ICCV 2021 Understanding Social Behavior in Dyadic and Small Group Interactions Challenge

Fact sheet: Behavior forecasting Track

I. TEAM DETAILS

- Team leader name: Nguyen Tan Viet Tuyen
- Username on Codalab: tuyennguyen
- Team leader affiliation: Social AI and Robotics Lab (SAIR), Department of Engineering, King's College London, United Kingdom.
- Team leader email: tan_viet_tuyen.nguyen@kcl.ac.uk
- Name of other team members (and affiliation):
  Oya Celiktutan, SAIR, Department of Engineering, King's College London, United Kingdom.
- Team website URL:
  https://sairlab.github.io

II. CONTRIBUTION DETAILS

A. Forecasting Nonverbal Social Signals during Dyadic Interactions with Generative Adversarial Neural Networks

We are approaching a future where social robots will progressively become widespread in many aspects of our daily lives, including education, healthcare, work, and personal use. All of such practical applications require that humans and robots collaborate in human environments, where social interaction is unavoidable. Along with verbal communication, successful social interaction is closely coupled with the interplay between nonverbal perception and action mechanisms, such as observation of one's gaze behaviour and following their attention, coordinating the form and function of hand-arm gestures. Humans perform nonverbal communication in an instinctive and adaptive manner, with no effort. For robots to be successful in our social landscape, they should therefore engage in social interactions in a human-like way [1], with increasing levels of autonomy. In particular, non-verbal gestures are expected to endow social robots with the capability of emphasizing their speech, or showing their intentions. Motivated by this, our research sheds a light on modeling human behaviors in social interactions, specifically, forecasting human non-verbal social signals during dyadic interactions, with an overarching goal of developing robotic interfaces that can learn to imitate human dyadic interactions. Such an approach will ensure the messages encoded in the robots' gestures could be perceived by interacting partners in a facile and transparent manner, which could help improve the interacting partner's perception and makes the social interaction outcomes enhanced [2].

B. Overview of the Proposed Approach

Fig. 1 illustrates the proposed training framework to forecast the non-verbal information $\hat{P}_{fo}^{k+1:T}$ $(k+1 \leq t \leq T)$ of the target person during dyadic interaction. The process is started by encoding face, body and hand landmarks $P_{ob}^{0:k}$ $(0 \leq t \leq k)$ of the interacting partner into a contextual vector c. It is followed by dividing the non-verbal features of the target person $P_{fo}^{0:k}$ at the same observed time window into three parts, namely face $P_{fo\_f}^{0:k}$, body $P_{fo\_b}^{0:k}$, and hands $P_{fo\_h}^{0:k}$, each of them is injected to a corresponding generator network. By combining with the contextual vector c, the three generator networks predict future landmarks of face $\hat{P}_{fo\_f}^{k+1:T}$, body $\hat{P}_{fo\_b}^{k+1:T}$, and hands $\hat{P}_{fo\_h}^{k+1:T}$. Those features are concatenated into the single vector $\hat{P}_{fo}^{k+1:T}$ representing the predicted motion of the target person. Finally, both of the real gesture $P_{fo}^{k+1:T}$ and the predicted one $\hat{P}_{fo}^{k+1:T}$ are fed to the Discriminator network. The following section will explain the proposed method in more detail.

C. Detailed method description

1) Encoder: : During social human-human interaction, people tend to use a wide range of non-verbal channels to communicate their intentions or emotions [3]. Such social signals would influence the other interlocutor's perception and actions, in particular, their non-verbal gestures. In other words, the non-verbal signals of the interacting partner should be treated as the essential stimuli to forecast the target person's motions. Inspiring from that, in the framework presented in Fig.1, a Recurrent Neural Network (RNN) *Encoder* is introduced to encode the interacting partner's motion $P_{ob}^{0:k}$ consisting of face, body, and hands represented by $2D$ joint coordinates into a vector c. The network is designed with a Long-Short Term Memory (LSTM) layer and followed by a fully connected layer to output the fix-length contextual information c. Finally, c is fed to the three *Generator* networks and treated as the conditional information to generate the predicted motion.

2) Generator: : The framework illustrated in Fig. 1 is equipped with three *Generator* networks, namely G_{face}, G_{body}, and G_{hand}. Those are implemented to forecast the motions of face, body, and hands. This strategy allows different motion features could be treated in appropriate manners. It should be noticed that we address the problem of motion prediction by creating a sequence-to-sequence network including $G_{Encoder}$ and $G_{Decoder}$, they are built upon
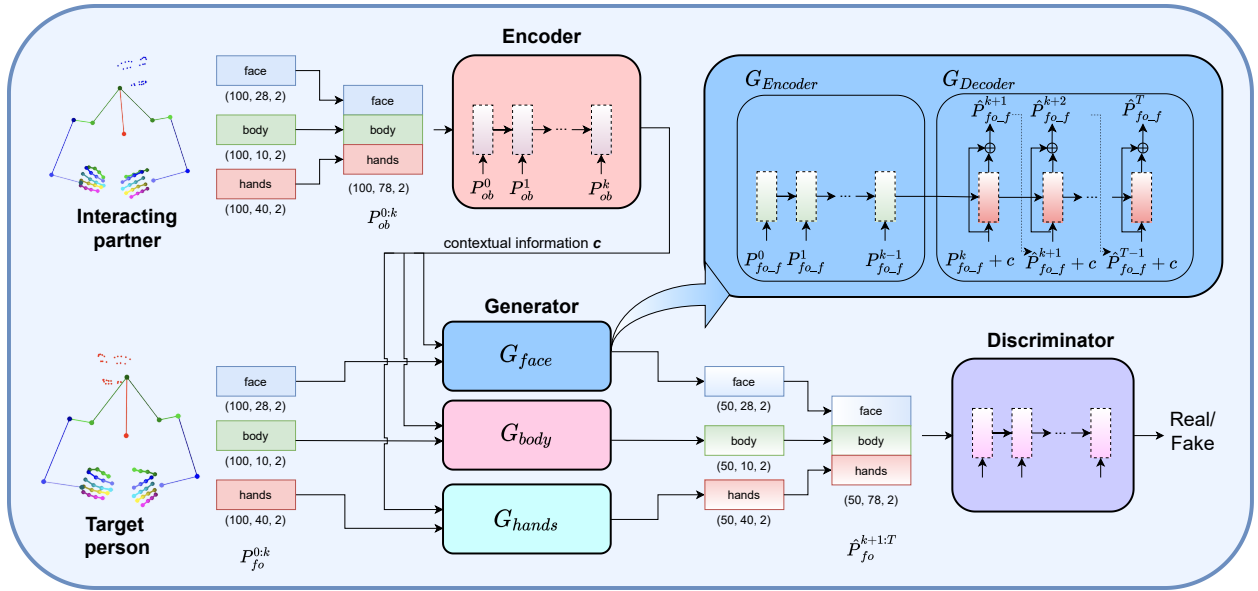
Fig. 1: The designed framework to forecast the non-verbal behaviors $\hat{P}_{fo}^{k+1:T}$ of the target person during socially dyadic interaction. Inputs consist of previous information of the target person $P_{fo}^{0:k}$ and the interacting partner $P_{ob}^{0:k}$.

LSTM layers. $G_{Decoder}$ receives the internal representation encoded by $G_{Encoder}$ and contextual vector c to generate the predicted motion. Noticed that a residual connection is added between the input and the output of each LSTM cell of $G_{Decoder}$, this approach allows the network to better model the velocity of motion [4]. At each time step, instead of employing the "teacher forcing" technique, $G_{Decoder}$ receives its own prediction to forecast the next motion frame.

3) Discriminator: : $Discriminator$ is created with a LSTM layer and a fully connected layer to produce an output probability indicating whether the input motion is real or fake. By training $Discriminator$ a capability of distinguishing between $P_{fo}^{k+1:T}$ and $\hat{P}_{fo}^{k+1:T}$, the adversarial loss [5] encourages $Generator$ to produce more realistic motions.

Overall, the framework demonstrated in Fig. 1 is trained with the loss functions defined in Eq. 1 and Eq. 2. We used $\mathcal{L}_G$ to train the $Encoder$ and $Generator$ while $\mathcal{L}_D$ is taken into account for optimizing $Discriminator$. Here, $\mathcal{L}_f^{MSE}$, $\mathcal{L}_b^{MSE}$, $\mathcal{L}_h^{MSE}$ are the mean square errors between the ground truth and the generated motions of face, body, and hands. α_1, α_2, α_3, and β are parameters to control the weights of the loss terms.

$$\begin{aligned} \mathcal{L}_G = \alpha_1 * \mathcal{L}_f^{MSE} + \alpha_2 * \mathcal{L}_b^{MSE} + \alpha_3 * \mathcal{L}_h^{MSE} \\ + \beta * \mathbb{E}\left[log(1 - D(\hat{P}_{fo}^{k+1:T}))\right] \end{aligned} \quad (1)$$

$$\mathcal{L}_D = -\mathbb{E}\left[log(D(P_{fo}^{k+1:T}))\right] - \mathbb{E}\left[log(1 - D(\hat{P}_{fo}^{k+1:T}))\right] \quad (2)$$

D. Challenge results

Pre-processing:

For training the framework, we used the training set of the UDIVA v0.5 [6]. On each interaction session, annotation data of two interlocutors were pad into 150 frames (6 seconds) ($P^{0:T} \in \mathbb{R}^{150 \times 78 \times 2}$). The first 100 frames (4 seconds) were used as the observed windows, and the last 50 frames (2 seconds) were used as the ground truth motions. As described in Eq. 3, joint coordinates of the motion frame $P^i (0 \leq i \leq 150)$ were normalized taking into account the mean μ and standard deviation σ of the whole motion sequence. Finally, we obtained 30964 training samples.

$$P^i = \frac{P^i - \mu}{\sigma + 10^{-8}} \quad (3)$$

Experimental results:

The training data was fed to the framework with a batch size of 1024. The model was trained for 1000 epochs. We used the Adam optimizer at the learning rate of 5×10^{-4} for $Encoder$, $Generator$, and $Discriminator$. Weights for the loss term were chosen empirically ($\alpha_1 = 10$, $\alpha_2 = 10$, $\alpha_3 = 10$, and $\beta = 1$). The adversarial loss mentioned in Eq. 1 was not used during the first 50 warm-up epochs.

At the testing phase, pairs of motion data collected from the target person $P_{fo}^{0:k}$ and the interacting partner $P_{ob}^{0:k}$ in a period of 4 seconds were sequentially filled in the framework. $Generator$ releases predicted motions of the target person $P_{fo}^{k+1:T}$ in the next 2 seconds. Fig. 1 presents a generated action conducted on the testing data. Using the evaluation metric defined by the DYAD challenge organizers, Table I presents the best result that we obtained on the testing set.
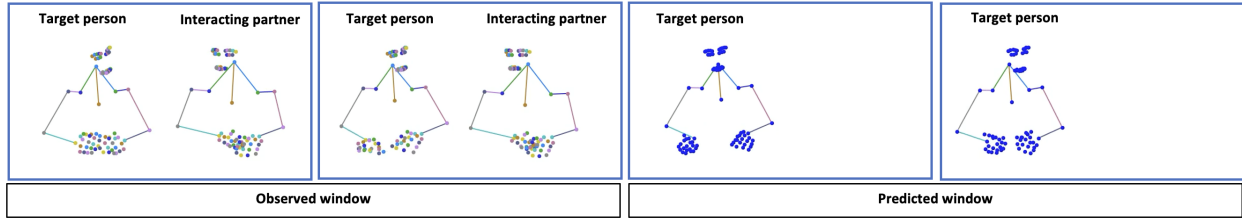
Fig. 2: The network receives the non-verbal data of the two interlocutors in a period of 4 seconds and predicts the action of the target person in the next 2 seconds.

TABLE I: The best results from Leaderboard (test phase) obtained by the proposed approach.

Avg. Rank	Face	Body	Hands
3.000	0.205	0.851	0.316

E. Final remarks

We observed that generated hand motions yielded lower performance than face and body. This problem could be explained by the randomness of hand gestures performed by the target person during social interaction. Such rhythmic movements are well-known as "beat gesture" [7] which are commonly used for stressing keywords or phrases of the communicator's speech. To better approach this motion, the audio features should be integrated into the proposed framework.

III. ADDITIONAL METHOD DETAILS

Please, reply if your challenge entry considered (or not) the following strategies and provide a brief explanation. For the non-binary questions, you can mark multiple options.

- **Mark with an X the modalities you have exploited.** ( ) Visual, ( ) Acoustic, ( ) Transcripts, ( ) Metadata, (X) Landmark annotations, ( ) Eye-gaze vectors.

- **In case you used metadata, mark with an X the types of metadata you have exploited.** ( ) Age, ( ) Gender, ( ) Country of origin, ( ) Max. level of education, ( ) Pre-session mood, ( ) Pre-session fatigue, ( ) Relationship among interactants, ( ) Task type, ( ) Task order, ( ) Task difficulty, ( ) Language, ( ) Other.
  If other, or if you have used just a subset of info for a given type of metadata (e.g., just a subset of mood values), please detail:

- **Mark with an X the tasks you used for training.** (X) Talk, ( ) Lego, ( ) Animals, ( ) Ghost.

- **Mark with an X the data representation type you used as input.** (X) Raw 2D coordinates, ( ) Raw 3D coordinates, ( ) Coordinate offsets (i.e. per-frame displacements), ( ) Velocity, ( ) Acceleration, ( ) Trajectories, ( ) Heatmaps, ( ) Other.
  Please detail: Annotation data represented by 2D co-

ordinates were pad into 150 frames. Each frame has a shape of (78,2) containing 28 face landmarks, 10 body landmarks, 20 left-hand landmarks, and 20 right-hand landmarks.

- **Mark with an X the data representation type you used as output.** (X) Raw 2D coordinates, ( ) Raw 3D coordinates, ( ) Coordinate offsets (i.e. per-frame displacements, ( ) Velocity, ( ) Acceleration, ( ) Trajectories, ( ) Heatmaps, ( ) Other.
  Please detail: The generated motion $P_{fo}^{0:k}$ has a shape of (50, 78, 2) including 50 motion frames. Each frame has a shape of (78,2) similar as the input.

- **Did you use information from previous sessions of the target interlocutor as a prior to model his/her behavior for a given (future) session?** ( ) Yes, (X) No
  If yes, please detail:

- **Did you use the minimum observable window of 4 seconds to predict future frames, or another approach?** (X) "observable window of 4 seconds", ( ) "another approach".
  If you used a different approach, please detail:

- **Did you use the provided validation set as part of your training set?** ( ) Yes, (X) No
  If yes, please detail:

- **Did you use any fusion strategy of modalities?** ( ) Yes, (X) No
  If yes, please detail:

- **Did you use the given/predicted personality labels?** ( ) Yes, (X) No
  If yes, please detail:

- **Did you treat the face, body and hands as different groups or did you predict all landmarks at once?** (X) "different groups", ( ) "all at once".

- **Did you use information from the other interlocutor (e.g., their visual info) to predict the future behavior of the target interlocutor?** (X) Yes, ( ) No.
  If yes, please detail: The motion data $P_{ob}^{0:k}$ of the other interlocutor was treated as a conditional input to forecast the action $P_{fo}^{k+1:T}$ of the target person.

- **Did you use pre-trained models?** ( ) Yes, (X) No

If yes, please detail:

- **Did you use any face / hand / body landmark detection, alignment or segmentation strategy, instead of or in addition to the landmark annotations provided by the dataset?** ( ) Yes, (X) No
  If yes, please detail:

- **Did you use external data?** ( ) Yes, (X) No
  If yes, please detail:

- **Did you use any regularization strategies/terms?** ( ) Yes, (X) No
  If yes, please detail:

- **Did you use handcrafted features?** ( ) Yes, (X) No
  If yes, please detail:

- **Did you exploit depth (i.e., Z component) information?** ( ) Yes, (X) No
  If yes, please detail:

- **Did you use any spatio-temporal feature extraction strategy?** ( ) Yes, (X) No
  If yes, please detail:

- **Did you use different weights for face, body and hands?** ( ) Yes, (X) No
  If yes, please detail:

- **Did you perform any data augmentation?**
  ( ) Yes, (X) No
  If yes, please detail:

- **Did you use any bias mitigation technique (e.g., rebalancing training data)?**
  ( ) Yes, (X) No
  If yes, please detail:

- **Did you use any input normalization technique (e.g., root-relative coordinates)?**
  (X) Yes, ( ) No
  If yes, please detail: The motion data were normalized/denormalized taking into account its mean and standard deviation values.

IV. CODE REPOSITORY

The codes used in this work are available at:
`https://github.com/TuyenNguyenTanViet/ForecastingNonverbalSignals`

V. ACKNOWLEDGMENT

REFERENCES

[1] C. L. Breazeal, *Designing sociable robots*. MIT press, 2002.
[2] M. Salem, F. Eyssel, K. Rohlfing, S. Kopp, and F. Joublin, "To err is human (-like): Effects of robot gesture on perceived anthropomorphism and likability," *International Journal of Social Robotics*, vol. 5, no. 3, pp. 313–323, 2013.
[3] F. Noroozi, D. Kaminska, C. Corneanu, T. Sapinski, S. Escalera, and G. Anbarjafari, "Survey on emotional body gesture recognition," *IEEE transactions on affective computing*, 2018.
[4] J. Martinez, M. J. Black, and J. Romero, "On human motion prediction using recurrent neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2891–2900.
[5] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.
[6] C. Palmero, J. Selva, S. Smeureanu, J. C. J. Junior, A. Clapés, A. Moseguí, Z. Zhang, D. Gallardo, G. Guilera, D. Leiva *et al.*, "Context-aware personality inference in dyadic scenarios: Introducing the udiva dataset." in *WACV (Workshops)*, 2021, pp. 1–12.
[7] D. McNeill, *Hand and mind*. De Gruyter Mouton, 2011.