The Tower Game Dataset: A Multimodal Dataset for Analyzing Social Interaction Predicates

Behjat Siddiquie, Mohamed R. Amer, Amir Tamrakar, David A. Salter, Ajay Divakaran
SRI International

Brian Lande
Polis LLC

Darius Mehri
UC Berkeley

***Abstract*—We introduce the *Tower Game Dataset* for computational modeling of social interaction predicates. Existing research in affective computing has focused primarily on recognizing the emotional and mental state of a human based on external behaviors. Recent research in the social science community argues that engaged and sustained social interactions require the participants to jointly coordinate their verbal and non-verbal behaviors. With this as our guiding principle, we collected the *Tower Game Dataset* consisting of multimodal recordings of two players participating in a tower building game, in the process communicating and collaborating with each other. The format of the game was specifically chosen as it elicits spontaneous communication from the participants through social interaction predicates such as such as *joint attention* and *entrainment*. The dataset will be made public and we believe that it will foster new research in the area of computational social interaction modeling.**

I. Introduction

Social interactions are an important part of human life. Humans interact and communicate with each other for a variety of reasons. These interactions, in turn, have a significant social and psychological impact on the people participating in them. The study of social interactions and their associated social and psychological implications have long been important research area in the social sciences community [1]–[4].

A better understanding of social interactions can benefit a number of applications. For example, consider aid-workers deployed in a foreign country. During the course of their deployment, these workers often have to interact with people with whom they share little in common in terms of language, customs and culture. Reducing friction as well as increasing engagement between the workers and the populations can have an important bearing on the success of their mission. Therefore, the ability to impart such professionals, with a general cross-cultural competency which would enable them to smoothly interact with the foreign populations they encounter would be extremely useful. With such an application in mind, we focus on identifying and automatically detecting predicates that facilitate social interactions irrespective of the cultural context. Since our interests lie in aspects of social interactions that reduce conflict and build trust, we focus on social predicates that support rapport: joint attention, temporal synchrony, mimicry, and coordination.

Given our focus on computational modeling of social interactions, we want to leverage research in social psychology, computer vision, signal processing, and machine learning to better understand human social interactions. Starting from a sound socio psychological theory, we collected a new *Tower Game Dataset*, consisting of audio-visual recordings of dyadic interactions. The dataset has been annotated with Essential Social Interaction Predicates (ESIPs).

Our orientation to social sensing departs significantly from existing methods [1], [5] that focus on inferring internal or hidden mental states. Instead, inspired by a growing body of research [6]–[8], we focus on the process of social interaction. This research argues that social interaction is more than the meeting of two minds, with an additional emphasis on the cognitive, perceptual and motor explanations of the joint and coordinated actions that occur as part of these interactions [9]. Our approach is guided by two key insights. The first is that apart from inferring the mental state of the other, social interactions require individuals to attend each other's movements, utterances and context to coordinate actions jointly with each other [10]. The second insight is that social interactions involve reciprocal acts, joint behaviors along with nested events (e.g. speech, gaze, gestures) at various timescales and therefore demand adaptive and cooperative behaviors of their participants [11].

Using the work of [12], which emphasizes the interactive and cooperative aspects of the social interactions, we focus on detecting rhythmic coupling (entrainment), mimicry (behavioral matching), movement simultaneity, kinematic turn taking patterns, and other measurable features of engaged social interaction. We established that behaviors such as *joint attention* and *entrainment* are the essential predicates of social interaction (ESIPs). We focus on developing computational models of social interaction, that utilize multimodal sensing and temporal deep learning models to detect and recognize these ESIPs as well as discover their actionable constituents.

With the aforementioned motivation, our aim was to collect recordings of dyadic interactions between individuals

containing behaviors such as *joint-attention* and *entrainment*. Given these requirements, we chose to have pairs of players participate in a tower building game. These tower games consist of both of the players cooperating, without verbally communicating, on building a tower using blocks of various shapes and colors. We record the interactions between the players playing these games using multiple audio-visual sensors enabling us to detect a range of fine-grained participant behaviors and actions, such as eye gaze and head and body motion. Furthermore, the dataset has been annotated with essential social interaction predicates (ESIPs). We expect this dataset to provide a new avenue for research in computational social interaction modeling.

Paper organization: The remainder of the paper is organized as follows. In sec. II we discuss prior work. In sec. III we describe the social psychological theory and methodology that motivates our approach. In sec. IV we specify our data collection procedure followed by a description of our dataset. Finally, we conclude in sec. V.

II. Related Work

Affective Computing refers to the study and development of systems that can automatically detect human affect [5], [13], has long been an active research area due to its utility in a variety of applications that require realistic Human Computer Interaction such as online tutoring [14] or health screenings [15]. The goal here is to detect the overall mental or emotional state of the person based on external cues. This is typically done based on speech [16], facial expressions [17], gesture and posture [18] and multimodal cues [19]–[21]. There has also been work on modeling activities and interactions involving multiple people [22]–[24]. However, most of this work deals with short duration task-oriented activities [22], [23] with a focus on their physical aspects. There has been a recent interest in modeling interactions with a focus on the rich and complex social behaviors that they elicit along with their affective impact on the participants [24].

The study of social interactions and their associated sociological and psychological implications has received a lot of attention from social science researchers [7], [9], [25]. Early research focused on the "Theory of Mind" according to which individuals ascribe mental states to themselves and others [1], a line of thinking that largely inspired much of the initial work on affective computing. However, more recent work has shown that apart from inferring each other's mental states, an important challenge for participants of a social interaction is to pragmatically sustain sequences of action where the action is tightly coupled to another via multiple channels of observable information (e.g. visible kinematic information, audible speech). In other words, social interactions require dynamically coupled interpersonal motor coordination from their participants [10]. Moreover, detecting coupled behaviors such as kinematic turn taking or simultaneity in movements can help in recognizing engaged social interactions [12].

To aid affective computing research, several datasets have been collected. These datasets have been collected primarily for developing computational models for detecting affect. Initially, these datasets focused on a single modality such as facial expressions [26], [27], paralinguistics [28], [29] and body pose and gestures [30], [31]. An important issue with some of these early datasets, for example [26] in the case of facial expressions, was that they contained acted or posed emotions and hence were not very reliable. To address this issue, researchers have collected datasets that contain spontaneous behavior and emotions [32], [33]. Furthermore, since humans express affect through multiple mediums simultaneously, several multimodal datasets have been collected [34]–[36] to facilitate development of multimodal models of affect.

In the field of activity recognition, there are several datasets and benchmarks – e.g., the Weizmann, Trecvid, PETS04, CAVIAR, IXMAS, Hollywood datasets, Olympic Sports and UCF-100 – however, these datasets contain relatively simple and repetitive actions involving a single person [37]. On the other hand, group activity recognition datasets such as UCLA Courtyard, UT-Interactions, Collective Activity datasets, and Volleyball dataset, lack rich social interactions and dynamics.

Other relevant datasets that focus on interactions include the Multimodal Dyadic Behavior (MMDB) dataset [24], which focuses on analyzing dyadic social interactions between adults and children in a developmental context. This dataset was collected in a semi-structured format, where children interact with an adult examiner in a series of pre-planned games. However, due to its narrow focus on analysis of social behaviors to diagnose developmental disorders in children, we believe it is not general enough. Another relevant dataset is the Mimicry database [38] which focuses on studying social interactions between humans with the aim of analyzing mimicry in human-human interactions.

There are a number of issues with the aforementioned affective computing, activity recognition and interaction datasets, including: (a) unnatural, acted activities and emotions in constrained scenes; (b) limited spatial and temporal coverage; (c) poor diversity of activity classes; (d) Lack of rich social interactions; (e) Narrow focus on a single behavior (e.g. mimicry); and (f) Unstructured or uncontrolled collection setup. Hence, we propose our new Tower Game Dataset to address the above issues.

III. Approach

In this section we describe the socio-psychological reasoning that motivates our approach towards analyzing social interaction processes, followed by a brief high-level overview of our approach.

Our orientation to social sensing departs significantly from existing methods [1], [5] that focus on inferring internal or hidden mental states (e.g. thoughts and feelings) following psychological accounts of "Theory of Mind" [1], [2], [39]–[41]. Instead, we follow a growing body of research in child development psychology [42], [43], primateology [44], [45], ethnomethodology and conversation analysis [46]–[49], sociolinguistics [7], and cognitive science that focuses on

the process of social interaction [3], [4], [8], [10], [50]. An emerging theme is that social interaction involves more than the meeting of two minds through complex "mind reading". While, "mind reading" or mind perception does occur, it is not the only type of social cognition or social perception. Researchers have begun to additionally emphasize cognitive, perceptual, and motor (kinematics and interpersonal motor coordination) explanations of joint and coordinated action that are independent of "mind reading" [9], [51].

Two themes have emerged. The first is that inferring mental states of one another is not the only problem in social interactions. The challenge for participants of a social interaction is to pragmatically sustain sequences of action where the action is tightly coupled to one another via multiple channels of observable information such as visible kinematic information, audible speech, etc. not all of which is purposefully expressive. In other words, individuals must attend to the form and timing of movements and also recognize which bodily movements are not relevant. They must also take into account each others utterances, and context to compute aligned behaviors, make predictions of one another's action [10], and coordinate actions jointly and over time with another. The task of interpersonal motor coordination requires different cognitive and perceptual resources and aims at a different goal compared to inferring mental states. The second theme is that the proper unit of study is the interaction, not the individual [11]. Social interaction has its own unique dynamics and constraints. Interaction involves verbal and non-verbal behaviors, places timing demands on participants (i.e. sequential acts), involves reciprocal acts, involves joint perceptions and actions, nests events at different timescales (e.g. speech, eye gaze, gestures, and behaviors all occur at different temporal time scales) and demands that behavioral patterns be non-stationary and adaptive [12].

The sensing and measurement methods we describe in this paper used the work of [12] as their starting point. They define social interaction as : *"Two or more autonomous agents co-regulating their coupling with the effect that their autonomy is not destroyed and their relational dynamics acquire an autonomy of their own"*. Similarly, they define coordination as: *"A non-accidental correlation in the activity of two or more systems that are coupled at present or were coupled in the past, or are or were coupled to another system in common, over and above what is expected from their normal behaviour in the absence of such couplings. A typical example of coordination between two people is synchronization of speech and bodily movements during a conversation. A situation where two people not directly influencing each other turn their attention to the same object at the same time because a strange sound is coming from it, is an example of coordination because of an external event"*.

Consequently we focus on detecting rhythmic coupling (also known as entrainment and attunement), mimicry (behavioral matching), movement simultaneity, kinematic turn taking patterns, and other measurable features of engaged social interaction. By engaged we mean the qualitative aspect of interaction where participants become absorbed or taken up in the interaction and the interaction is seen and felt to have "taken over." Other studies have used video based measurement tools to automatically detect rhythmic features of interaction [52]–[54] and richly elaborated coding frameworks have been developed to measure temporal synchrony, affect attunement [55], [56], behavioral matching, joint engagement [57], and joint gaze patterns [58]–[60]. Given this background, we approach the problem of computational modeling of social interactions as follows:

- Design and carry out experiments that:
  - Evoke the aforementioned essential predicates in dyadic interactions.
  - Provide a sensing framework that provides a rich multimodal capture of the interaction.
- Manually annotate videos with labels for basic predicates.
- Train statistical classifiers on multimodal data for detecting these predicates.
- Analyze classifiers to uncover constituent behaviors of coordination.

In this paper, we focus on steps 1-2, leaving steps 3 and 4 for future work. The final step of analyzing classifiers would enable us to uncover actionable constituents of key predicates of social interaction, which serves the original social interaction training application that motivated our work.

IV. Dataset

In this section, we describe the collection, processing and annotation of our *Tower Game Dataset*, which contains recordings of two players building a tower and in the process engaging in a variety of interactive behaviors.

A. Tower Game

The tower game is a simple game of tower building often used in social psychology to elicit different kinds of interactive behaviors from the participants [61]. It is typically played between two people working with a small fixed number of simple toy blocks that can be stacked to form various kinds of towers. There are many different variants of the game. We settled on two variants designed to elicit maximum communication between the players, namely, (i) the *architect-builder* variant and (ii) the *distinct-objective* variant. Furthermore, in order to maximize the amount of non-verbal communication, we prohibited the participants from verbally communicating with each other.

The *architect-builder* variant involves one participant playing the role of the architect, who decides the kind of tower to build and how to build it. The second participant is the builder, who has control of all the building blocks and is the only one actually manipulating the blocks. The goal here is for the architect to communicate to the builder how to build the tower so that builder can build the desired tower.

The *distinct-objective* variant is slightly more complicated and is designed to elicit more interaction between the players. Each player is given half of the building blocks required to build the tower. Each player is also given a particular rule,

restricting the configuration of the tower being built, that they are required to enforce. An example rule could be that no two blocks of the same color may be placed such that they are touching each other. We limit the number of rules to 6. To make the play interesting, each player only knows their own rule and is not aware of rule given to the other player. The rules are selected at random from a small rule book. While certain combinations of rules may result in some conflict between the objectives of the two players, this is typically not the case. However, since each player needs to adhere to their rule, it means that they will need to correct an action taken by the other if it conflicts with their rule. In the process, each player also tries to figure out the rule assigned to the other player so that the process of building the tower is more efficient. Also, when the subjects played multiple sessions of this game, the blocks used were changed.

We choose these tower games as they force the players to engage and communicate with each other in order to achieve the objectives of the game, thereby evoking behaviors such as *joint-attention* and *entrainment*. The game, due to its simplicity, allows for total control over the variables of an interaction. Due to the small number of blocks involved, the number of potential moves (actions) is limited. Also since the game involves interacting with physical objects, *joint-attention* is mediated through concrete objects. Furthermore, only two players are involved, ensuring that we can stay in the realm of dyadic interactions.

B. Data Capture Setup

Our sensors include a pair of Kinect cameras that record color videos, depth video and track skeletons of the players and a pair of GoPro cameras mounted on the chest of each player (Fig. 1). External lapel microphones were attached to the GoPro cameras. However, the audio captured from them was used only for data synchronization purposes. Since the players were not allowed to verbally communicate with each other, very little speech (or paralinguistic) data exists.

In order to ensure optimal data capture from the Kinect cameras (i.e. minimal occlusions and optimal skeleton tracking), they were mounted on tripods facing one another, slightly to the right and behind each of the participants and slightly elevated, ensuring that each camera got an unobstructed view of the other participant. The overhead layout is shown in Fig. 1. These videos are of VGA resolutions (640x480) and were captured at 30Hz. The GoPro cameras were set to capture at full HD (1920x1080) resolution. They were placed on the harnesses rotated 90 degrees so as to capture the face of the other player as well as the blocks on the table (Fig. 2).

In each session, the subjects play the game by standing at either end of a small rectangular table as shown in Fig. 2. The person supervising the data collection enters player information and other meta-data about the game session into a form and then starts recording. He/she then instructs the players to begin. They first manually activate the GoPro cameras to start recording and then clap their hands before starting their sessions. These claps were used to automatically synchronize the GoPro videos with the Kinect videos.

C. Data Pre-Processing

Our custom data collection software recorded the Kinect videos and skeletons from both the cameras in a synchronized manner in real-time and later imported the corresponding recordings from the GoPro cameras after the session was finished. We used a number of computer vision algorithms to reconstruct the scene and to extract a number of features that would enable understanding the social interactions. We now describe each of these steps in detail.

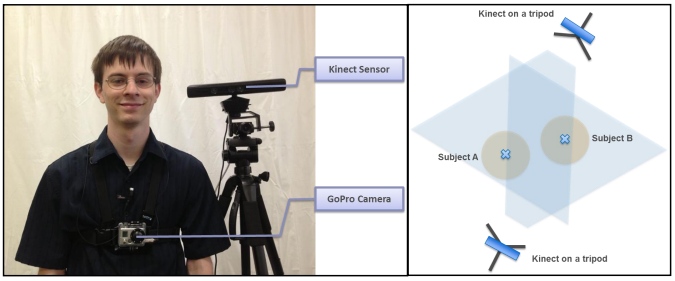

Fig. 1. Data capture setup: The image on the left shows our capture setup which includes a GoPro camera mounted on each participant's chest and a kinect mounted on a tripod. The diagram on the right provides an overhead view of our capture setup involving the two participants.

Fig. 2. Merged Video: The image outlined in solid red shows the image captured from the GoPro camera on player A (green shirt), while the image outlined in dashed red shows the image captured from the kinect behind player A . Similarly the image outlined in solid green is the image captured from the GoPro on player B (green shirt) and the image outlined in dashed green is the image captured from the kinect behind player B.

Synchronization: We synchronized the GoPro videos with the Kinect videos based on the player's claps at the start of the game. In order to make that work, audio streams from the Kinect sensors were also synchronized with the Kinect videos. A custom clap detection algorithm was used to precisely locate the onset of claps in the audio streams. This algorithm was based on detecting a sudden large increase in the energy of the audio signal and the distribution of the energy over the full range of frequencies in the spectrum. It worked very well on most clips. However, it did produce occasional false alarms, in which case, the synchronization points were manually selected

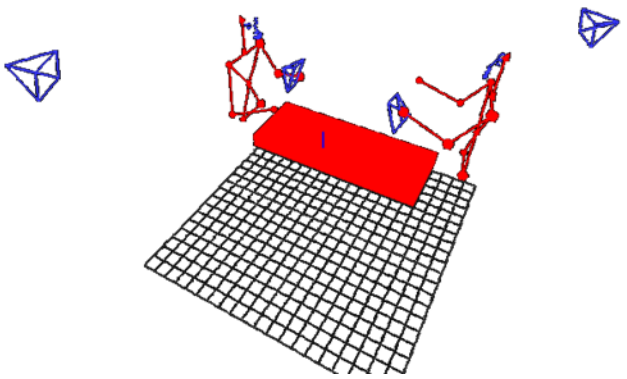

Fig. 3. Data Visualization: A view of our collected data projected in a unified coordinate framework as generated by our visualization software.

from a small set of likely candidates. Note that the audio frames used for computing the power spectra were about 5ms in length. Hence, the clap localizations thus produced were able to provide a frame-level accurate synchronization of the videos. All the videos were then clipped to include just the duration of the game. Additionally, a frame synchronized merged video, as shown in Fig. 2, was also generated for each game session, primarily to help with the annotation.

Camera Calibration: We calibrated the GoPro cameras using the Camera Calibration Toolbox for Matlab [62], obtaining the intrinsic camera matrices as well as their distortion coefficients. Using these we were able to generate distortion corrected versions of the GoPro videos. The intrinsic camera matrices for the Kinect cameras were obtained directly from the Kinect sensors.

Determining Kinect Camera Pose: The extrinsic pose (Rotation and Translation) of the static Kinect cameras were obtained in the world coordinate system using the 3D point clouds generated from the Kinect depth videos. The origin of the world coordinate system was set to a point on the floor beneath the center of the table. Furthermore, the world coordinate system was defined to be aligned with the walls of the room, simplifying the pose extraction process. From each point cloud, the wall and floor planes were estimated using robust plane fitting algorithms. The known dimensions of the room, specifically the distance between the opposite walls, allowed for the extraction of the orientation of the camera as well as two of the three coordinates for the position of the cameras - the height from the floor and the distances from the walls. Since the corners of the room were not visible, the only way to extract the final position coordinates was by aligning the point clouds from the two Kinects using the ICP (Iterated Closest Point) algorithm [63]. As the two Kinects were facing opposite directions, there was very little overlap between their two point clouds. However, since there was a single unknown, the alignment algorithm was constrained to search in the space of the only unknown dimension, i.e., the left-right shift. A large number of frames of the depth videos were used for computing the alignments to ensure robustness. The resulting alignment provided the full extrinsic calibrations (poses) of both the Kinect cameras.

Determining GoPro Camera Pose: We next locate both the GoPro cameras in the same world coordinate system. Note that the GoPro cameras are non-stationary hence their poses need to be extracted at each frame. This was done using a two stage algorithm. First, the extrinsic GoPro world poses were extracted on a small set of frames based on feature correspondences between the GoPro videos and the Kinect videos. These correspondences were manually selected to ensure accuracy. The feature points on the Kinect video were translated into 3D world coordinates using the depth data and the previously computed Kinect poses. Now, given the 2D image points on the GoPro video frames and the corresponding 3D points in the world along with the GoPro intrinsic camera matrices, their world poses were computed using the POSIT algorithm [64]. Due to the manual correspondences required, this was only done on a small subset of frames. Hence these frames were used as initializations of poses, and from these initializations, the extrinsic poses of the GoPro frames were tracked using standard monocular Visual-Odometry algorithms [65] which rely on local solutions of Structure-From-Motion (SFM) algorithms [66] that compute relative motions between frames. These algorithms make use of all available features in the images for tracking and thus are more robust to occlusions. Finally, bundle adjustment [67] was performed to optimize the tracked camera poses over the whole length of the video since camera poses extracted using Visual Odometry tend to drift slowly over time.

Facial Landmark Tracking and Eye Gaze: We tracked the faces of the participants opportunistically in the GoPro videos. We performed face detection and used a state-of-the-art facial landmark tracking software [68] to track 51 facial landmark points. We used a subset of these tracked landmarks points that are relatively stable (e.g. the corners of the eyes, the tip of the nose), to compute each playerss head pose (position and orientation relative to the GoPro cameras) using a standard rigid object pose estimation algorithm (POSIT) [64] by aligning the facial landmarks with the vertices of a generic face mesh model. The players head pose and location information obtained from the face tracking were transformed into the world coordinate frame using the extrinsic poses of the GoPro cameras. Additionally, when possible, the players eye gaze directions were also determined from the high resolution GoPro videos using gaze tracking algorithms which track the elliptical dark irises within the white sclera of the eyes. These gaze vectors were also translated into the world coordinate system. The eye positions and gaze vectors in world coordinate frame enable identifying the target of the players gaze. This information can be used for determining joint attention as well as eye contact. Furthermore, the facial landmark tracking may be used for facial expression recognition such as detecting smiles and frowns, while the head pose may be used for recognizing gestures such as head nods and shakes.

Tracking Blocks: We track the toy blocks used in the tower games over the entire duration of the game. The tracking was

done by a human annotator using a semi-automated annotation tool where the annotator sporadically marked the bounding boxes on the blocks and an automated tracking algorithm filled in the gaps. The tracking algorithm predicted the object track between pairs of annotated key frames based on color histogram features and feature point and edge tracking. The annotator was required to visually verify the veracity of the automated track interpolation and, if necessary, annotate more key frames to improve the tracks. The annotation was done in 2D on the GoPro videos and hence the precise 3D locations of the blocks are not available. However, using the previously estimated GoPro extrinsic poses, and the 2D tracking on the two GoPro camera videos, coarse 3D positions of the blocks were obtained by triangulating from the two GoPro views opportunistically.

D. Dataset Description

The dataset contains 190 game sessions, with average session length 3 minutes. Out of these, 112 have been preprocessed and annotated. We have a total of 39 unique subjects in our participant pool and $\binom{39}{2}$ pairs of players. Each player-pair typically participated in eight different game sessions - two of those sessions were of the **architect-builder** variant where the subjects switched roles; the next six were of the **distinct-objective** variant where the rules given to the subjects were changed every two sessions of the game along with the set of blocks available. In the annotated set, there are 100 instances of the **architect-builder** variant and 90 instances of the **distinct-objective** variant.

The final dataset consists of the following data types for each game session:

- Two Kinect videos (RGB).
- Two depth videos (depth encoded in RGB).
- Two GoPro videos (distortion corrected).
- Intrinsic and extrinsic calibrations for the Kinect cameras.
- Intrinsic calibrations and video frame aligned sequences of camera poses for the GoPro cameras.
- Kinect tracked skeletons for the two participants.
- Face and head pose tracking for the two individuals from the GoPro cameras when visible.
- Eye Gaze information (3D vectors) for the two participants whenever available.
- Object positions (2D bounding boxes) and tracks for all the blocks.

A dataset visualization software (Fig. 3) is available and will be released with the dataset.

E. Data Annotation

Since our focus is on *joint attention* and *entrainment*, we annotated 112 videos which were divided into 1213 10-second segments indicating the presence or absence of these two behaviors in each segment. To annotate the videos, we developed an innovative annotation schema drawn from concepts in the social psychology literature [25], [57]. The annotation schema is a series of questions, that could be used as a guideline to assist the annotators.

The questions regarding joint attention are:

- Are the two actors looking at one another during the engagement?
- Are the two actors looking at the same object during the engagement?
- During engagement, do the actors use body language while interacting?

The questions regarding entrainment are:

- During engagement, were the actors engaged in simultaneous movement?
- Rate the degree to which the two people in the clip seem to be "marching to the beat of the same drummer"
- How smoothly do the actors flow of behavior intertwine, or mesh evenly and smoothly?
- During the scene, how often do the actors imitate each other's motions and or body language?

The annotation schema associates high level social interaction predicates with more objectively perceptible measures. For example, *Joint attention* involves eye gaze and body language, and *entrainment* involves simultaneous movement, tempo similarity coordination and imitation. Each measure (except for imitation) was rated using a six-point Likert scale (0,0.2,0.4,0.6,1.0), where a zero indicates the absence of the measure during the segment, while a 1.0 denotes its presence during the entire 10 second segment, with values in between denoting intermediate durations of the measure. Imitation was measured according to the incidents that it occurred in, within the segment. A group of six undergraduate sociology and psychology students annotated the videos. The students were given a general introduction to the annotation schema and were then asked to code representative samples of the videos. The videos were annotated after ensuring that all the students as a group were annotating the sample videos accurately and reliably.

The initial results on the dataset will be published in a separate publication and will be available publicly along with the dataset.

V. Conclusions

We have introduced a new *Tower Game Dataset* consisting of audio-visual recordings of players engaging in dyadic interactions. Starting from a sound socio psychological basis, we focus on the interactive and cooperative aspects of the social interactions. We collect the *Tower Game Dataset* that consists of multimodal recordings of two players collaborating with each other to build a tower, using a rich array of sensors. Furthermore, we pre-process the dataset and extract a variety of low-level participant behaviors. The dataset has been annotated with the presence of social interaction predicates such as *joint attention* and *entrainment*. We believe that this dataset will enable developing of computational approaches to modeling and analysis of social interactions.

References

[1] S. Baron-Cohen, "Mindblindness: An essay on autism and theory of mind," in *MIT*, 1997.

[2] S. M. Schaafsma, D. W. Pfaff, R. P. Spunt, and R. Adolphs, "Understanding other minds: Perspectives from developmental social neuroscience," in *Oxford*, 2014.

[3] K. Shockley, D. Richardson, and R. Dale, "Conversation and coordinative structures," in *TCS*, 2009.

[4] K.L. Marsh et al., "Contrasting approaches to perceiving and acting with others," in *EP*, 2006.

[5] R. W. Picard, *Affective Computing*, MIT Press, 1995.

[6] Michael Tomasello, *The Cultural Origins of Human Cognition*, Harvard University Press, 2001.

[7] Stephen C. Levinson, "On the human "interaction engine"," in *Roots of Human Sociality Culture, Cognition and Interaction*. Berg, 2006.

[8] M. M. Louwerse, R. Dale, E. G. Bard, and P. Jeuniaux, "Behavior matching in multimodal communication is synchronized," *Cognitive Science*, 2012.

[9] Ezequiel Di Paolo and Hanne De Jaegher, "The interactive brain hypothesis," *Frontiers in Human Neuroscience*, 2012.

[10] Natalie Sebanz and Guenther Knoblich, "Prediction in joint action: What, when, and where," *Topics in Cognitive Science*, 2009.

[11] Valentina Fantasia, Hanne De Jaegher, and Alessandra Fasulo, "We can work it out: an enactive look at cooperation," *Frontieres in Psychology*, 2014.

[12] H. De Jaegher, E. Di Paolo, and S. Gallagher, "Can social interaction constitute social cognition?," *Trends in Cognitive Science*, 2010.

[13] Rafael Calvo and Sidney D'Mello, "Affect detection: An interdisciplinary review of models, methods, and their applications," in *IEEE Transactions on Affective Computing*, 2010.

[14] Sidney D'Mello, Rosalind W. Picard, and Arthur Graesser, "Toward an affect-sensitive autotutor," in *IEEE Intelligent Systems*, 2007.

[15] S. Ghosh, M. Chatterjee, and L.-P. Morency, "A multimodal context-based approach for distress assessment," in *ICMI*, 2014.

[16] Mohamed Amer, Behjat Siddiquie, Colleen Richey, and Ajay Divakaran, "Emotion detection in speech using deep networks," in *ICASSP*, 2014.

[17] Ying li Tian, Takeo Kanade, and Jeffrey F. Cohn, "Recognizing action units for facial expression analysis," in *IEEE PAMI*. 2001.

[18] Selene Mota and Rosalind W. Picard, "Automated posture analysis for detecting learners interest level," in *CVPRW*, 2003.

[19] B. Siddiquie, S. Khan, A. Divakaran, and H. Sawhney, "Affect analysis in natural human interactions using joint hidden conditional random fields," in *ICME*, 2013.

[20] G. Ramirez, T. Baltrusaitis, and L. P. Morency, "Modeling latent discriminative dynamic of multi-dimensional affective signals," in *ACII*, 2011.

[21] Mohamed Amer, Behjat Siddiquie, Saad Khan, Ajay Divakaran, and Harpreet Sawhney, "Multimodal fusion using dynamic hybrid models," in *WACV*, 2014.

[22] M.S. Ryoo and J.K. Aggarwal, "Semantic representation and recognition of continued and recursive human activities," in *IJCV*, 2009.

[23] T. Lan, Y. Wang, W. Yang, S. Robinovitch, and G. Mori, "Discriminative latent models for recognizing contextual group activities," in *PAMI*, 2012.

[24] James M. Rehg and et . al, "Decoding children's social behavior," *CVPR*, 2013.

[25] Frank J. Bernieri, "Coordinated movement and rapport in teacher-student interactions," *Journal of Non-Verbal Behavior*, 1988.

[26] T Kanade, J. F Cohn, and Y. Tian, "Comprehensive database for facial expression analysis," in *FG*, 2000.

[27] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended cohn-kanade dataset (ck+): A complete expression dataset for action unit and emotion-specified expression," in *CVPR4HB*, 2010.

[28] M. Grimm, K. Kroschel, and S. Narayanan, "The vera am mittag german audio-visual emotional speech database," in *ICME*, 2008.

[29] B. Schuller and et. al., "Audiovisual recognition of spontaneous interest within conversations," in *ICMI*, 2007.

[30] H. Gunes and M. Piccardi, "A bimodal face and body gesture database for automatic analysis of human nonverbal affective behavior," in *ICPR*, 2006.

[31] A. Kleinsmith, P. Ravindra De Silvaa, and N. Bianchi-Berthouze, "Cross-cultural differences in recognizing affect from body posture," in *Interacting with Computers*, 2006.

[32] S. M. Mavadati, M. H. Mahoor, K. Bartlett, and Jeff Cohn, "Disfa: A spontaneous facial expressions dataset," in *IEEE Transactions on Affective Computing*, 2013.

[33] Xing Zhang and et. al., "Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database," in *IVC*, 2014.

[34] B. Schuller and et al., "Avec 2011 -the first international audio visual emotion challenge," in *ACII*, 2011.

[35] C. Busso and et. al., "Iemocap: Interactive emotional dyadic motion capture database," in *Journal of Language Resources and Evaluation*, 2008.

[36] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic, "A multimodal database for affect recognition and implicit tagging," in *IEEE Transactions on Affective Computing*, 2012.

[37] J. M. Chaquet, E. J. Carmona, and A. Fernndez-Caballero, "A survey of video datasets for human action and activity recognition," in *CVIU*, 2013.

[38] Xiaofan Sun, Jeroen Lichtenauer, Michel Valstar, Anton Nijholt, and Maja Pantic, "A multimodal database for mimicry analysis," in *ACII*, 2011.

[39] D. Premack and G. Woodruff, "Does the chimpanzee have a theory of mind?," in *BBS*, 1978.

[40] A. Waytz, K. Gray, N. Epley, and D. M. Wegner, "Causes and consequences of mind perception," in *TCS*, 2010.

[41] S. Baron-Cohen, M. Lombardo, H. Tager-Flusberg, and D. Cohen, "Deconstructing and reconstructing theory of mind," in *TCS*, 2013.

[42] V. Reddy, "How infants know minds. harvard university press," in *Harvard*, 2011.

[43] A. Fogel and A. Garvey, "Alive communication," in *ICD*, 2007.

[44] M. Tomasello, "The cultural origins of human cognition," in *Harvard*, 2009.

[45] M. Tomasello, "Constructing a language: A usage-based theory of language acquisition.," in *HarvardSP*, 20094.

[46] A. Kendon, "Some functions of gaze-direction in social interaction," in *AP*, 1967.

[47] M. Kidwell and D. H. Zimmerman, "Joint attention as action.," in *JP*, 2007.

[48] E. A. Schegloff, "Sequence organization in interaction: Volume 1: A primer in conversation analysis," in *Cambridge*, 2007.

[49] C. Goodwin, "Gestures as a resource for the organization of mutual orientation," in *Semiotica*, 1986.

[50] T. L. Chartrand and R. Van Baaren, "Human mimicry," in *AESP*, 2009.

[51] M. J. Richardson, S. J. Harrison, R. W. Kallen, A. Walton, B. A. Eiler, E. Saltzman, and R. C. Schmidt, "Self-organized complementary joint action: Behavioral dynamics of an interpersonal collision-avoidance task.," in *JEP*, 2015.

[52] K. et al. Grammer, "Fuzziness of nonverbal courtship communication unblurred by motion energy detection," in *JPSP*, 1999.

[53] F. Ramseyer and W. Tschacher, "Synchrony: A core concept for a constructivist approach to psychotherapy," in *CHS*, 2006.

[54] Ronald Poppe, Sophie Van Der Zee, Dirk Heylen, and Paul. Taylor, "Amab: Automated measurement and analysis of body motion," in *BRM*, 2013.

[55] Karen Bartling, Franziska Kopp, , and Ulman Lindenberger, "Maternal affect attunement: Refinement and internal validation of a coding scheme," in *EJDS*, 2010.

[56] J. M. Gottman, K. McCoy, J. A. Coan, and H. Collier, "The specific affect coding system (spaff) for observing emotional communication in marital and family interaction," in *Erlbaum*, 1995.

[57] Lauren Adamson, Roger Bakeman, Deborah Deckner, and Nelson Booke, "Rating parent-child interactions: Joint engagement, communication dynamics and shared topics in autism, down syndrome, and typical development," *Journal of Autism Development Disorders*, 2012.

[58] S. R. Langton, R. J. Watt, and V. Bruce, "Do the eyes have it? cues to the direction of social attention.," in *TCS*, 2000.

[59] F. Rossano, "Gaze in conversation," in *HCA*, 2013.

[60] A. P. Bayliss and et. al., ""gaze leading": Initiating simulated joint attention influences eye movements and choice behavior.," in *JEP*, 2012.

[61] Julie Shah and Cynthia Breazeal, "An empirical analysis of team coordination behaviors and action planning with application to human-robot teaming," in *Human Factors*, 2010.

[62] Jean-Yves Bouguet, "Camera calibration toolbox for matlab," in *http://www.vision.caltech.edu/bouguetj/calib$_d$oc/*, 2013.

[63] Zhengyou Zhang, "Iterative point matching for registration of free-form curves and surfaces," in *IJCV*, 1994.

[64] Daniel F. DeMenthon and Larry S. Davis, "Model based object pose in 25 lines of code," in *IJCV*, 1995.

[65] D. Nister, O. Naroditsky, and J. Bergen, "Visual odometry," in *CVPR*, 2004.
[66] Richard Hartley and Andrew Zisserman, "Multiple view geometry in computer vision," in *Cambridge University Press*, 2003.
[67] B.Triggs, P. F. McLauchlan, R. Hartley, and A. Fitzgibbon, "Bundle adjustmenta modern synthesis," in *Vision algorithms: theory and practice*, 2000.
[68] Xuehan Xiong and Fernando De la Torre, "Supervised descent method and its applications to face alignment," in *CVPR*, 2013.