A GAN-based Approach to Communicative Gesture Generation for Social Robots

Nguyen Tan Viet Tuyen, Armagan Elibol, and Nak Young Chong

***Abstract*— People use a wide range of non-verbal behaviors to signal their intentions in interpersonal relationships. Being echoed by the proven benefits and impact of people's social interaction skills, considerable attention has been paid to generating non-verbal cues for social robots. In particular, communicative gestures help social robots emphasize the thoughts in their speech, describing something or conveying their feelings using bodily movements. This paper introduces a generative framework for producing communicative gestures to better enforce the semantic contents that social robots express. The proposed model is inspired by the Conditional Generative Adversarial Network and built upon a convolutional neural network. The experimental results confirmed that a variety of motions could be generated for expressing input contexts. The framework can produce synthetic actions defined in a high number of upper body joints, allowing social robots to clearly express sophisticated contexts. Indeed, the fully implemented model shows better performance than the one without Action Encoder and Decoder. Finally, the generated motions were transformed into the target robot and combined with the robot's speech, with an expectation of gaining broad social acceptance.**

I. INTRODUCTION

People use a wide range of non-verbal channels, including facial expressions, body gestures, and similar others to signal their intention during human-human interaction. Those modalities help the communicators' messages transmit to interacting partners in a facile and transparent manner [1]. Being echoed by the influence of human social behaviors, considerable attention has been paid to generate non-verbal cues for social robots that are appealing and familiar to human interacting partners. In particular, communicative gestures endow the social robots with capabilities of emphasizing certain keywords in their speech, describing something, or conveying their intention. By adding communicative gestures to the robots' interactive behaviors, this strategy improves the user's perception of robots' speech and makes the social interaction outcomes enhanced [2].

The generation of communicative non-verbal behaviors for robots could be briefly categorized into two groups: rule-based approach and data-driven approach. Behavior Expression Animation Toolkit (BEAT) [3] is a well know approach that receives input text and releases non-verbal behaviors. In the BEAT toolkit, the connections between context input and behavior output are established based on a set of predefined rules. A similar approach can be found in [4], where the model of communicative gesture generation for a humanoid robot is introduced. Analyzing the input text, output patterns could be selected out from a list of manually designed robot gestures. Likewise, social robots have been equipped with the capability of performing communicative gestures for supporting their speech. Such gestures are manually designed by animation experts to ensure the familiarity and human-likeness of the motions. In contrast to the rule-based approach, where human labors are needed for modeling all possible contexts of interaction, the data-driven approach sidesteps such requirement by capturing the connections between non-verbal behaviors and corresponding neutral language context of various communication topics in an autonomous manner. In [5], the bidirectional mapping between gestures and natural language has been investigated. By feeding an input text description to the proposed framework, a synthetic action is released, and vice versa. However, the generated actions are defined in joint space of the Master Motion Map (MMM) model [6], it is difficult to transfer such generated gestures to other robots whose kinematics structures are different from the MMM framework. To overcome this problem, our generative framework produced output actions defined in $3D$ motion space, allowing them to be converted into various robot platforms. Recently, Generative Adversarial Network (GAN) [7] has been investigated for the generation of communicative gestures [8], [9]. Different from [8], in our approach, we consider the input text as conditional information for producing communicative gestures, this approach allows the contexts of the input are better expressed by generated body gestures. On the other hand, compared to [9], our designed framework is constructed by a convolutional neural network (CNN), which has been applied with great success in various domains including video generation [10], audio generation [11] and especially, image generation [12], [13].

In this paper, we aim at generating co-speech gestures for social robots to convey the semantic contents of their speech. The framework inspired by Conditional Generative Adversarial Network (CGAN) [14] built upon CNN with Action Encoder and Decoder, considering the input sentence as an essential condition for producing robots' communicative gestures. Thus, the connections between the input context and the output action are better addressed. The generated actions are defined in $3D$ space, allowing them to be easily implemented on various robotic platforms. Through the designed Transformation model, we demonstrate the synthetic gestures on the Pepper humanoid robot. Details of the framework are described in Section II. In Section III, the model was validated on two public datasets.

The authors are with the School of Information Science, Japan Advanced Institute of Science and Technology, Ishikawa, Japan {ngtvtuyen, aelibol, nakyoung}@jaist.ac.jp

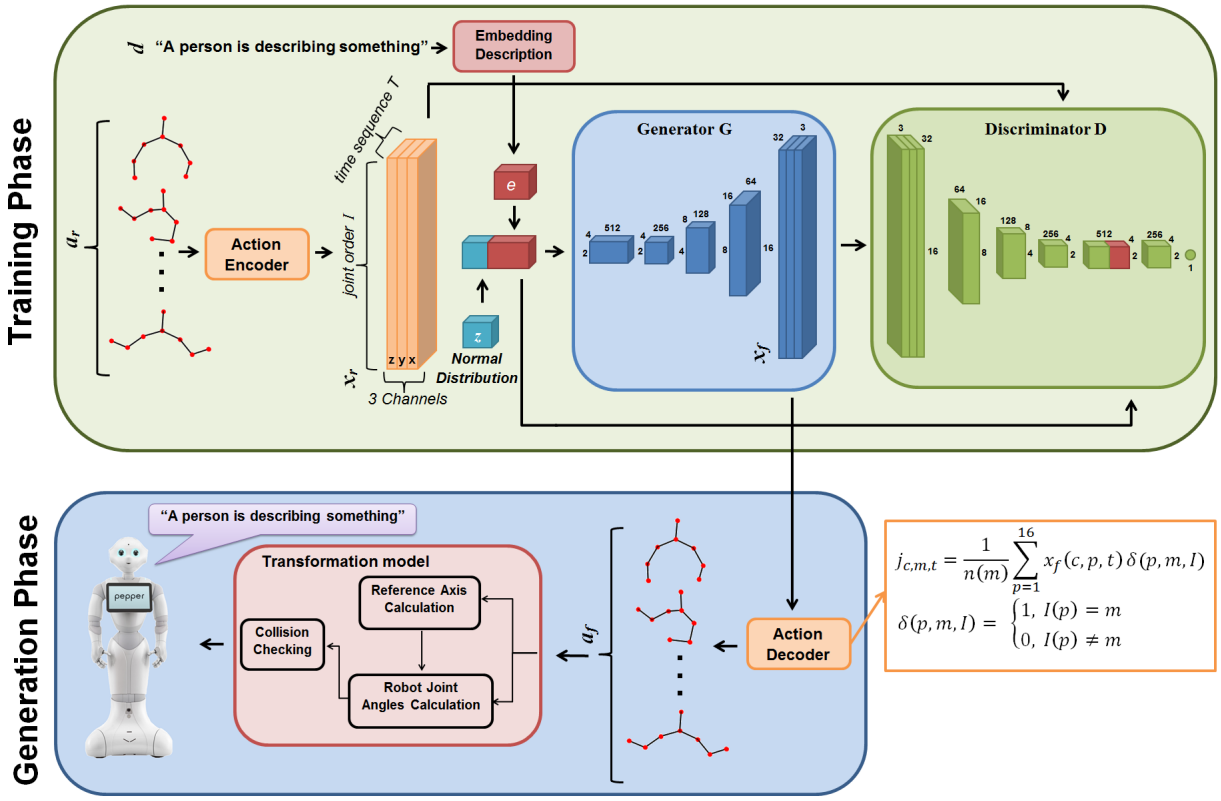


Fig. 1: The designed framework for producing action a_f synthesized with text d. Through the Transformation model, a_f is transformed into the Pepper robot's motion and associated with the robot's speech.

Here, we further extended the experiments conducted in our previous work [15] to better validate the performance of the designed framework in different aspects. Finally, the research conclusion and future works are explained in Section III.

II. METHODOLOGY

Fig. 1 illustrates the designed framework including the training and generation phase. $a_r = [S_1, S_2, S_3, ..., S_T]$ ($a_r \in \mathbb{R}^{3\times 8\times T}$) presents a real action consisting of T motion frames. $d = [w_1, w_2, w_3, ..., w_k]$ is a raw sentence synthesized with a_r. The training process starts by encoding a_r into x_r using Action Encoder. d is encoded into a fixed-length embedding vector e utilizing the encoder phase of the skip-thoughts model [16]. e is then concatenated with a noise vector z sampled from a normal distribution function. The concatenated vector is given to the Generator network for producing a fake action x_f conditioned to e. At the training phase, Generator aims to produce a fake action x_f as similar as x_r to fool the Discriminator whilst Discriminator attempt to distinguish between x_r and x_f taking heed of the condition e.

At the generation phase, an action x_f is generated by feeding the raw text input d into G. The generated action x_f is decoded to $a_f = [S'_1, S'_2, S'_3, ..., S'_T]$ defining human motion in 3D space using the Action Decoder. Through the Transformation model, a_f is transformed into the target robot motion, defined by a set of the robot's joint angles. The robot action is synchronized with the robot speech, being the robot communicative gesture as presented in Fig. 1. The following section will explain the Action Encoder, Action Decoder, Generator, and Discriminator model in detail.

A. Action Encoder and Decoder

Action Encoder: CNN-based approach has been widely applied with great success in action recognition tasks [17], [18], [19]. Inspiring from that, in our work [20], human actions are displayed as 2D matrices consisting of 3 channels representing for x, y, z. On each channel, the horizontal axis covers the temporal information of the action while the vertical axis represents a sequence of joints at a specific timestamp. However, the chain order of joints I on the vertical axis influences the spatial information represented in x_r. To better capture spatial information of the relative joints of the action a_r, our proposed Action Encoder [15] locates its adjacent joints near each other. With this idea, through the Action Encoder, a_r is encoded to x_r as illustrated in Fig. 1. On each channel $c \in \{x, y, z\}$ of x_r, the horizontal axis contains the motion sequence T, while the vertical axis is a series of joints denoted as I at a certain timestamp $t \in [1, T]$. Hence, rather than feeding the raw input a_r to D as applied in our previous work [20], using Action Encoder, the spatial-temporal features of a_r are better represented by x_r.

Action Decoder: At the generation phase, through the Action Decoder, generated action x_f is decoded to a_f as the formula presented Fig. 1. In the fake action a_f, $j_{c,m,t}$

denotes the joint index m, on the channel c, at the time stamp t. Noticed that $n(m)$ is the number of times the joint index m presented in the order I of x_f.

B. Generator and Discriminator Network

Generator: The effectiveness of transposed convolutional layers has been validated in a wide range of contexts such as image generation [12], [13], video generation [10], audio generation [11], and recently, motion generation [20], [8]. Inspiring from those previous works, this research investigates the convolution operation for generating robots' actions conditioned to their input speech. As the network architecture of G presented in Fig. 1, the combined vector between e and z is firstly passed through a fully connected layer. It is followed by four transposed convolutional layers for upsampling the data to the output target $x_f \leftarrow \mathrm{G}(z, e)$. On each convolutional layer, we applied batch normalization to stabilize the learning process, which helps to reduce the number of training epochs required. It is followed by Rectified Linear Unit (ReLU) activation [21] except for the last layer, where the *tanh* activation is implemented before producing x_f.

Discriminator: D is designed by five convolutional layers. Batch normalization and ReLU are implemented for all layers except the output one, where the sigmoid function is used. Discriminator D receives either real action a_r or fake action a_f as an input. D also takes into account the information of e, this is done by concatenating e with the output value of the fourth layer. At the last layer, D produces an output probability representing the realistic of the input action.

D is trained to distinguish between x_r and x_f, this is done by training D to maximize the output probability $y_r \leftarrow \mathrm{D}(x_r, e)$ when the real action x_r synthesis with e is injected to the network. Vice versa, by feeding x_f and e to the network, D is trained to minimize $y_f \leftarrow \mathrm{D}(x_f, e)$. In the training set, the miss-matching description $\hat{d}$ is also collected. $\hat{d}$ is considered as a sentence incorrectly annotates the motion x_r. When a pair of real motion x_r and miss-matching embedding $\hat{e}$ is given to Discriminator, D is targeted to minimize $y_m \leftarrow \mathrm{D}(x_r, \hat{e})$, implying that x_r does not appropriately synthesize with $\hat{d}$. Overall, the miss-classification error L_D of D network is summarized in Eq. 1. Concerning the training strategy for the Generator network, G is targeted to produce the fake action x_f synthesized with e as much realistic as possible to fool D, it is done by training the network to maximize the output probability y_f. The miss-classification error L_G of G network is described in Eq. 2. The binary cross-entropy is applied to compute the error of both L_D and L_G. During the training process, the parameters of D are firstly updated while fixing the parameters of G constant. Then, G network is updated while remaining parameters of D unchanged.

$$L_D = log(y_r) + log(1 - y_m) + log(1 - y_f); \quad (1)$$

$$L_G = log(y_f); \quad (2)$$

III. EXPERIMENTAL RESULTS AND DISCUSSION

A. Experimental Setup

The designed framework was firstly validated on the MSR-VTT dataset [22] as similarly conducted in [9]. The dataset includes $2,822$ actions $a_r \in \mathbb{R}^{3\times8\times32}$ defined human upper body motion in 3D space and $31,863$ corresponding description d to convey that motion (one action is associated with more than one text description). For encoding d into an embedding vector e as presented in Fig. 1, we utilized the encoder phase of skip-thought model trained on BookCorpus dataset [23]. Additionally, a description $\hat{d}$ is randomly picked up from the dataset. $\hat{d}$ is defined as a miss-matching description if the dissimilarity between e and $\hat{e}$ is equal or higher than our predefined threshold. Totally, $29,663$ pairs of actions a_r, text descriptions d, and miss-matching descriptions $\hat{d}$ were obtained. We split the dataset into 90% for training and 10% for testing.

During the training process, real actions a_r, matching texts d, and miss-matching ones $\hat{d}$ were fed into the training framework with a batch size of 100. We applied the Adam optimizer [24] for both G and D network at the learning rate $\alpha = 2 \times 10^{-5}$, and they were trained for 700 epochs. Once the training process is completed, an action x_f could be generated by feeding a raw text d and a noise vector z to the G network. Through Action Decoder, x_f is decoded to a_f defining human motion in 3D space.

B. Evaluation Metrics

Consider that $a_r = [S_1, S_2, S_3, ..., S_T]$ is the real action synthesized with the text d. On the other hand, the fake action $a_f = [S'_1, S'_2, S'_3, ..., S'_T]$ is produced by feeding the description d to the Generator network. Since of both of a_r and a_f are associated with a same annotation d, thus, it is reasonable for measuring the similarity between a_r and a_f in order to verify the synthesis between a_r and d. The evaluation is started by encoding both of a_r and a_f into feature descriptors C_r and C_f using covariance description with temporal hierarchical construction [25] as illustrated in Eq. 3. This approach ensures the spatial and temporal features of action is well presented by a fixed-length descriptor. Noticed that $\overline{S}$ is the sample mean of S_i computed over the time T while $^\intercal$ stands for the transpose operator. Finally, we measure the similarity between C_r and C_f using cosine similarity as given in Eq. 4.

$$C = \frac{1}{t-1}\sum_{i=1}^{T}(S_i - \overline{S})(S_i - \overline{S})^\intercal \quad (3)$$

$$Similarity(C_r, C_f) = \frac{C_r \cdot C_f}{||C_r||\,||C_f||} \quad (4)$$

C. Variety of Generated Actions to Express an Input Context

We firstly examined how different generated actions are when altering the raw text input while retaining the same context. Fig 2 shows a sequence of skeleton frames of the generated action a_f synthesized with the description d *"a young woman demonstrates example of lifting exercises."*,

Fig. 2: "a young woman demonstrates example of lifting exercises."

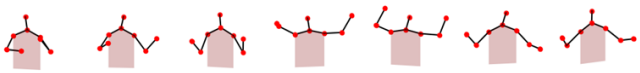


Fig. 3: *"a girl practices lifting exercise at the gym."*

Fig. 4: *"a woman performs weight lifting exercises."*

Fig. 5: *"I was practicing lifting exercises at the gym."*

which belongs to the testing set. Then, three modified versions of d were used as *"a girl practices lifting exercise at the gym."*, *"a woman performs weight lifting exercises."*, and *"I was practicing lifting exercises at the gym."*. Figs. 3, 4 and 5 present the generated motions synthesized with the aforementioned modified annotations. The resulting action in Fig. 2 looks like a person is lifting something by pushing their two arms up and down several times. Similarly, the motions displayed in Figs. 3, 4, and 5 seem to express a same physical meaning of "lifting exercise" although skeleton frames of those motions are not exactly matched to each other at a specific timestamp.

In the second example illustrated in Fig. 6, we examined the variety of generated actions by feeding the same raw text and different random noise vectors to the Generator network. As presented in Fig. 6, the text input *"one girl is dancing to music"*, which belongs to the testing set, was given to the Generator network with three different noise vectors z_1, z_2, and z_3. A closer look at the three generated motions, it can be seen that they are not exactly similar to each other at specific timestamps. However, those bodily expressions could be perceived as someone is performing exaggerated movements of two hands while dancing. Overall, the results demonstrated in the two examples suggest that our designed generative framework does not simply memorize and reproduce the actions learned from the training phase. Instead, G is able to generate a diverse set of actions expressing a certain input context. For social robots, this property would allow them to generate novel body gestures overtime to convey a certain context of their speech rather than performing stereotyped action patterns. It is widely known that the diversity of robots' non-verbal behaviors is considered as a key role for maintaining user engagement in social human-robot interaction [26].

D. Generation of "high resolution" Actions

The designed framework was also validated on KIT dataset [27], a higher dimensional dataset compared to the MSR-VTT dataset that we used in the previous section.

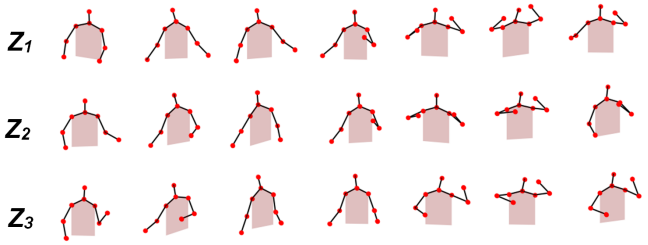


Fig. 6: Generated actions for "one girl is dancing to music," produced from noise vector z_1, z_2 and z_3, respectively.

The KIT dataset allows for assessing the performance of the network for generating human motions displayed in a higher "resolution". In total, $51,360$ pairs of real actions $a_r \in \mathbb{R}^{3\times 20\times 240}$ and matching descriptions d were obtained. Similar to the previous experiment, we also used the encoder phase of the skip-thought model trained on BookCorpus dataset [23] as Embedding Description and collected miss-matching descriptions $\hat{d}$. The dataset was divided into 90 % for training and 10% for testing. The framework was trained for $1,200$ epochs using Adam optimizer [24] at the learning rate of $\alpha = 2 \times 10^{-5}$ for both Generator and Discriminator.

At the generation phase, by feeding the following sentence *"A person waves with both hands"*, which is included in the testing set, to the G network, the resulting action (full model) and the corresponding ground truth one (GT) are illustrated in Fig. 7. Firstly, it is clear that generated action well expresses the content of input speech by performing waving motions with two hands. Indeed, the generated motion is similar to the real one over the time sequence, although the corresponding poses at a certain timestamp are not exactly matched to each other. It is noticed that by producing the fake human upper body actions defined in a higher number of joints, sophisticated input contexts can be expressed in a transparent manner. Fig. 8 shows the generated motion synthesized with the description *"A person makes motion as if playing violin"*. In addition to the movements of two hands for playing the violin, that bodily expression is further strengthened by equipping with head movements that the action looks like someone turning their head to the left for holding the violin on their shoulder.

Fig. 9 presents the tSNE projection [28] of a_f on the 2 dimensional space. Here, each plot represents a generated action a_f synthesized with description d, which is included in the testing set. This projection allows to qualitatively visualize similarities among generated motions taking into account the synthesis context on a low dimensional space. By taking into consideration the keywords of the raw sentences, generated actions were categorized into different groups according to the type of motion performed, such as waving, dancing, bowing, walking, and similar others. On the other hand, due to the large variety of motion types available in the testing set, unclassified motions were label as others and colored in gray. The visualization shown in Fig. 9 indicates that generated motions synthesized with similar text descriptions are located near to each other. In particular, motions related to waving hands, bowing, or dancing have

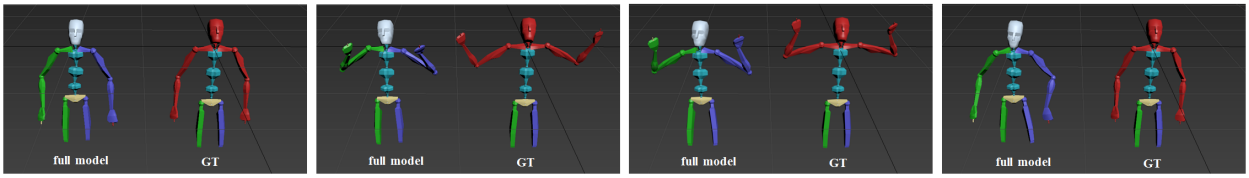


Fig. 7: *"A person waves with both hands"*

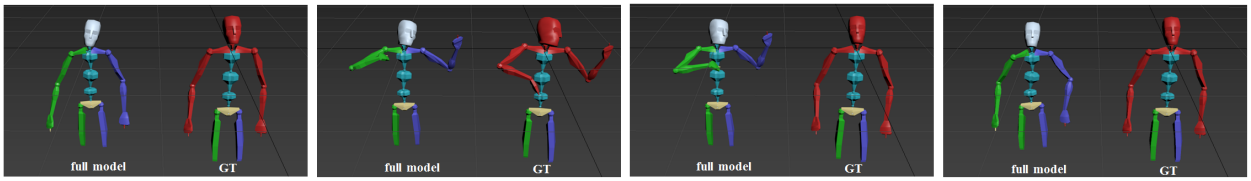


Fig. 8: *"A person makes motion as if playing violin"*

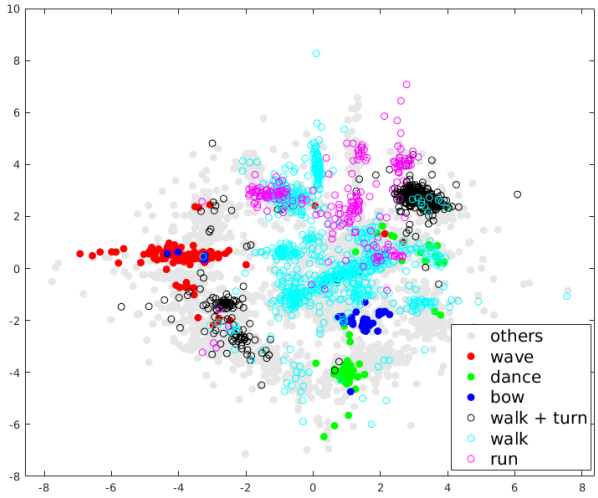


Fig. 9: 2-dimensional tSNE projection of generated action a_f, colored by their motion types.

TABLE I: Comparison between the designed framework without Action Encoder/Decoder (w/o E/D) and the fully implemented model (full model). The evaluation was conducted on the MSR-VTT and the KIT dataset.

	w/o E/D	full model
MSR-VTT dataset	0.5060	0.5287
KIT dataset	0.6364	0.6603

denser clusters with less variation than locomotion actions such as walking or running. It should be highlighted that our approach concentrates on the generation of upper body motions, thus, the lack of lower body joints highly affected the quality of generated locomotion actions. In other words, bodily expressions of locomotion actions such as walking or running are less transparent compared to the upper body movements such as waving hands, dancing or bowing.

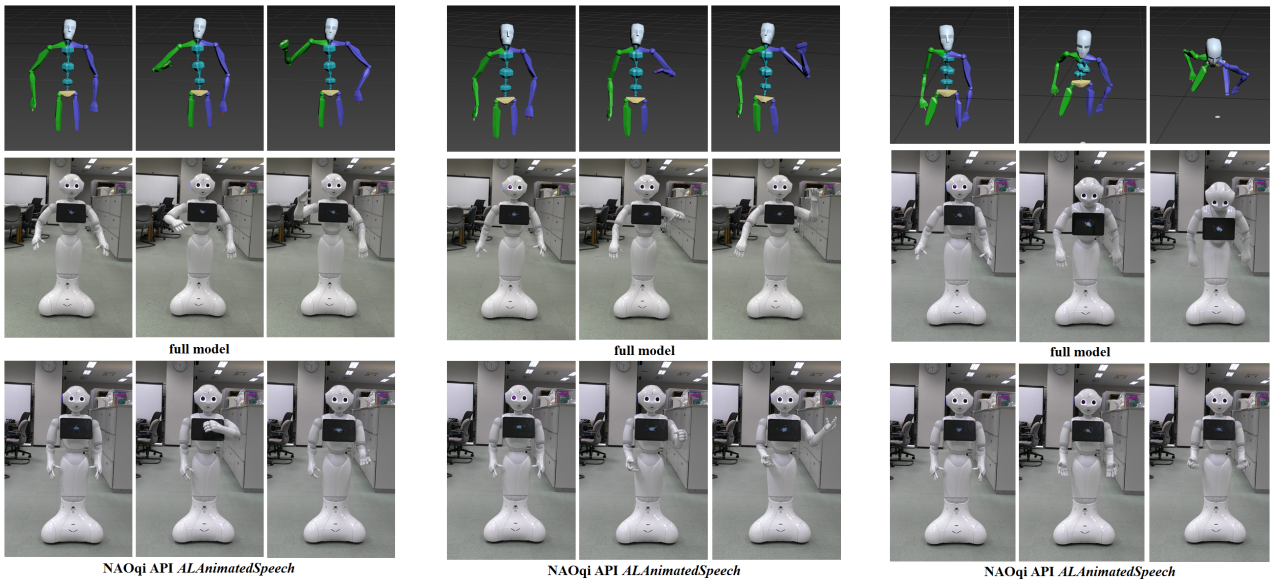
E. Quantitative Evaluation of Generated Actions

For quantitative evaluation of generated actions, from the testing set of the MSR-VTT dataset, d was fed to the G network for generating a_f. Additionally, we also examined the proposed framework illustrated in Fig. 1 without Action Encoder and Decoder. Particularly, the raw action a_r was given to the training phase without passing through Action Encoder. At the generation phase, a_f could be produced from G without using Action Decoder. By applying the evaluation metric discussed in III-B, we obtained the average similarity between ground truth actions a_r and the actions a_f, which are produced from the simplified framework without Action Encoder/Decoder and the fully implemented model. The evaluation was also carried with the testing set of the KIT dataset introduced in III-D. Finally, the experimental results are summarized in Table. I.

The results presented in Table I reveals that the fully implemented model exhibits better performance than the framework without Action Encoder and Decoder. The experiment underlined that by adopting the simplified version without Action Encoder and Decoder for the training phase, the training process is sped up. The main reason is that Action Encoder encodes a_r into x_r, which is a higher dimension matrix. However, with the fully implemented model, by furnishing the generative framework with Action Encoder, relative joints of the human upper body are distributed near each other on the vertical axis while the time sequence of the motion is captured by the horizontal axis. This representation allows for the spatial and temporal information of the action a_r displaying better. As the result, Discriminator can detect the action features faster and more efficient. Then, D could provide more informative feedback to the G network for optimizing the generated actions.

F. Transferring generated actions into the target robot

Through the transformation model illustrated in Fig. 1, a_f synthesized with d was transformed into the Pepper robot motion space. Additionally, the robot off-the-shelf module *ALTextToSpeech* was integrated into the action generation phase as this function enables the robot to utter the text d while performing bodily expression to support for its speech. In order to see the differences on the robot co-speech gestures produced from our approach and the ones generated from the robot on-board module, the same context d was given to the robot NAOqi API *ALAnimatedSpeech*. The results of comparison is demonstrated in Fig.10. It can be noticed that the following sentences *"I am performing a waving motion*



(a) *"I am performing a waving motion with the right hand"* (b) *"Someone over there is waving with the left hand"* (c) *"Someone is performing an upper body bow"*

Fig. 10: Differences between gestures from the proposed approach and the robot off-the-shelf module.

with the right hand", *"Someone over there is waving with the left hand"*, and *"Someone is performing an upper body bow"* were edited from the original descriptions, which are included in the KIT testing data, while keeping the messages of "waving right hand", "waving left hand", and "bowing" intact.

It can be seen that gestures produced from the robot on-board module *ALAnimatedSpeech* are mostly not related to the semantic content of the robot's speech. The robot gestures in Figs. 10a and 10c could be interpreted in such way that a person is explaining something with slight movements of their hands. To some extent, the robot action shown in Fig. 10b could be understandable as "waving left hand". It should be noticed that NAOqi API *ALAnimatedSpeech* consists of a set of robot gestures handcrafted by animation experts. When feeding an text input, a random motion could be selected out from the predefined list if certain keywords are not detected from the given sentence. This approach may limit the robot ability for performing bodily expressions to convey the semantic contents of its speech. In contrast to the robot on-board module, the proposed framework utilizes the embedding vectors capturing the semantic and syntax of raw sentence inputs for producing robot co-speech gestures. As the result, the robot gestures produced from our approach are more appropriately fitted to the contents of the robot' speech. It can be seen in Figs 10a and 10b that the robot is able to perform suitable waving hand movements with an upright posture. On the other hand, as displayed in Fig. 10c, the robot collapsed their upper body downward while uttering the context input. However, the limitation of the robot's physical configuration constrains the range of their bending motion, falling to reach the extent as performed by the generated human-like action.

IV. CONCLUSION AND FUTURE WORKS

This work presented an approach to communicative gesture generation. The generative framework inspired by CGAN built upon CNN with Action Encoder and Decoder. The model receives an input text and releases a synthetic action expressing the meaning of input context. The framework was firstly validated on MSR-VTT - a low dimensional co-speech action dataset as similar as conducted in our previous work [15]. Additionally, the experiment was carried out on the KIT dataset to confirm the network's capability of generating "high resolution" actions. Overall, the experimental results suggested that a variety of actions could be generated to express an input context. For social robots, this ability would allow them to perform various communicative gestures supporting long-term human-robot interaction. Indeed, the generative framework is able to produce synthetic actions defined in a high number of joints, which makes sophisticated input contexts possible to be clearly expressed. The comparative results indicated that the fully implemented model yields better performance than the one without Action Encoder and Decoder. Finally, the generated actions were converted into the target robot's motion and combined with the robot's speech. Compared to the robot's off-the-shelf module, it was shown that communicative gestures produced by our approach are better connected to the semantic contents of the robot's speech, although sometimes the limitation of the robot's physical configuration affected the robot's bodily expressions.

Non-verbal behaviors are complex and cover various communication topics. Rather than establishing a large set of rules for modeling all possible interaction contexts, we urge that the proposed approach could be used for capturing the connection between human actions and corresponding

neutral language contexts in an efficient manner. This idea endows robots with an ability to perform communicative gestures which are more acceptable to human interacting partners. In our future work, the current framework will be extended for editing styles of robots' communicative gestures.

ACKNOWLEDGMENT

The authors are grateful for financial support from the Air Force Office of Scientific Research under AFOSRAOARD/FA2386-19-1-4015 and the Shibuya Science, Culture, and Sports Foundation 2019 Grant Program.

REFERENCES

[1] M. L. Knapp, J. A. Hall, and T. G. Horgan, *Nonverbal communication in human interaction*. Cengage Learning, 2013.

[2] S. Saunderson and G. Nejat, "How robots influence humans: A survey of nonverbal communication in social human–robot interaction," *Int'l Journal of Social Robotics*, vol. 11, no. 4, pp. 575–608, 2019.

[3] J. Cassell, H. H. Vilhjálmsson, and T. Bickmore, "Beat: the behavior expression animation toolkit," in *Life-Like Characters*. Springer, 2004, pp. 163–185.

[4] V. Ng-Thow-Hing, P. Luo, and S. Okita, "Synchronized gesture and speech production for humanoid robots," in *IEEE/RSJ Int'l Conf. on Intelligent Robots and Systems*, 2010, pp. 4617–4624.

[5] M. Plappert, C. Mandery, and T. Asfour, "Learning a bidirectional mapping between human whole-body motion and natural language using deep recurrent neural networks," *Robotics and Autonomous Systems*, vol. 109, pp. 13–26, 2018.

[6] Ö. Terlemez, S. Ulbrich, C. Mandery, M. Do, N. Vahrenkamp, and T. Asfour, "Master motor map (mmm)—framework and toolkit for capturing, representing, and reproducing human motion on humanoid robots," in *IEEE-RAS Int'l Conf. on Humanoid Robots*, 2014, pp. 894–901.

[7] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, 2014, pp. 2672–2680.

[8] Y. Nishimura, Y. Nakamura, and H. Ishiguro, "Human interaction behavior modeling using generative adversarial networks," *Neural Networks*, vol. 132, pp. 521–531, 2020.

[9] H. Ahn, T. Ha, Y. Choi, H. Yoo, and S. Oh, "Text2action: Generative adversarial synthesis from language to action," in *IEEE Int'l Conf. on Robotics and Automation*, 2018, pp. 5915–5920.

[10] C. Vondrick, H. Pirsiavash, and A. Torralba, "Generating videos with scene dynamics," in *Advances In Neural Information Processing Systems*, 2016, pp. 613–621.

[11] C. Donahue, J. McAuley, and M. Puckette, "Adversarial audio synthesis," *arXiv preprint arXiv:1802.04208*, 2018.

[12] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, 2015.

[13] Z. He, W. Zuo, M. Kan, S. Shan, and X. Chen, "Attgan: Facial attribute editing by only changing what you want," *IEEE Transactions on Image Processing*, vol. 28, no. 11, pp. 5464–5478, 2019.

[14] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.

[15] N. T. V. Tuyen, A. Elibol, and N. Y. Chong, "Conditional generative adversarial network for generating communicative robot gestures," in *IEEE Int'l Conf. on Robot and Human Interactive Communication*, 2020, pp. 201–207.

[16] R. Kiros, Y. Zhu, R. R. Salakhutdinov, R. Zemel, R. Urtasun, A. Torralba, and S. Fidler, "Skip-thought vectors," in *Advances in Neural Information Processing Systems*, 2015, pp. 3294–3302.

[17] Q. Ke, M. Bennamoun, S. An, F. Sohel, and F. Boussaid, "A new representation of skeleton sequences for 3d action recognition," in *IEEE Conf. on Computer Vision and Pattern Recognition*, 2017, pp. 3288–3297.

[18] Z. Yang, Y. Li, J. Yang, and J. Luo, "Action recognition with visual attention on skeleton images," in *IEEE Int'l Conf. on Pattern Recognition*, 2018, pp. 3309–3314.

[19] C. Caetano, F. Brémond, and W. R. Schwartz, "Skeleton image representation for 3d action recognition based on tree structure and reference joints," in *IEEE Conf. on Graphics, Patterns and Images*, 2019, pp. 16–23.

[20] N. T. V. Tuyen, A. Elibol, and N. Y. Chong, "Learning from humans to generate communicative gestures for social robots," in *Int'l Conf. on Ubiquitous Robots*, 2020, pp. 284–289.

[21] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Int'l Conf. on Machine Learning*, 2010, pp. 807–814.

[22] J. Xu, T. Mei, T. Yao, and Y. Rui, "Msr-vtt: A large video description dataset for bridging video and language," in *IEEE Conf. on Computer Vision and Pattern Recognition*, 2016, pp. 5288–5296.

[23] Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler, "Aligning books and movies: Towards story-like visual explanations by watching movies and reading books," in *IEEE Int'l Conf. on Computer Vision*, 2015, pp. 19–27.

[24] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[25] M. E. Hussein, M. Torki, M. A. Gowayyed, and M. El-Saban, "Human action recognition using a temporal hierarchy of covariance descriptors on 3d joint locations," in *Int'l Joint Conf. on Artificial Intelligence*, vol. 13, 2013, pp. 2466–2472.

[26] I. Leite, C. Martinho, and A. Paiva, "Social robots for long-term interaction: a survey," *Int'l Journal of Social Robotics*, vol. 5, no. 2, pp. 291–308, 2013.

[27] M. Plappert, C. Mandery, and T. Asfour, "The kit motion-language dataset," *Big data*, vol. 4, no. 4, pp. 236–252, 2016.

[28] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of Machine Learning Research*, vol. 9, no. 11, 2008.