Automatic Generation of Eye Expressions with End-to-End Learning

Ung Park, Eui Jun Hwang*, JongSuk Choi*

University of Science and Technology
Korea Institute of Science and Technology
{wpark, ktye2220, cjs}@kist.re.kr

***Abstract*— The generation of eye expressions in robots is very important to enhance their ability to find appropriate expressions required for a situation among the major social facilities. In particular, it is the main research subject to be addressed to achieve a positive human–robot interaction. We used a learning-based facial expression generation method utilizing human facial expression data, rather than a conventional heuristic method where all the pre-rules should be defined by experts. We designed a neural network model based on an end-to-end learning model, which showed effective results for learning gestures from given image data. Our proposed model consists of an encoder for processing speech and a decoder for generating facial expressions. The model successfully generates eye expressions corresponding to input sentences. Through this, we propose a new approach for enhancing human–robot interaction and studying the robot's eye expressions.**

I. INTRODUCTION

Since the early days of human–robot interaction research, it has been suggested that the ability to form and maintain social relationships based on reciprocity and reactivity is an important asset that social robots must possess. [27] Similarly, research on the expressions of the robot's eyes has received relatively minimal attention, although it has been argued that the ability to combine non-verbal signals (such as the speaker's gaze) with other signals (such as language) is important in human–robot interaction. [10] Only recently, it has been discovered that proper eye contact and facial expressions of the robot toward the user contribute to the formation of a positive rapport between human and robot. As evidence suggests, the robot recognizing human's non-verbal cues and expressing the internal state of itself affects the development of a positive relationship between humans and robots. [13] Therefore, robot eye expression in human–robot interaction is important.

Accordingly, social robot researchers have advanced various mechanisms to more accurately implement robot facial expressions, including eye expressions. Kismet [4] demonstrated an advance design with a variety of facial movements, and Eddie [21] demonstrated a facial movement similar to Kismet while adding a mechanism to mimic the facial expressions of human observers. In addition, mood expressions have been implemented using facial expressions in virtual agents, such as Valerie [9]. These studies have significant contribution to improving the expressive functions of robots. Recently, robots emulating more sophisticated emotions, such as SEER, have been introduced. However, many robots still use heuristic methods designed by human experts. Heuristic expression can be expressed naturally in a similar manner to a human, but there is a limitation as only a predefined expression can be generated. Moreover, a considerable artificial work is required to establish a connection between sentences and expressions. This paper presents a model for generating natural eye expressions of robots through data-based learning to minimize the manpower required to generate robot expressions and to allow robots to generate eye expressions related to their internal state as natural as possible.

As a result, we propose the Speech Eye Motion Data (SEMD), a dataset that can be used for these facial expressions studies. We have also attempted to contribute to human–robot interaction research by proposing a generator that can automatically generate various types of facial expressions related to the internal state of the robot without any rules predefined by humans.

II. RESEARCH TRENDS

Research on automatic motion generation for speech or sentences can be broadly divided into three categories: rule-based systems, statistical modeling methods, and machine learning approaches. [8] The rule-based system has been widely applied not only to the expression of the robot but also to motion generation. In this system, the developer can design semantic contents considering the context of the situation and the dialogue. However, this system has a limitation as it requires a considerable amount of manpower to formulate it as if–then rules for all situations. In addition, methods that rely on probabilistic modeling of annotated features using semantic tags for a given speech have been proposed, but they share a common limitation, that is, expert subjective judgment and manpower are required.

To overcome this limitation, a recent deep neural network (DNN) approach was used to learn the relationship between a

*: These authors contributed equally.

[1] https://github.com/WoongDemianPark/Speech-Eye-Motion-Dataset

speaker's speech and head movement from news data and attempted to map speech to head motion. [7] In addition, there have been attempts to learn the relationship between gestures obtained by attaching a motion sensor to an actor's body and speech. [8] These studies were similar to ours because they used repetitive neural networks to generate human movements from text. However, to the best of our knowledge, our paper is the first paper on generating motion coordinates to generate robot eye expressions from a given text to express a robot's internal state.

III. Theoretical Background

A. Definition of Eye Expressions Range

The fundamental task of recognizing or expressing facial expressions is to define the meaning of human facial expressions and the range of each component, with certain facial features involved in facial identification. Thus, the cognitive science community has attempted to define the classification, components, and scope by attempting a semantic language analysis of facial expressions; there are some studies involving eyebrows in the range of eye expressions. [25] In addition, it has been proved through experiments [18] that the combination of these eyebrow traits, not just the eyeball, not only determines the final facial expression but also affects the human facial recognition mechanism. Moreover, it has also been proven that eyebrows affect emotion perception through facial expressions. [20] Accordingly, in this study, the area including the size of the vertical eye aperture, distance between the eyebrows, and slope of the eyebrows as well as the direction of the gaze were defined as a range of eye expressions.

B. Sequence-to-Sequence Model

Representative examples of technology-oriented research for expressions include expression generation based on photographic data and studies on how to analyze emotions in a continuous video sequence. However, because most of these research methods are focused on the problem of recognizing human expressions rather than robot expressions or handling static data, such as photographs, applying them to the study of generating expressions of robots considering context is not appropriate. The initial method used to solve this problem was to define and use a set of prototypical expressions in advance [2]. However, as described above, this heuristic method has limitations in defining expression components for all behaviors and changes in situations. Furthermore, even if a number of components are defined, combining them to generate appropriate expressions for situations also has the same limitations.

Because expression in human–human and human–robot interaction is a dynamic region rather than a static region, a method for generating an expression considering a sequence in a video region rather than an image is required. In particular, on an interactive interface, such as a robot with animated avatars or physical shapes, facial expression generation considering a given context is required. To meet these requirements, we defined facial expressions as a sequential problem that depend on the order of speech. At this time, speech is a sequence that can be divided into a series of letters or words, whereas video is a sequence that can be divided into consecutive images. Hence, we expected that facial expressions could be predicted if the generator learns the relationship between speech and facial expressions from video including speech.

By applying this method, we propose a model capable of generating eye expressions related to a given robot's internal state without prior expertise or manpower for heuristic design. We used speech to consider context when generating eye expressions, and in this case, audio was not used because it could not be implemented for existing text-to-speech (TTS) functions used in robots.

C. Data Collection: English Education Dataset

Co-speech facial expressions are found in almost all video footage where humans appear. Various movies, dramas, and TV series as well as everyday videos through social network services include facial expressions along with speech. Nevertheless, data on how to retain the eyes on the front during a speech and track accurate eye movements are very limited. We considered the application of an open dataset, such as RAVDESS [12], Oulu-CASIA [29], or TV Human Interaction Dataset [15]. However, they we not sufficient to track eye movements with various voice contents providing the frontal expression of the speaker, and the datasets have long playing time.

For this reason, we developed the SEMD by collecting new co-speech facial expression data from YouTube. We used the English education content registered on YouTube, and these data have the following advantages compared to existing datasets:

(1) It is possible to collect a large amount of video data, including various topics and contents on YouTube.

(2) Most speakers are staring at the camera, and tracking eye movements, including facial expressions, is relatively easy.

(3) Owing to the nature of educational material, the contents are transmitted using accurate pronunciation and expressions.

(4) It is advantageous for data collection automation because images are provided on the same platform.

(5) It is advantageous for annotation automation through YouTube subtitles.

We collected a total of 763 videos and subtitles from YouTube's English education channel. We extracted facial expression feature points from the data collected using dlib and OpenPose [6] and filtered images that did not contain the feature points to be extracted. The exclusion conditions were as follows:

(1) Image without movement.

(2) Very short video.

(3) Images without pupil visible.

(4) Video featuring two or more people.

(5) An image with a face that is too small to hold a feature point.

Under these conditions, we removed as many images that could negatively affect the learning process, and we obtained 33,195 pairs of sequences and eye motions. We divided the pairs into approximately 80% for the training set, 10% for the validation set, and 10% as test partitions. Moreover, the dataset was not large enough, so we conducted k-fold cross-validation to verify our model.

IV. Proposed Model

A. Data Pre-processing and Semantic Analysis

In this study, the feature points extracted through dilib were represented by 1 for eye pupil, 6 for eye shapes including eyelids, and 5 for eyebrows. The extracted skeletons were normalized so that the middle of the forehead was located at the center of the screen. These facial expression data were converted into a seven-dimensional vector using principal component analysis (PCA), and seven components were identified. We can observe meaningful expressions in the PCA space, for example, the first component is the rotational movement around the facial y-axis, the second component is the rotational movement around the z-axis, the third and fourth components show eyebrow movement, the fifth component is related to eye opening and closing, and the sixth and seventh components are related to pupil movement. We excluded the first and second components because they were not related to facial expressions.

B. Model Structure (Seq2Eye)

We used a sequence-to-sequence architecture [22] consisting of an encoder and decoder, inspired by the neural machine translator (NMT). However, because it is difficult to map word sequences to eye movement sequences using the existing model, we modified parts of it. In the modified model, the encoder uses a bidirectional RNN and processes input word by word. In addition, the decoder uses the encoder output to predict a sequence of motions y, and it is trained to minimize the negative log-likelihood of the target sequence y_hat. As a result, the decoder generates eye motions using Luong attention, a post layer, and a residual connection.

Both encoder and decoder use 2 × LSTM with 200 hidden layers. Our model generated a fixed number of motions. Sliced input sequences were processed to prevent gradient vanishing or exploding problems, and the decoder network generated 20 motions using 10 forwarded eye motion sequences.

C. Training details

We used mean squared error (MSE) as a loss function, and 33,195 pairs of sequences of words and eye motion from YouTube were used for the training. We used the Adam optimizer with a learning rate of 0.0001. The batch size was 512, and gradients were clipped to 2.0 to prevent gradient exploding. A dropout rate of 0.1 was applied to the encoder and decoder, and the train iteration was 1400 epochs. The training took approximately 15 h with an NVIDIA GTX 1080 ti GPU. In addition, we used k-fold cross-validation to prevent data bias, underfitting, and improve accuracy.

V. Experiment

To classify sentences used as input, generate expressions, and infer the relationship, specific criteria for the classification were required. Accordingly, we randomly selected sentences corresponding to eight emotions from the annotated corpora datasets for Emotion Classification in Text of Boston and Klinger [3] to classify sentences to be used as input, and 20 sentences were tested for each emotion classification (a total of 160 sentences). At this time, the generator generates facial expressions corresponding to the text sequence based on the results learned without the emotion labeling process, and expressions such as language and facial expressions are dependent on the social context [14]. Therefore, analyzing the exact correlation between these inputs and the generated facial expressions requires in-depth research from a more diverse perspective. However, language is a means to grasp situation and context and express emotions more precisely than other means of expression. Therefore, Boston's study, which classified emotional words, suggested an appropriate criterion for classifying the type of sentences to be input into the generator and confirming the results.

VI. Result

For the analysis of the generated facial expressions, we observed facial expressions generated based on the following five eye features, including gaze direction, as Table 1.

VII. Discussion

We found that the trained network generated different facial expressions for basic emotions, such as joy, trust, fear, anticipation, sadness, disgust, and anger, for voice inputs, but not in the training set. In particular, in the case of joy, significant generated facial expressions with universal facts were observed, such as a phenomenon in which vertical eye aperture decreases and eyebrow curvature increases due to eye laughter. These results show that the network successfully captures the differences in sentences and can generate various facial expressions corresponding to them. However, it is a challenge to evaluate how organically the generated expressions connect with spoken text.

VIII. Conclusion

We proposed a Seq2Seq model that can generate facial expressions corresponding to a given text. This model trained a given YouTube English education dataset without predefined contents for facial expression generation. The result of generating an appropriate expression for the given text showed the possibility of automatically generating the expression of the robot that matches the text without a heuristic method devised by an expert. In future experiments, we intend to conduct a subjective assessment of multiple subjects to measure the correlation between text and facial expressions generated by this model.

References

[1] Ahn, H., Lee, D.-W., Choi, D., Lee, D., Hur, M., Lee, H., & Shon, W. (2011). Development of an android for singing with facial expression. IECON 2011 - 37th Annual Conference of the IEEE Industrial Electronics Society, (pp. 104–109). Melbourne, VIC.

[2] Blanz, V., & Vetter, T. (1999). A morphable model for the synthesis of 3D faces. SIGGRAPH '99: Proceedings of the 26th annual conference on Computer graphics and interactive techniques, pp. 187–194.

[3] Boston., L. A., & Klinger, R. (2018). An Analysis of Annotated Corpora for Emotion Classification in Text. Proceedings of the 27th International Conference on Computational Linguistics, (pp. 2104–2119). Santa Fe, New Mexico, USA.

[4] Breazeal, C. (2003). Emotion and Sociable Humanoid Robots. International Journal of Human-Computer Studies, 59(1–2): pp. 119–155.

[5] Bush, L. E. (1973). Individual differences multidimensional scaling of adjectives denoting feelings. Journal of Personality and Social Psychology, 25(1): pp. 50–57.

[6] Cao, Z., Simon, T., Wei, S.-E., & Sheikh, Y. (2017). Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), (pp. 7291–7299).

[7] Ding, C., Xie, L., & Zhu, P. (2014). Head motion synthesis from speech using deep neural networks. Multimedia Tools and Applications, 74: 9871-9888(2015).

[8] Ferstle, Y., & McDonnell, R. (2018). Investigating the use of recurrent motion modelling for speech gesture generation. IVA '18: Proceedings of the 18th International Conference on Intelligent Virtual Agents (pp. 93-98). New York, NY: Association for Computing Machinery.

[9] Gockley, R., Forlizzi, J., & Simmons, R. (2006). Interactions with a moody robot. HRI '06: Proceedings of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction, (pp. 186–193). Salt Lake City Utah USA.

[10] Hall, J., Coats, E., & LeBeau, L. (2005). Nonverbal behavior and the vertical dimension of social relations: a meta-analysis. Psychological Bulletin, 131(6): 898–924.

[11] Kuzmanovic, B., Georgescu, A., Eickhoff, S., Shah, N., Bente, G., Fink, G., & Vogeley, K. (2009). Duration matters: dissociating neural correlates of detection and evaluation of social gaze. Neuroimage, 46: pp. 1154–1163.

[12] Livingstone, S. R., & Russo, F. A. (2018). The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. PLoS One, 13(5): e0196391.

[13] Miles, L. K., Nind, L. K., & Macrae, N. C. (2009). The rhythm of rapport: Interpersonal synchrony and social perception. Journal of Experimental Social Psychology, 45(3): pp. 585–589.

[14] Pantic, M., & Rothkrantz, L. (2003). Toward an affect-sensitive multimodal human-computer interaction. in Proceedings of the IEEE, 91(9): pp. 1370–1390.

[15] Patron-Perez, A., Marszalek, M., Reid, I., & Zisserman, A. (2012). Structured Learning of Human Interactions in TV Shows. IEEE Transactions on Pattern Analysis and Machine Intelligence, 34(12): pp. 2441–2453.

[16] Pelachaud, C. (2009). Modelling Multimodal Expression of Emotion in a Virtual Agent. Philosophical Transactions of The Royal Society B Biological Sciences, 364(1535): pp. 3539–3548.

[17] Pelphrey, K., Singerman, J., Allison, T., & McCarthy, G. (2003). Brain activation evoked by perception of gaze shifts: the influence of context. Neuropsychologia, 41(2): pp. 156–70.

[18] Sadr, J., Jarudi, I., & Shinha, P. (2003). The role of eyebrows in face recognition. Perception, 32: pp. 285–293.

[19] Saerbeck, M., Schut, T., Bartneck, C., & Janse, M. D. (2010). Expressive Expressive robots in education: varying the degree of social supportive behavior of a robotic tutor. CHI '10: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (pp. 1613-1622). New York, NY: Association for Computing Machinery.

[20] Schroeder, B. L. (2011). Eyes, eyebrows and their effect on the facial perception of hostility. Modern Psychological Studies, Vol. 16: No. 2, Article 4.

[21] Sosnowski, S., Bittermann, A., Kuhnlenz, K., & Buss, M. (2006). Design and Evaluation of Emotion-Display EDDIE. 2006 IEEE/RSJ International Conference on Intelligent Robots and Systems, (pp. 3113–3118). Beijing.

[22] Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to Sequence Learning with Neural Networks. Advances in Neural Information Processing Systems 27 (NIPS 2014).

[23] Todo, T. (2018). SEER: simulative emotional expression robot. ACM SIGGRAPH 2018 Emerging Technologies. Vancouver, BC, Canada.

[24] Tomasello, M., Hare, B., Lehmann, H., & Call, J. (2007). Reliance on head versus eyes in the gaze following of great apes and human infants: the cooperative eye hypothesis. Journal of Human Evolution, Volume 52, Issue 3, pp. 314–320.

[25] Wierzbicka, A. (2000). Semantics of Human Facial Expressions. Pragmatics and Cognition, 8(1): pp.147–183.

[26] Wilkins, A. S. (2017). Making Faces: The Evolutionary Origins of the Human Face. Belknap Press.

[27] Wills, T. (1991). Social support and interpersonal relationships. In M. Clark, Review of personality and social psychology, Vol. 12. Prosocial behavior (pp. 265–289). Sage Publications, Inc.

[28] Zee, D. S., Leigh, J. R., & Marthieu-Millaire, F. (1980). Cerebellar control of ocular gaze stability. Annals of Neurology, 7(1): pp. 37–40.

[29] Zhao, G., Huang, X., Taini, M., Li, S. Z., & Pietikainen, M. (2011). Facial expression recognition from near-infrared videos. Image and Vision Computing, 29(9): pp. 607–619.