Towards Social Artificial Intelligence: Nonverbal Social Signal Prediction in A Triadic Interaction (Supplementary Material) [*]

Hanbyul Joo[1†] Tomas Simon[1‡] Mina Cikara[2] Yaser Sheikh[1]
[1]Carnegie Mellon University [2]Harvard University
{hjoo, tsimon, yaser}@fb.com, mcikara@fas.harvard.edu

1. The Haggling Game Protocol

To evoke natural interactions, we involved participants in a social game named the *Haggling* game. We invent this game to simulate a haggling situation among two sellers and a buyer. The triadic interaction is chosen to include interesting social behaviors such as turn taking and attention changes, which are missing in previous dyadic interaction datasets [8]. During the game, two sellers are promoting their own comparable products for selling, and a buyer makes a decision about which product he/she buys between the two. The game lasts for a minute, and the seller who has sold his/her product is awarded $5. To maximize the influence of each seller's behavior on the buyer's decision-making, the items assigned to sellers are similar products with slightly different properties. Example items are shown in Table 1.

For every capture, we follow the protocol described below. We randomly recruited participants using the CMU Participant Pool[1]. Over the 8 days of captures, 122 subjects participated and 180 haggling sequences were captured (about 3 hours of data). The participants arrive at the lab for the capture and first sign the IRB consent form with an agreement to publicly release the data for research purposes only. A unique identification number is assigned to each participant, and participants are also equipped with a wireless microphone. Then, all subjects are informed of the rules of the Haggling game by watching a pre-recorded presentation video together. Notably, they are not instructed about how to behave during the game, nor is their clothing or appearance controlled. All motions in the sequences are spontaneous social behaviors based on the informed game rules. After introducing the game rules, participants are asked to spend time inside the studio (as shown in Figure 1) so that they can be accustomed to the interior view of the Panoptic Studio [4, 5]. Before starting the capture, groups and roles are randomly assigned, and participants line up based on their numerical orders. We provide written descriptions to sellers about the items they will be selling in small cards 1 minute before the game, and the sellers return the card before entering the studio. With a starting signal, participants in a group enter the studio and start the haggling game immediately. The positions and orientations of the groups inside the system are also spontaneously decided (no instructions are given). During the capture, their social signals including voice, positions, orientations, and body motions are recorded. We send a signal by ringing a bell 10 seconds before the end of the game, and send the same alarm at the end of the game. After the capture, the buyer annotates the decision between the two items in the prepared result sheet. The captured sequences contain many voluntary social behaviors of diverse people in a common social context. Example scenes are shown in Figure 9 and our supplementary video.

Figure 1: Before starting the social game capture, participants are instructed the game rules and also spent time to be accustomed to the Panoptic Studio environment, as shown in these photos. We follow a common and strict protocol during all captures to avoid any potential bias.

[*] Website: http://domedb.perception.cs.cmu.edu/ssp
[†] H. Joo is now at Facebook AI Research (FAIR)
[‡] T. Simon is now at Facebook Reality Labs
[1] https://cbdr.cmu.edu/

Items	Seller 1	Seller 2
Phone	Light weight Medium storage	Medium weight Large storage
Laptop	Light weight Medium speed	Medium weight Fast speed
Tablet PC	Large storage Medium speed	Medium storage Fast speed
Speaker	High quality audio Wired	Medium quality audio Wireless

Table 1: Examples of items assigned to sellers in our Haggling games.

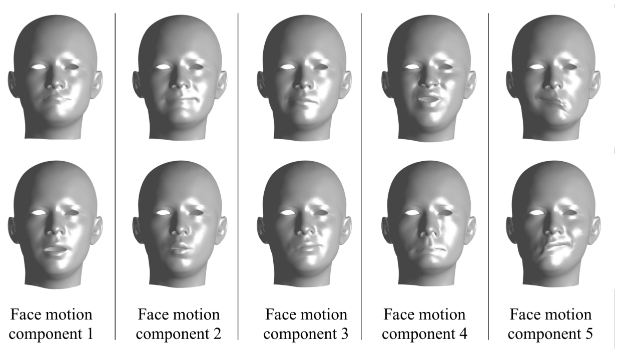

Figure 2: The five face motion components (showing parameter weights with -0.3 on the top and 0.3 on the bottom) used in our social signal modeling.

2. Face Motion Parameters

For the face motion signal, we first fit the deformable face model of [1] and use the initial 5 dimensions of the facial expression parameters, because we found the remaining dimensions have an almost negligible impact on our reconstruction quality. Note that the face expression parameters in [1] are sorted by their influence by construction and the initial components have more impact in expressing facial motion. To this end, face motion at a time instance is represented by a 5-dimensional vector, $\mathbf{F}(t) \in \mathbb{R}^5$, and example facial expressions expressed by each component are shown in Figure 2.

3. Implementation Details of Social Signal Predictions

In this section, we discuss the details of our neural network architectures to implement the social signal prediction models for each sub-task.

3.1. Predicting Speaking

Our neural network is composed of four 1-D convolutional layers (see Figure 3). The first three layers output 128, 256, and 512 dimensional features respectively with ReLU activation functions, and the last layer has 1×1 convolutions with a sigmoid activation layer. Dropout [9] is also applied for the second and third layers with a probability of 0.25. Our model does not require a fixed window size for the input (since it is fully convolutional), but we separate input data into small clips with a fixed size (denoted by f) for efficiency during training. During testing time, our models can be applied to the input of arbitrary length. We use $f = 120$ (4 seconds) for the input window size for training, and we use an arbitrary length of input data for testing. The feature dimension of the input to our network is the concatenation of face motion and body motion (78 dimensions). If fewer cues are used (e.g., face only or body only), we mask out the unused channels as their average values computed in the training set and keep the same network structure. We use an adaptive gradient descent algorithm, AmsGrad algorithm [7], implemented in PyTorch [6], along with l^1 regularization loss with 0.001 as the regularization strength.

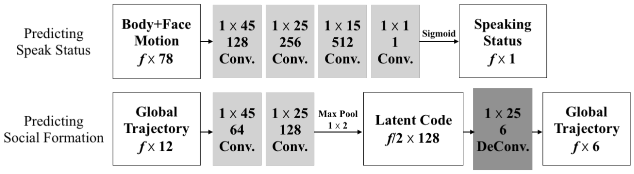

Figure 3: Network architectures for predicting speaking status and predicting social formation. We use fully convolutional networks for both sub-tasks.

3.2. Predicting Social Formations

Our neural network has an autoencoder structure, where the encoder is composed of two 1-D convolutional layers followed by a max pooling layer with stride 2, and the decoder is composed of a single 1-D transposed convolution layer (see Figure 3). The output feature dimensions are 64, 128, and 6 respectively. Dropout [9] is also applied in front of all layers with a probability of 0.25. Similar to the speaking status prediction, our model does not require a fixed window size for the input, but we separate input data into small clips with a fixed size ($f = 120$, or 4 seconds) for the efficiency in training. The input is the concatenation of the cues of the other two communication partners (12 dimensions) with a fixed order (buyer and then the right seller), and the output of our network is the position and orientations of the target individual, the left seller (6 dimensions). Similar to the previous prediction task, if fewer cues are used (e.g., position only), we mask out the unused channels as their average values computed in the training set and keep the same network structure. We use an adaptive gradient descent algorithm, AmsGrad algorithm [7], implemented in

PyTorch [6], along with l^1 regularization loss with 0.1 as the regularization strength.

3.3. Predicting Body Gestures (Kinesic Signals)

As in the work of Holden et al. [3], we first train an body motion autoencoder (as shown in the first row of Figure 4) to find the motion manifold space, so that the decoded output from the latent space can express a reasonable human body motion. Then, we keep the decoder part of this network (shown as the blue boxes in Figure 4) for the gesture prediction, which uses the latent codes generated by the following two different approaches as input.

From the Body Trajectory: We regress the latent code for the gesture prediction from the estimated trajectory information of the target person (position and body orientation). The network architecture is shown in the second row of Figure 4. Note that we freeze and do not train the decoder part (the blue box) which is taken from the body motion autoencoder. As input, the model uses the velocities of position and body orientation (relative root movements with respect to the previous frame), which is a subpart of our body motion representation (the first 3-dimensions out of the 73-dimensional vector). For training, we use ground truth body motion data, by using the subpart representing relative root movements as input, and all dimensions for body motion as output. During testing, we convert the social formation prediction output (global position and orientation) into this velocity representation (relative position and orientation), and use it as the input for this network.

From Other Body Gestures: In this case, we use the other two partners' body motions as input to generate the latent code, and decode it to predict the target individual's body gesture, similar to the previous approach. The network architecture is shown in the third row of Figure 4.

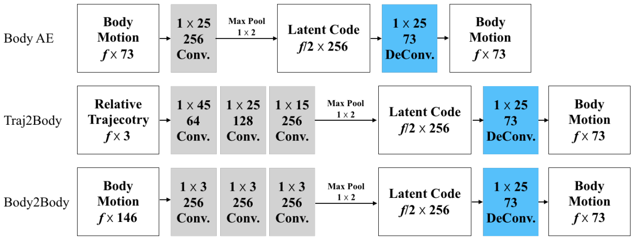
Body AE
Body Motion $f \times 73$
1 × 25 256 Conv.
Max Pool 1 × 2
Latent Code $f/2 \times 256$
1 × 25 73 DeConv.
Body Motion $f \times 73$
Traj2Body
Relative Trajecotry $f \times 3$
1 × 45 64 Conv.
1 × 25 128 Conv.
1 × 15 256 Conv.
Max Pool 1 × 2
Latent Code $f/2 \times 256$
1 × 25 73 DeConv.
Body Motion $f \times 73$
Body2Body
Body Motion $f \times 146$
1 × 3 256 Conv.
1 × 3 256 Conv.
1 × 3 256 Conv.
Max Pool 1 × 2
Latent Code $f/2 \times 256$
1 × 25 73 DeConv.
Body Motion $f \times 73$

Figure 4: Network architectures for body gesture prediction. We adopt the autoencoder architecture of Holden et al. [3] to learn the latent space for by motion shown in the first row. We consider two different approaches to generate the latent code, from the predicted social formation of the target individual or from the communication partners' body motion. Note that the decoder part shown as the blue boxes are frozen after training the body motion autoencoder.

Table 2: Average distances (cm) between subjects. B, RS, and LS denote buyer, right seller, and left seller respectively.

	Avg. dist.	Std.	Min	Max
B-RS	148.11	27.26	99.03	265.52
B-LS	151.45	29.62	104.24	284.85
LS-RS	124.13	24.05	77.70	206.26

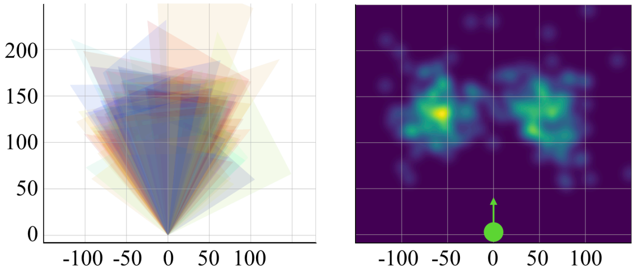

Figure 5: Visualizing social formations in the haggling sequences as triangles (left) and a heat map (right). The formation is normalized w.r.t the buyer's location, and the green circle on the right shows the buyer location (origin) and orientation (z-axis).

4. Revisiting Proxemics

Our dataset has the measurement of fully spontaneous motions (including the position and orientation of groups) of interacting people, and enables us to revisit the well-known proxemics theory [2]. We first compute the average, minimum, and maximum distances between a pair of subjects: (1) buyer and right sellers (B-RS), (2) buyer and left seller (B-LS), and (3) left seller and right seller (LS-RS). The results are shown in Table 2. We found that the result approximately follows the social distance categories defined in the Hall's categorization [2]. The average distances among subjects are within the close phase of social distance ranges (from 120 cm to 210 cm) and the max distances are within the far phase of social distance (from 210 cm to 370 cm) in [2]. To analyze the shape of the social formation, we plot the average formation of games in a person-centric coordinate by a buyer. The results are shown in the Figure 5, showing that the formation is often similar to isosceles triangles with relatively far distances between a buyer and two sellers than the distance between sellers.

5. Further Analysis on Speaking Status Prediction

Result on Inter-personal Signals. As shown in the second column (other seller's input) of Table 1 of the paper, the result clearly shows that there exists a strong link between

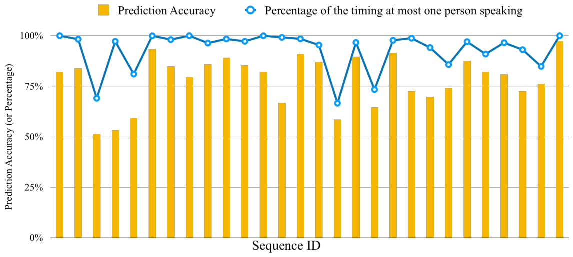

Figure 6: Comparison between the performance of speaking status prediction and turn-taking status for each sequence. Each column shows the prediction performance (yellow bar) where the other seller's face and body signals are used as input. The blue curve represents how well the turn-taking rule is satisfied, which is defined by counting the percentage of the timing where at most one person is speaking.

interpersonal social signals. The other seller's facial motion shows a strong predictive power for the target person's speaking status, where the accuracy is higher than the case of using the target person's own body signals as input, presumably due to the turn-taking property in social communication. For example, we can assume that the target person is not speaking, when the other seller is speaking. We can further investigate this by checking how well the turn-taking rule is satisfied during each social game scene, along with its predicting performance. As a way to measure the turn-taking status, we consider the percentage of the timing at which at most one person speaks, which defined by:

$$\frac{\sum_t \delta\left(\mathbf{S}^0(t) + \mathbf{S}^1(t) < 2\right)}{T}, \qquad (1)$$

where T is the total time of a Haggling game, $\mathbf{S}$ is the speaking status for sellers, and δ is a function that returns 1 if the condition satisfies and returns 0 otherwise. In this measurement, 100% means that there is no time that both sellers are speaking at the same time, where the turn-taking rules are perfectly satisfied. We compute this measurement to check the turn-taking status for each testing sequence as the blue curve in Figure 6. In this figure, we also plot the speaking prediction accuracy for each testing sequence by using the other seller's both face and body signals as input, which is shown as yellow bars. As shown in the figure, the prediction performance shows a very similar pattern to this turn-taking status, and this means that this implicit social "rule" is a source of linking the social signals across individuals. Example qualitative results are shown in the Figure 7.

6. Qualitative Results

Example results of speaking classification and social formation prediction are shown in Figure 7 and Figure 8. Results are best seen in the supplementary videos.

References

[1] Chen Cao, Yanlin Weng, Shun Zhou, Yiying Tong, and Kun Zhou. Facewarehouse: A 3d facial expression database for visual computing. In *TVCG*, 2014. 2

[2] Edward Twitchell Hall. The hidden dimension. Doubleday & Co, 1966. 3

[3] Daniel Holden, Jun Saito, and Taku Komura. A deep learning framework for character motion synthesis and editing. In *TOG*, 2016. 3

[4] Hanbyul Joo, Hao Liu, Lei Tan, Lin Gui, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic studio: A massively multiview system for social motion capture. In *CVPR*, 2015. 1

[5] Hanbyul Joo, Tomas Simon, Xulong Li, Hao Liu, Lei Tan, Lin Gui, Sean Banerjee, Timothy Godisart, Bart Nabbe, Iain Matthews, et al. Panoptic studio: A massively multiview system for social interaction capture. In *TPAMI*, 2017. 1

[6] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *ICLR*, 2017. 2, 3

[7] Sashank J Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of adam and beyond. In *ICLR*, 2018. 2

[8] James Rehg, Gregory Abowd, Agata Rozga, Mario Romero, Mark Clements, Stan Sclaroff, Irfan Essa, O Ousley, Yin Li, Chanho Kim, et al. Decoding children's social behavior. In *CVPR*, 2013. 1

[9] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. In *The Journal of Machine Learning Research*, 2014. 2

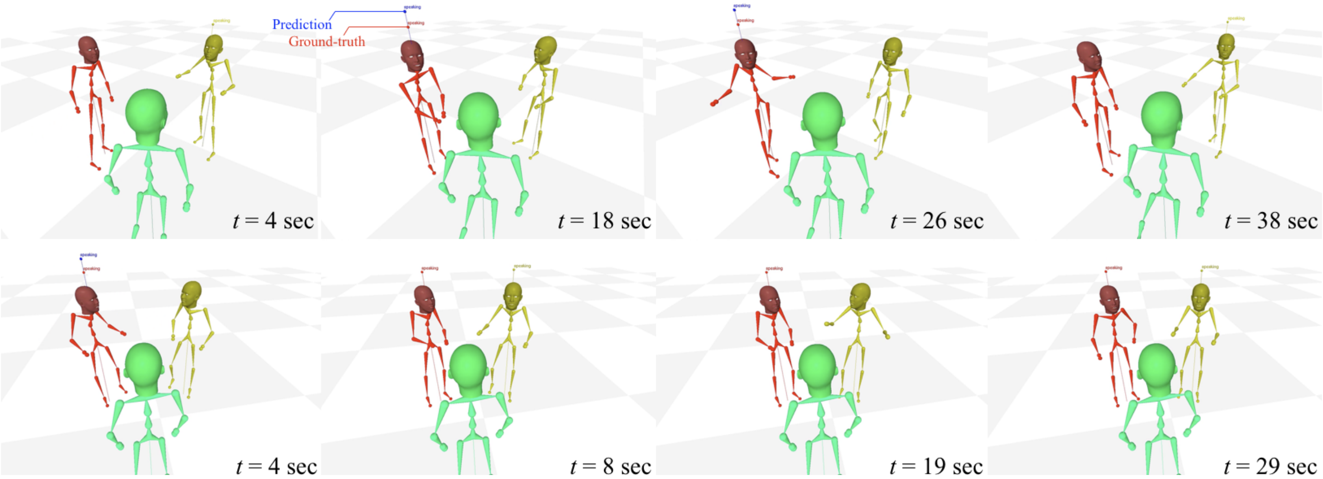

Figure 7: Qualitative results of the speaking prediction of the target person (red) by using the other seller's (yellow) face and body motions as input. The speaking prediction output is shown as the blue "speaking" label above the target person's head, while the ground truth speaking status is shown as the red label. The prediction is accurate, if both blue and red labels are shown or not shown. Examples scenes of two haggling games (top and bottom) are shown, where the sequence on the top has high accuracy (89%) and the sequence on the bottom has low accuracy (58%). In the haggling game shown on the top, both sellers follow turn taking almost always, while both sellers frequently speak at the same time in the sequence shown on the bottom.

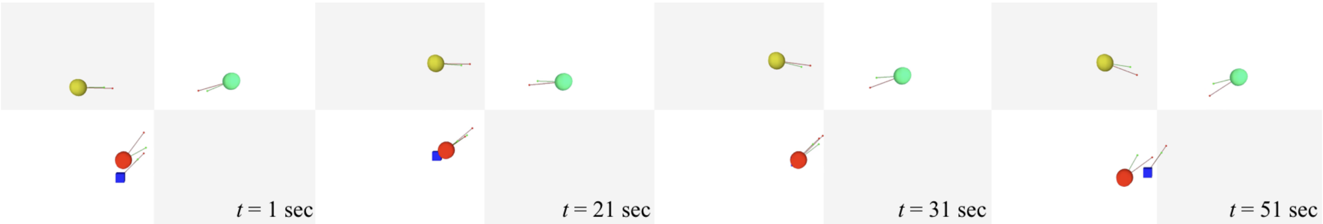

Figure 8: Qualitative results of the social formation prediction of a haggling game, visualized from the view at the top. The target person is shown as red spheres. The cues from other people (yellow and cyan spheres) are used as input for the prediction, and the prediction output is shown as the blue cube. The red lines represent body orientations, and the green lines represent face orientations.

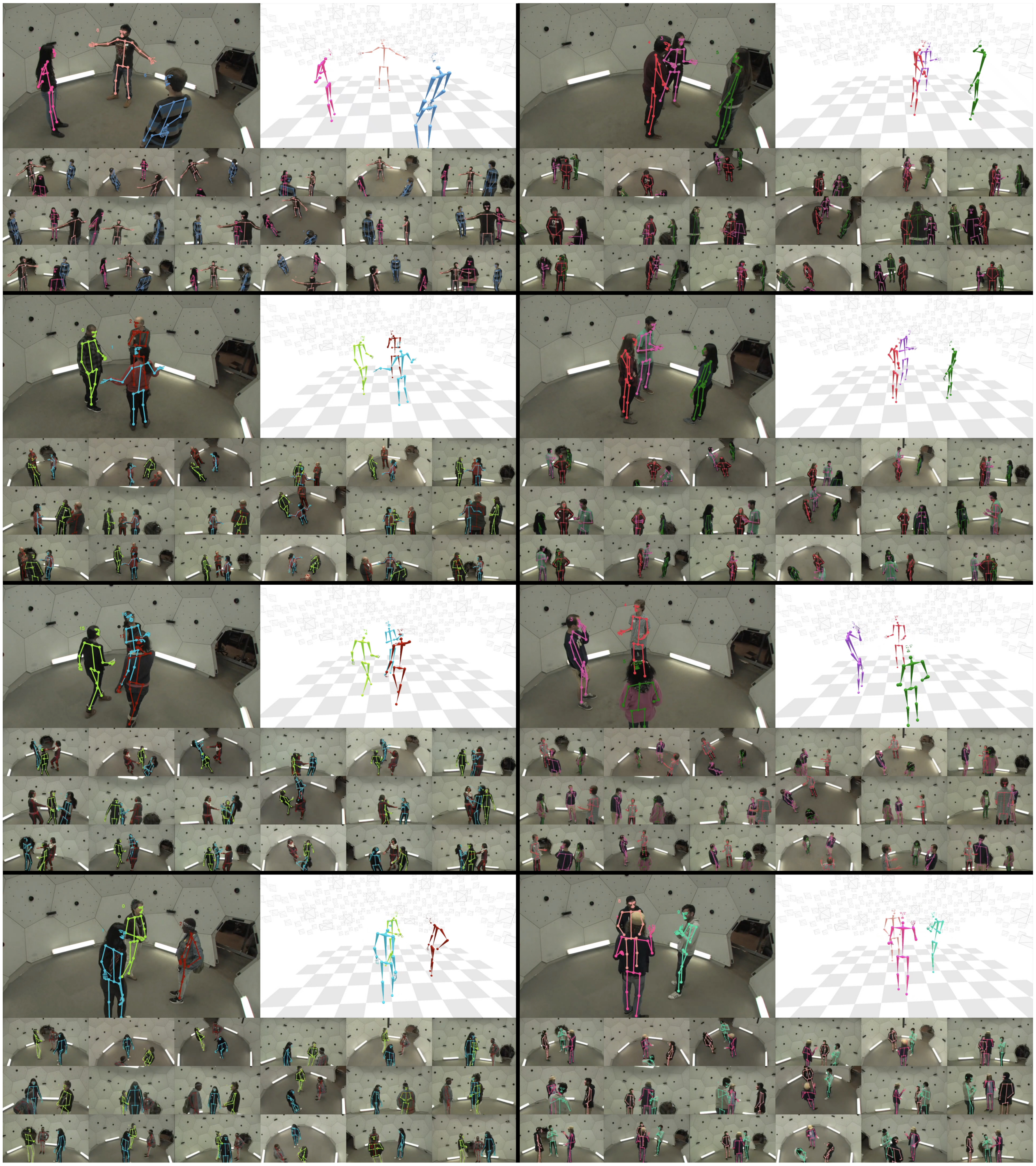

Figure 9: Example scenes of haggling sequences with social signal measurements. For each example, HD images overlaid by the projections of 3D anatomical keypoints (from bodies, faces, and hands) are shown, along with a 3D view of the social signal measurements (top right).