces *skeletal trajectories* with an associated set of labeled 3D trajectories for each body part. We present a bottom-up sampling-based approach that fuses low-level appearance and motion cues of local landmarks into progressively compounded constructions—from node proposals (e.g., left shoulder), to part proposals (e.g., upper arm), to part trajectory proposals (e.g., rigid motion of the upper arm), to skeletal trajectory proposal (e.g., multi-part motion of one individual).

To produce evidence of the location of different anatomical landmarks, we compute appearance-based 2D human pose detection [42] for each view and at each time instance. The i-th 2D skeleton in a camera view c at time t is denoted by $s_i^c(t) \in \mathbb{R}^{30}$, which is composed of fifteen 2D anatomical landmarks or *nodes* (3 for the head/torso and 12 for the limbs), and the j-th node of $s_i^c(t)$ is denoted by

[1] Each module has 4 HDDs integrated as RAID-0 to have sufficient write speed without data loss, which ends up with 80 HDDs for 20 modules.



(a) An example view (b) Node proposals (c) Part proposals (d) Part trajectory proposals (e) Skeletal trajectory proposals

Figure 3. Several levels of proposals generated by our method. (a) An example view out of 480 views. (b) Node proposals generated after Non-Maxima Suppression. (c) Part proposals by connecting a pair of node proposals. (d) Part trajectory proposals generated by propagating part proposals. All part trajectory proposals at a time instance is shown in the left rectangle. (e) Skeletal trajectory proposals generated by piecing together part trajectory proposals. Locations of each skeletal proposals at a time instance are shown in the left image. In (b-d), color means part labels: neck (red), head (blue), torso (black), shoulder (green), upper arm (cyan), lower arm (magenta), hip (yellow), upper leg (orange), and lower leg (gray). In (e), color means subject's label.

$s^c_{ij}(t) \in \mathbb{R}^2$. The $s^c_i(t)$ is also associated with its detection score $\alpha^c_i(t) \in \mathbb{R}$ and a scale $\sigma^c_i(t) \in \mathbb{R}$, provided by the 2D pose detector. Given the detected 2D skeletons in view c at time t, we generate a 2D score map $\phi^c_j(z,t)$ for each node j, where $z \in \mathbb{R}^2$ indexes 2D image space. The 2D score maps of node j from all views are then combined into a 3D score map $\mathbf{\Phi}_j(Z,t)$, where $Z \in \mathbb{R}^3$ indexes 3D event space. To produce evidence of the motion of different anatomical landmarks, we compute a set of dense 3D trajectories $\mathbf{F} = \{f_i\}_{i=1}^{N_F}$, or a *3D trajectory stream*, by tracking each 3D particle independently. Each 3D trajectory f_i is initiated at an arbitrary time, and tracked for an arbitrary duration using the method of Joo et al. [22].

Our approach generates several levels of proposals. A set of *node proposals* for a node j is denoted by $\mathbf{X}_j(t)$, and the k-th proposal $\mathbf{X}^k_j(t) \in \mathbb{R}^3$ is a putative 3D position of that anatomical landmark at time t. Similarly, the set of *part proposals* at time t is denoted by $\mathbf{P}_{uv}(t)$, where $(u,v) \in \mathbf{B}$ is the set of all parts composing a skeleton hierarchy. Since our skeleton is a tree structure and has fifteen nodes, $|\mathbf{B}| = 14$. The k-th part proposal, $\mathbf{P}^k_{uv}(t) = (\mathbf{X}^{k_1}_u(t), \mathbf{X}^{k_2}_v(t)) \in \mathbb{R}^6$, is a body part connecting two node proposals, $\mathbf{X}^{k_1}_u(t)$ and $\mathbf{X}^{k_2}_v(t)$, at time t. Our method also estimates trajectory proposals for nodes and parts. We refer to the k-th *node trajectory proposal* as $\mathbf{Y}^k_j = \{\mathbf{Y}^k_j(t)\}_t$, and the k-th *part trajectory proposal* as $\mathbf{Q}^k_{uv} = \{\mathbf{Q}^k_{uv}(t)\}_t$. In our method, a part trajectory proposal is generated by selecting an initial part proposal, and propagating it across time using a set of associated trajectories. Note that a part trajectory proposal is composed of a pair of selected node-trajectories proposals. As a final output, our algorithm produces *skeletal trajectory proposals*; we refer to the k-th proposal as $\mathbf{S}^k = \{\mathbf{Q}^{k_{uv}}_{uv}\}_{uv \in \mathbf{B}}$. The $\mathbf{S}$ can be directly converted to a set of fifteen node-trajectories $\{\mathbf{Y}_j\}_{j=1}^{15}$, by fusing corresponding common nodes of the neighboring part trajectory proposals. Our method associates a set of labelled trajectories $\mathbf{F}^k_{uv}$ out of $\mathbf{F}$ corresponding to each $\mathbf{Q}^k_{uv}$ of a subject. These trajectories determine a series of rigid transformations, $T(t \mid \mathbf{F}^k_{uv}, t_0) \in SE(3)$, between any time t and the initiating time instance t_0 of $\mathbf{Q}^k_{uv}$; the part trajectory proposal $\mathbf{Q}^k_{uv}$ is generated by propagating a part proposal using the $T(t \mid \mathbf{F}^k_{uv}, t_0)$.

5. Skeletal Proposal Generation

We adopt an incremental approach to estimating skeletal motion, fusing appearance and motion cues across the set of views. In this section, we describe how the proposals are generated and built upon from these cues.

5.1. Node Proposals

A single-view 2D pose detector is computed on all 480 views at a time instant, and is used to generate 2D score maps for each node in each image. These 2D score maps from all views are combined in 3D via a spatial voting method, similar to 3D volumetric reconstruction. For 2D pose detection, we use the publicly available pose detector of [42] without additional training. Since we do not assume any prior knowledge about the number of people, each image may have multiple people, and, thus, we keep all the 2D skeletons above a fixed detection threshold in every view. Each 2D skeleton $s^c_i(t)$ in a view c and time t contains a tree like skeletal hierarchy composed of 15 nodes, as shown in Figure 4 (a)[2]. For clarity, we will consider a fixed time instant t, and drop the time variable. From the detected 2D skeletons, we generate a 2D score map for each node j in each view, by convolving a Gaussian kernel on the node locations s^c_{ij}. The score map of a node j in a view c is defined as

$$\phi^c_j(z) = \max_i \alpha^c_i \mathcal{G}(z \mid s^c_{ij}, \sigma^c_i), \tag{1}$$

where $z \in \mathbb{R}^2$ is a 2D location, and $\mathcal{G}$ is a Gaussian kernel centered on s^c_{ij} with covariance σ^c_i, and scaled by the detection score α^c_i. Note that we have a score map for each node and for each view. However, we do not distinguish left-side node with right-side node, because they are dependent on the camera view point. We treat the left-side nodes

[2]We modify the skeleton hierarchy of [42] to have a single torso bone, by taking the center of the two hip nodes as a body center node.

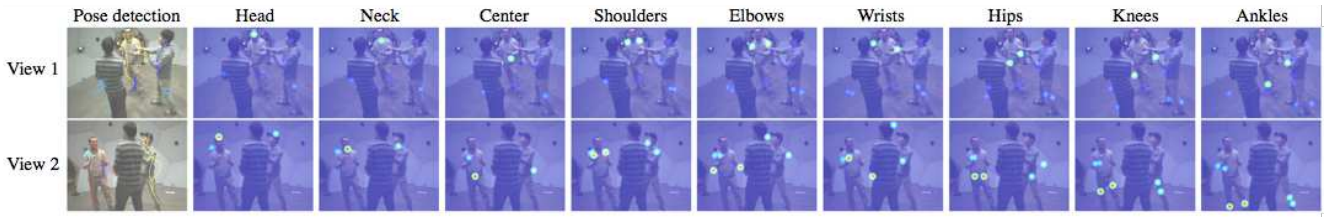


Figure 4. 2D pose detection and score map generation. (Column 1) Example views out of 480 views with proposals by the pose detector (Column 2-10) Score map for each node on each view. Pose detection results are noisy due to occlusions among people.

and corresponding right-side nodes together, producing 9 probability maps (3 for head/torso nodes, and 6 for limbs) in each view. Score maps of example views are shown in Figure 4. Note that pose detection results are noisy, due to the challenging hand gestures and occlusions among people (e.g., see wrists and elbows).

To combine 2D node score maps from multiple views, we generate a 3D score maps for each node using a spatial voting method. We first index the 3D working space into a voxel grid, and compute the *node-likelihood* score of each voxel by projecting the center of the voxel to all views and taking the sum of the 2D scores at the projected locations. The 3D score map Φ_j for a node j at the 3D position Z is defined as

$$\Phi_j(Z) = \sum_c \phi_j^c \left(\frac{\mathbf{M}^c \hat{Z}}{(\mathbf{M}^c \hat{Z})_3} \right), \tag{2}$$

where the $\mathbf{M}^c$ is a projection matrix for view c, and $\hat{\cdot}$ is a homogeneous coordinate representation. The $(\cdot)_3$ means the third column of the vector. Note that 3D score map for each node is computed separately, producing nine 3D score maps at each time. We perform this process at every frame independently.

From the 3D score map for each node at each time instance, we perform Non-Maxima Suppression (NMS), and keep all the candidates above a fixed threshold. The results are shown in the Figure 3(b). Each 3D point, denoted as $\mathbf{X}_j^k$ for the node j, is a putative candidate for the j-th anatomical landmark of a participant, which we refer to as a node proposal.

5.2. Part Proposals

Given the generated node proposals, we infer part proposals by estimating connectivity between each pair of nodes consisting of a body part. The 2D detector as [42] uses appearance information during the inference, and, thus, the result tends to preserve the connectivity information (e.g., left knee is connected to left foot). Although this information is noisy in a single view, our approach fuses them in 3D, by voting for 3D node score maps. More specifically, we define a connectivity score between a pair of node proposals by projecting them on to all views and checking

(a) Epsilon-ball (b) Rigid (ours) (c) Associated trajectories

Figure 5. Trajectory association to body parts. We use the approximate rigidity and spatial proximity to associate unlabelled points trajectories to body parts.

their connectivity in the pose detection for that view. The connectivity score of a part $\mathbf{P}_{uv}$ composed of between two node proposals $(\mathbf{X}_u^{k_1}, \mathbf{X}_v^{k_2})$, where $(u, v) \in \mathbf{B}$, is defined as

$$\mathbf{L}(\mathbf{P}_{uv}) = \sum_c \delta_{uv}^c \left(\frac{\mathbf{M}^c \hat{\mathbf{X}}_u^{k_1}}{(\mathbf{M}^c \hat{\mathbf{X}}_u^{k_1})_3}, \frac{\mathbf{M}^c \hat{\mathbf{X}}_v^{k_2}}{(\mathbf{M}^c \hat{\mathbf{X}}_v^{k_2})_3} \right),$$

where

$$\delta_{uv}^c(\mathbf{x}_u, \mathbf{x}_v) = \begin{cases} 1 & \text{if } \|\mathbf{x}_u - s_{iu}^c\| < \sigma_i^c \text{ and } \|\mathbf{x}_v - s_{ic}^c\| < \sigma_i^c \\ 0 & \text{otherwise.} \end{cases}$$

The $\mathbf{L}$ function counts how many connections exist in the 2D pose results across all views. The line segments in Figure 3(c) represent examples of part proposals. We compute this connectivity score for every pair of nodes, and retain the parts above a fixed threshold, which we call part proposals. Intuitively, each part proposal is a putative candidate of a body part at a time instance.

5.3. Part Trajectory Proposals

Using a part proposal as an initialization, we estimate a part *trajectory* proposal by propagating it backwards and forwards using the 3D trajectory stream $\mathbf{F}$. A part trajectory proposal is a potential candidate of a moving body parts, preserving its (approximate) rigidity. To propagate a part proposal, we associate it with a set of trajectories, and the trajectories constrains the computation of a part specific series of rigid transformations. Using the 3D score maps generated in subsection 5.1 in every time instance, we can score the validity of the part trajectory proposals to discriminate true parts from outliers. Examples are shown in the Figure 3(d).

3D Trajectory Stream Generation. We briefly overview the method of [22] to generate dense 3D trajectories. Given an initial 3D point reconstructed by feature matching and triangulation, it is projected on all views where the point is visible, and optical flow is computed from the projected 2D positions. Next 3D position is reconstructed by back-projecting the tracked 2D flow positions using RANSAC, and this process is iterated. The core idea to fully leverage large number of views is to reason about time-varying camera visibility for each point. The visibility is optimally estimated in a MAP framework combining photometric consistency, motion consistency, and visibility regularization priors. See [22] for more details.

Trajectory Association. To propagate a part proposal generated at a time t_0, we find related trajectories f corresponding to the part, out the $\mathbf{F}$. To select an initial trajectory set, we first use proximity information by selecting all trajectories within an epsilon-ball from the part proposal. Usually, this selection contains trajectories that originated from other body parts especially when they are close each other, as shown in Figure 5(a). Although solving this problem is challenging for a point cloud at a single time instance, it can be distinguishable in our case by analyzing the trajectories for long duration of time. To score the likelihood that two trajectories $\mathbf{f}_1$ and $\mathbf{f}_2$ originated from the same rigid part, we define a distance

$$d(\mathbf{f}_1, \mathbf{f}_2) = \max_t \|\mathbf{f}_1(t) - \mathbf{f}_2(t)\| - \min_t \|\mathbf{f}_1(t) - \mathbf{f}_2(t)\|. \tag{3}$$

When trajectories arise from the same rigid body part, this distance is close to zero, whereas trajectories from different parts have a large error once they move with a distinct motion. To find the correct inlier set given the initial trajectories by fixed radius thresholding, we perform RANSAC based on this distance. In each iteration, we select a reference trajectory and find corresponding inliers in a distance lower than a threshold. An example result is shown in Figure 5.

From a set of trajectories $\mathbf{F}_{uv}$ associated with a part proposal $\mathbf{P}_{uv}$, we can estimate a series of rigid transformations $T(t \mid \mathbf{F}_{uv}, t_0)$ from t_0 to any time t where a rigid transformation can be estimated from trajectories. Our approach then generates a part trajectory proposal as,

$$\begin{aligned}\mathbf{Q}_{uv}(t) &= T(t \mid \mathbf{F}_{uv}, t_0) \cdot \mathbf{P}_{uv} \\ &= (T(t \mid \mathbf{F}_{uv}, t_0) \cdot \mathbf{X}_u, T(t \mid \mathbf{F}_{uv}, t_0) \cdot \mathbf{X}_v) \\ &= (\mathbf{Y}_u(t), \mathbf{Y}_v(t)),\end{aligned}$$

where $\mathbf{P}_{uv} = (\mathbf{X}_u, \mathbf{X}_v)$. Note that the $\mathbf{Y}_u$ and the $\mathbf{Y}_v$ move rigidly since they are propagated by same rigid transformations.

Part Trajectory Scoring. We compute the score of each part trajectory proposal using the 3D score maps generated in subsection 5.1. The 3D score map at a time instance measures the *node likelihood* at that 3D location. Thus, we can compute the score of the part trajectory proposal by aggregating the 3D scores of all locations where the part traverses. That is,

$$\Theta_{uv}(\mathbf{Q}_{uv}) = \sum_t \big(\Phi_u(\mathbf{Y}_u(t)) + \Phi_v(\mathbf{Y}_v(t))\big). \tag{4}$$

This measurement means that we favor part trajectory proposals that go though the region of high detection scores, with rigidity constraint between two end nodes (note that the two end nodes are rigid by construction).

5.4. Skeletal Trajectory Proposals

Each part trajectory proposal is a candidate for moving body parts. Skeletal trajectory proposals are obtained by selecting the best combination of part trajectory proposals. We use Dynamic Programming (DP) over part trajectory proposals. We can consider our model as an undirected graph $G = (V, E)$, where each vertex is a part trajectory proposal $\mathbf{Q}^k_{uv}$, and the graph edges are defined by the parts in a child-parent relationship. The edge score is defined using the distance of Equation 3. For example, an instance of upper arm part trajectory has an edge with a lower arm part trajectory, and the edge score between them is determined by the elbow node trajectories of both parts, which should ideally be coincident. Using DP, we maximize the following objective:

$$\Theta_{\mathbf{S}}(\mathbf{S}^k) = \sum_{(u,v)\in\mathbf{B}} \Theta_{uv}(\mathbf{Q}^k_{uv}) - \lambda \sum_{(u,v),(v,w)\in\mathbf{B}} \Psi(\mathbf{Q}^k_{uv}, \mathbf{Q}^k_{vw}),$$

where

$$\Psi(\mathbf{Q}_{uv}, \mathbf{Q}_{vw}) = d(Y^v_-, Y^v_+).$$

$\Psi(\mathbf{Q}_{uv}, \mathbf{Q}_{vw})$ is a pairwise term of two part trajectory proposals, where Y^v_- is a node trajectory from $\mathbf{Q}_{uv}$ and Y^v_+ is a node trajectory from $\mathbf{Q}_{vw}$ (e.g., $\mathbf{Q}_{uv}$ is an upper arm and $\mathbf{Q}_{vw}$ is a lower arm and both Y^v_- and Y^v_+ are elbow's trajectories from the two different parts respectively). We subtract two terms, because the first term is a score and the second term is a distance. The λ is a weight factor balancing between them. As mentioned, a skeletal trajectory proposal can be represented by 15 node trajectories $\{\mathbf{Y}_j\}_{j=1}^{15}$. We perform dynamic programming on our part trajectory pools, and retain all skeletal trajectories after NMS and thresholding.

6. Trajectory Optimization and Reassociation

From the method described above, initial skeletal trajectory proposals and initial part association for trajectories are

generated. Using these as an initialization, we refine the estimates by optimizing the skeletal trajectories and subsequently re-associating trajectories. Each skeletal trajectory is optimized as:

$$\underset{\{\mathbf{Y}_i\}}{\operatorname{argmin}} \sum_{i=1}^{N_Y} \Phi_i(\mathbf{Y}_i),$$

where

$$\mathbf{Y}_i(t) = T(t \mid \mathbf{F}_i, t_0) \cdot \mathbf{Y}_i(t_0).$$

The $\mathbf{Y}_i(t_0)$ is the initial 3D node location of $\mathbf{Y}_i$ and the $T(t \mid \mathbf{F}_i, t_0)$ is transformations determined by the associated trajectories $\mathbf{F}_i$. Assuming $\mathbf{F}_i$ is fixed, we can optimize this objective by varying $\{\mathbf{Y}_i(t_0)\}$. For trajectory re-association, we use a distance measurement similar to Equation 3 between a trajectory and a part trajectory proposal as

$$d_{\mathbf{Q}}(\mathbf{f}_1, \mathbf{Q}_{uv}) = \max_t \|\mathbf{f}_1(t) - \mathbf{Q}_{uv}(t)\| - \min_t \|\mathbf{f}_1(t) - \mathbf{Q}_{uv}(t)\|,$$

where the $\|\mathbf{f}_1(t) - \mathbf{Q}_{uv}(t)\|$ represents orthogonal distance at time t from a 3D location to a body part (line segment). Our method iteratively performs skeleton location optimization and trajectory re-association, and, as output, produces refined skeletal reconstruction with labelled trajectories.

7. Results

7.1. Dataset

We capture people engaged in various social interactions using our massive camera system with 480 views. To evoke natural interactions, we involved participants in various games: *Ultimatum* (with 3 subjects), *Prisoner's dilemma* (with 8 subjects), *Mafia* (with 8 subjects), *Haggling* (with 3 subjects), and *007-bang game* (with 5 subjects)[3]. The first three games are used in experimental economics and psychology to study conflict and cooperation, and we select additional two games where rich natural interactions can be induced. The number of participants in each session varies from three to eight. From captured data, we selected 5 vignettes containing interesting non-verbal interactions among people. We will publicly share these dataset with all 480 synchronized camera feeds, calibration, 3D trajectory stream, 3D pose reconstruction, and articulated non-rigidity representation result.

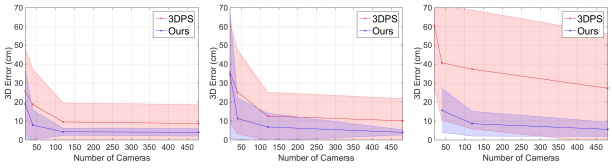
7.2. Quantitative Evaluation

We compare our method with two different baselines: multiple Kinects and 3D pictorial structure method similar to [8, 5]. We use 5 Kinect IIs calibrated in the same coordinate with our cameras. Since Kinects do not accept external time signal for sync, we manually align them to our VGA cameras up to frame level. The 3D pose estimation of Kinects are performed individually. For the 3D pictorial structure (3DPS) method, we implement the spatial message passing algorithm based on our skeletal hierarchy in the reconstructed 3D volume space of our method. Since the 3DPS method does not produce temporally coherent subject identity, we retain all the proposals in every frame using a sufficiently low threshold, and compute 3D error by finding the proposal with lowest 3D error from ground truth data.

We select the haggling and 007-bang sequences, and manually generate ground truth data by annotating subject's node locations in multiview. Table 1 and 2 show average errors from all methods. As shown in the results, all individual Kinects shows failures on at least one subject, because people are facing each other, and, thus, they frequently occlude each other's frontal view, which severely affects Kinect's performance. To see the limit of multiple Kinect system, we also compute the lowest error at each time among all Kinects, assuming that an Oracle selects the best view for each node at each time, which is still outperformed by ours. The 3D pictorial structure also shows frequent failures, because it only relies on appearance cue, while our method fuses motion cue together. The appearance cue becomes often weak because: (1) there exist severe occlusions and interference among people; (2) many views are observing only parts of human body (e.g., upper

Table 1. Average 3D errors (cm) of Haggling Sequence (subject 1: short hair, subject 2: grey hoodie, subject 3: striped sweater).

Subject	K1	K2	K3	K4	K5	OracleKinect	3DPS	Ours
1	14.24	10.35	35.51	15.84	12.04	5.08	8.67	**3.94**
2	12.40	9.10	60.05	66.39	94.56	5.84	10.11	**5.18**
3	81.64	78.40	55.58	13.62	8.65	5.74	27.28	**5.52**



(a) Subject 1 (b) Subject 2 (c) Subject 3

Figure 7. Average 3D error on varying number of cameras in the Haggling sequence. A significant number of cameras (more than 100 in this case) are necessary to achieve accurate motion capture. The result of subject 3 in (c) shows higher errors as the pose detection severely failed due to occlusions. With 20 cameras, our algorithm failed to find subject 3, and 3D Pictorial Structure (3DPS) also shows large error.

Table 2. Average 3D errors (cm) in 007-bang sequence. We selected the challenging subject on the center (grey hoodie).

K1	K2	K3	K4	K5	OracleKinect	3DPS	Ours
102.48	13.26	11.99	89.67	206.28	7.80	13.11	**6.25**

[3]Refer the supplementary material for the descriptions of the games and our capture procedures

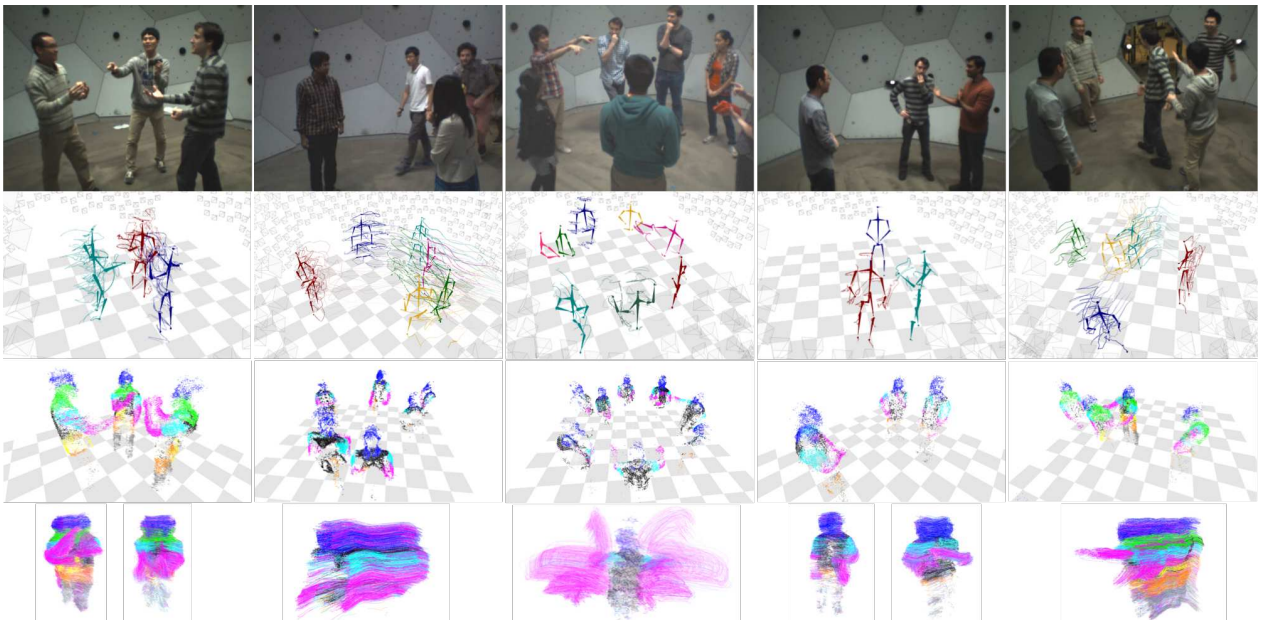


Figure 6. We perform our method to capture social interactions of multiple people on 5 vignettes: Haggling (column 1), Prisoner's dilemma (column 2), Mafia (column 3), Ultimatum (column 4), and 007-bang game (column 5). (Row 1) Example views; (Row 2) Skeletal structure reconstruction with visualized node trajectories; (Row 3 and Row 4) Labelled 3D trajectories representing articulated non-rigid body parts of each subject, where colors represent same parts as in Figure 3.

body only); (3) camera views may not be consistent with the dataset used to train the pose detector of [42]. Note that directly retraining the detector for each view of our system would require a large annotated training set for each view.

Accuracy on Varying Camera Number. We apply our method and the 3DPS algorithm for the Haggling sequence with varying number of cameras (20, 40, 120, and 480). As shown in the Figure 7, if a small number of cameras is used, both our method and 3DPS have higher 3D error. The error saturates around 100 cameras, which depends on the complexity of the scene. The difference in performance between ours and 3DPS is caused by the tolerance on pose detection inaccuracy. By using dense 3D trajectories, our method can better utilize temporal relation, and outperforms methods relying on 2D pose detection only.

7.3. Qualitative evaluation

We apply our method to capture social motion of multiple interacting people in all of our dataset, and the results are shown in Figure 6. Our approach automatically reconstructs each subject's moving skeletal structure and its non-rigid part models by associating trajectories. The test scenes contain naturally emerged social motion of people, including subtle gestures, gaze direction changes, and topological changes. Note that the results are generated without knowing the number of subjects or individual specific information such as body shape and bone-length. Our results demonstrate the robustness in capturing rich social signals in various challenging scenarios.

8. Discussion

We present a system to capture the social interaction of multiple people. Our system is composed of a massively multiview camera system in a modularized design, and a novel algorithm fusing "weak" detection and tracking cues from multiple views for robust human skeletal pose estimation, associated with detailed labelled trajectories for each body parts. Our method is well suited for sociological analyses since both our hardware and software systems are designed to be unobtrusively robust to occlusions that emerge during socials interactions. The fact that our system does not require any time-consuming model generation is also a crucial advantage to be used as a tool for behavioral analysis. Additional, the labelled trajectories associated to each body part can provide further detail of motions. There are two major failure cases of our method. It generates failure if either detection or tracking cues completely fails. Examples are: (1) detection consistently produces a strong false positives; (2) no trajectory is reconstructed due to the lack of texture such as dark pants in our data set.

Acknowledgements

This research is supported by the National Science Foundation under Grants No. 1353120 and 1029679, and in part using an ONR grant 11628301.

References

[1] A. Alahi, V. Ramanathan, and L. Fei-Fei. Socially-aware large-scale crowd forecasting. *CVPR*, 2014.

[2] I. Arev, H. S. Park, Y. Sheikh, J. Hodgins, and A. Shamir. Automatic editing of footage from multiple social cameras. *SIGGRAPH*, 2014.

[3] H. Aviezer, Y. Trope, and A. Todorov. Body cues, not facial expressions, discriminate between intense positive and negative emotions. *Science*, 2012.

[4] A. Baak, M. Müller, G. Bharaj, H.-P. Seidel, and C. Theobalt. A Data-Driven Approach for Real-Time Full Body Pose Reconstruction from a Depth Camera. *ICCV*, 2011.

[5] V. Belagiannis, S. Amin, M. Andriluka, B. Schiele, N. Navab, and S. Ilic. 3D pictorial structures for multiple human pose estimation. *CVPR*, 2014.

[6] R. Birdwhistell. Kinesics and context: Essays on body motion communication. *University of Pennsylvania Press*, 1970.

[7] T. B. Brazelton, B. Koslowski, and M. Main. The origins of reciprocity: The early mother-infant interaction. *Oxford, England: Wiley-Interscience*, 1974.

[8] M. Burenius, J. Sullivan, and S. Carlsson. 3D pictorial structures for multiple view articulated pose estimation. *CVPR*, 2013.

[9] I. Chakraborty, H. Cheng, and O. Javed. 3d visual proxemics: Recognizing human interactions in 3d from a single image. *CVPR*, 2013.

[10] W. S. Condon and L. W. Sander. Synchrony demonstrated between movements of the neonate and adult speech. *Child development*, 1974.

[11] C. Darwin. The expression of the emotions in man and animals. *John Murray*, 1872.

[12] E. de Aguiar, C. Stoll, C. Theobalt, N. Ahmed, H.-P. Seidel, and S. Thrun. Performance capture from sparse multi-view video. *SIGGRAPH*, 2008.

[13] P. Ekman and W. Friesen. Facial action coding system. *Consulting Psychologists Press*, 1977.

[14] A. Elhayek, C. Stoll, N. Hasler, K. I. Kim, H.-P. Seidel, and C. Theobalt. Spatio-temporal motion tracking with unsynchronized cameras. *CVPR*, 2012.

[15] A. Fathi, J. K. Hodgins, and J. M. Rehg. Social interactions: A first-person perspective. *CVPR*, 2012.

[16] Y. Furukawa and J. Ponce. Dense 3d motion capture from synchronized video streams. *CVPR*, 2008.

[17] J. Gall, C. Stoll, E. de Aguiar, C. Theobalt, B. Rosenhahn, and H.-P. Seidel. Motion capture using joint skeleton tracking and surface estimation. *CVPR*, 2009.

[18] M. Gross, S. Würmlin, M. Naef, E. Lamboray, C. Spagno, A. Kunz, E. Koller-Meier, T. Svoboda, L. Van Gool, S. Lang, K. Strehlke, A. V. Moere, and O. Staadt. Blue-c: A spatially immersive display and 3d video portal for telepresence. *SIGGRAPH*, 2003.

[19] E. T. Hall. Proxemics: The study of man's spatial relations. *International Universities Press*, 1962.

[20] M. Hofmann and D. Gavrila. Multi-view 3d human pose estimation in complex environment. *IJCV*, 2012.

[21] C. E. Izard. The face of emotions. *New York: Appleton-Century-Crofts*, 1971.

[22] H. Joo, H. S. Park, and Y. Sheikh. Map visibility estimation for large-scale dynamic 3d reconstruction. *CVPR*, 2014.

[23] T. Kanade, P. Rander, and P. Narayanan. Virtualized reality: Constructing virtual worlds from real scenes. *IEEE Multimedia*, 1997.

[24] Y. Liu, J. Gall, C. Stoll, Q. Dai, H.-P. Seidel, and C. Theobalt. Markerless motion capture of multiple characters using multiview image segmentation. *TPAMI*, 2013.

[25] T. Matsuyama and T. Takai. Generation, visualization, and editing of 3d video. *3DPVT*, 2002.

[26] W. Matusik, C. Buehler, R. Raskar, S. J. Gortler, and L. McMillan. Image-based visual hulls. *SIGGRAPH*, 2000.

[27] R. McDonnell, S. Jrg, J. McHugh, F. Newell, and C. O'Sullivan. Evaluating the emotional content of human motions on real and virtual characters. *Symposium on Applied Perception in Graphics and Visualization*, 2008.

[28] H. K. M. Meeren, C. C. R. J. van Heijnsbergen, and B. de Gelder. Rapid perceptual integration of facial expression and emotional body language. *National Academy of Sciences of the United States of America*, 2005.

[29] E. Muybridge. The human figure in motion. *Courier Corporation*, 1887.

[30] H. S. Park, E. Jain, and Y. Sheikh. 3d gaze concurrences from head-mounted cameras. *NIPS*, 2012.

[31] H. S. Park, E. Jain, and Y. Sheikh. Predicting primary gaze behavior using social saliency fields. *ICCV*, 2013.

[32] B. Petit, J.-D. Lesage, E. Boyer, and B. Raffin. Virtualization Gate. *SIGGRAPH Emerging Technologies*, 2009.

[33] J. S. Philpott. The relative contribution to meaning of verbal and nonverbal channels of communication. *Unpublished master's thesis, University of Nebraska*, 1983.

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

[35] E. Sapir. The unconscious patterning of behavior in society. *Selected Writings of Edward Sapir in Language, Culture, and Personality*, 1949.

[36] J. Shotton, A. Fitzgibboan, M. Cook, and T. Sharp. Real-time human pose recognition in parts from single depth images. *CVPR*, 2011.

[37] L. Sigal, S. Bhatia, S. Roth, M. J. Black, and M. Isard. Tracking loose-limbed people. *CVPR*, 2004.

[38] C. Stoll, N. Hasler, J. Gall, H.-P. Seidel, and C. Theobalt. Fast articulated motion tracking using a sums of gaussians body model. *ICCV*, 2011.

[39] D. Vlasic, I. Baran, W. Matusik, and J. Popović. Articulated mesh animation from multi-view silhouettes. *SIGGRAPH*, 2008.

[40] R. Williams. The geometrical foundation of natural structure: A source book of design. *Dover Publications*, 1979.

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

[42] Y. Yang and D. Ramanan. Articulated Human Detection with Flexible Mixtures-of-Parts. *TPAMI*, 2012.

[43] A. L. Yarbus. Eye movements and vision. *Plenum press*, 1967.