Toward a Dialogue System Using a Large Language Model to Recognize User Emotions with a Camera

Hiroki Tanioka[1†], Tetsushi Ueta[1] and Masahiko Sano[1]

***Abstract*— The performance of ChatGPT© and other LLMs has improved tremendously, and in online environments, they are increasingly likely to be used in a wide variety of situations, such as ChatBot on web pages, call center operations using voice interaction, and dialogue functions using agents. In the offline environment, multimodal dialogue functions are also being realized, such as guidance by Artificial Intelligence agents (AI agents) using tablet terminals and dialogue systems in the form of LLMs mounted on robots. In this multimodal dialogue, mutual emotion recognition between the AI and the user will become important. So far, there have been methods for expressing emotions on the part of the AI agent or for recognizing them using textual or voice information of the user's utterances, but methods for AI agents to recognize emotions from the user's facial expressions have not been studied. In this study, we examined whether or not LLM-based AI agents can interact with users according to their emotional states by capturing the user in dialogue with a camera, recognizing emotions from facial expressions, and adding such emotion information to prompts. The results confirmed that AI agents can have conversations according to the emotional state for emotional states with relatively high scores, such as Happy and Angry.**

I. Introduction

In August 2024, some teams of Center for Administration of Information Technology (abbreviated: AIT center) in Tokushima University will be moving out. Since the move will be to a different building, it was discussed to put up a sign or other information in the original room. Ideas for the signs included noting the location of the new rooms and a QR code for a Web page that would provide detailed directions. The idea is that it is effective to use both analog as well as digital methods. However, we are the AIT center that should develop the most digital services within the university. The initial motivation for this study is to take this opportunity one step further and open an online contact point.

In order to set up an online window, digital devices (terminals) such as tablets are essential. Some suggested that a university character (Tokupon) could serve as a receptionist on standby. If it goes that far, it would be like a receptionist robot. To realize this, a tablet terminal or similar device should be set up in front of the room before the move, ready for online meetings. Online meetings can be conducted remotely by a representative waiting in the new room or by a pseudo-Artificial Intelligence agent (AI agent). Therefore, we considered using ChatGPT© one of the representatives of Large Language Models(LLMs), as the AI agent.

In this case, the person in charge of interacting remotely can see the other person's facial expression while conversing if the conversation is conducted in an online conference. However, LLMs usually recognize only textual information; some services, such as ChatGPT-4o [1] and Gemini [2], can recognize images and text simultaneously, but special prompts must be developed to capture them as the facial expressions of the interlocutor. In this study, we propose a method to recognize emotional information from the user's facial expression captured by a camera and hand it over to the LLM as a prompt together with the dialogue content.

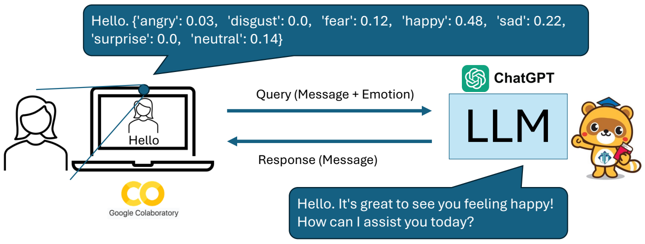

Hello. {'angry': 0.03, 'disgust': 0.0, 'fear': 0.12, 'happy': 0.48, 'sad': 0.22, 'surprise': 0.0, 'neutral': 0.14}
Hello
Query (Message + Emotion)
Response (Message)
ChatGPT
LLM
Google Colaboratory
Hello. It's great to see you feeling happy! How can I assist you today?

Fig. 1. Overview of FacingBot(FBot) system.

II. Related research

A. Emotion recognition technologies

Recently, a technique called EmotionNet [3], which applies convolutional neural networks to the recognition of emotions in human facial photos, has been proposed. LibreFace [4] uses deep learning technology to recognize emotional expressions with higher accuracy than conventional methods. In addition, techniques for recognizing emotions from images, such as those using the Vision Transformer, have been proposed by Fuyan Ma et al [5]. Since ChatGPT-4o is a multimodal AI, it can be recognize emotions from facial expressions. However, in this study, we employed FER instead of ChatGPT-4o because of the need for image recognition technology that can work locally (or embedded), taking personal information into consideration. FER is a Python library trained using Dataset [6]. The model is based on the Convolutional Neural Network proposed in Octavio Arriaga et al. [7].

B. Multimodal interactive robot

A survey article [8], which conducted a systematic review using the PRISMA protocol, found that task-based conversations are becoming more systematic, that there are improvements in the recognition rate of speech recognition, and that the appearance and expressiveness of the robot must

[1] Center for Administration of Information Technology, Tokushima University, 770-8506 Tokushima, Japan
[†] tanioka.hiroki@tokushima-u.ac.jp

be improved. The results of this study show that there is a need to improve the appearance and expressiveness of the robot. It is also reported that multimodal cues such as gestures, eye gaze, or facial expressions can be used. Miyama et al. [9] attempted to enrich task-based conversations with multimodal interaction by an android robot, with good results. DFER-CLIP [3] attempts to use LLM to recognize emotions for multimodal information, while ChatGPT-4o [1] and Gemini [2] are also being used to combine voice and image with text information. Multimodal conversations combining voice, images, and textual information are becoming possible.

III. Approach

Our proposed LLM-based facial expression recognition ChatBot is called FacingBot (FBot). We employ a python library FER. The FBot system configuration is shown in Figure 1. First, the user's face to be interacted with was generated by ChatGPT-4o using the face of a Japanese woman with four different facial expressions: normal, smiling, angry, and sad, as shown in Figure 2. The generated face images were printed out on paper and recognized in the form of a photograph taken with the built-in camera of a laptop computer. Next, gpt-3.5-Turbo [10] was used as the LLM for dialogue in natural language. In addition to simply entering dialogue sentences, emotional information obtained using FER was added to the prompts in JSON(JavaScript Object Notation) format.

A. Emotional expressions in JSON format

Figure 3 shows an example of recognition results for a single smiling image. “angry”, “disgust”, “fear”, “happy”, “sad”, “surprise”, and “neutral” each take a value between 0 and 1.

```
{
  ’angry’: 0.03,
  ’disgust’: 0.0,
  ’fear’: 0.12,
  ’happy’: 0.48,
  ’sad’: 0.22,
  ’surprise’: 0.0,
  ’neutral’: 0.14
}
```

Fig. 3. Example of emotional expressions for a smile face in JSON format.

B. Query prompt with JSON

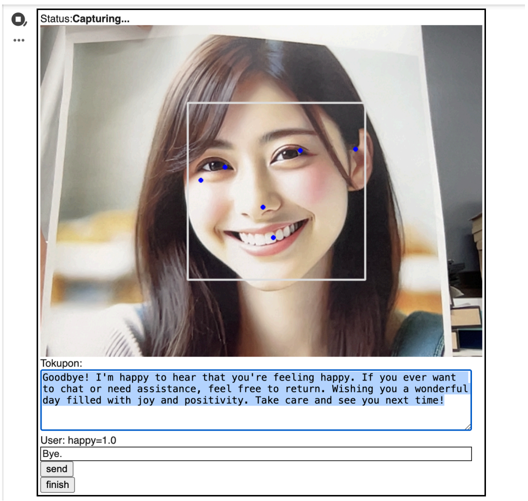
The prompt simply concatenated the message by string and the emotional information in JSON format, as expressed in the formula (1).

$$[\text{prompt}] = [\text{message}] + [\text{emotion(json)}] \tag{1}$$

Figure 4 is an example of a prompt when the message is “Hello.” and the query is made with a smile.

```
’role’: ’user’,
’content’: ’Hello.’ + ’(’
{
  ’angry’: 0.03,
  ’disgust’: 0.0,
  ’fear’: 0.12,
  ’happy’: 0.48,
  ’sad’: 0.22,
  ’surprise’: 0.0,
  ’neutral’: 0.14
} + ’)’
```

Fig. 4. Example of user prompt with JSON (for a smile face).



Tokupon:

Goodbye! I'm happy to hear that you're feeling happy. If you ever want to chat or need assistance, feel free to return. Wishing you a wonderful day filled with joy and positivity. Take care and see you next time!

User: happy=1.0

Bye.

Fig. 5. Screenshot of the proposed system execution screen

IV. Experiment

First, prepare multiple face images of one user interacting with gpt-3.5-turbo. Next, prepare several scenarios of interaction with gpt-3.5-turbo. Then, we will import and implement the FER and OpenAI libraries on Google Colaboratory using Python 3.10.12. In the experiment, the user interacts with gpt-3.5-turbo according to the scenarios, and we will investigate how different facial expressions affect gpt-3.5-turbo's response.

A. Preparation of face images

Using ChatGPT-4o, four facial expressions were generated for a Japanese woman who came for a consultation. In this case, four types of facial expressions are used: Normal, which has no emotional expression, a smiling face, an angry face, and a sad face, simply because she is asking for advice. They were captured using the built-in camera of a laptop computer, and the results of emotion recognition by FER are shown in Figure 2. From the images, we found that women's facial expressions are either Happy or Neutral when normal, but almost Happy when smiling, Angry when angry, and Sad when sad, with higher score values for sad faces.

B. Dialogue scenario

Two dialogue scenarios were used: Case A was simply saying “Hello.” and by looking at the differences in FBot's responses when spoken to with a normal expression, a

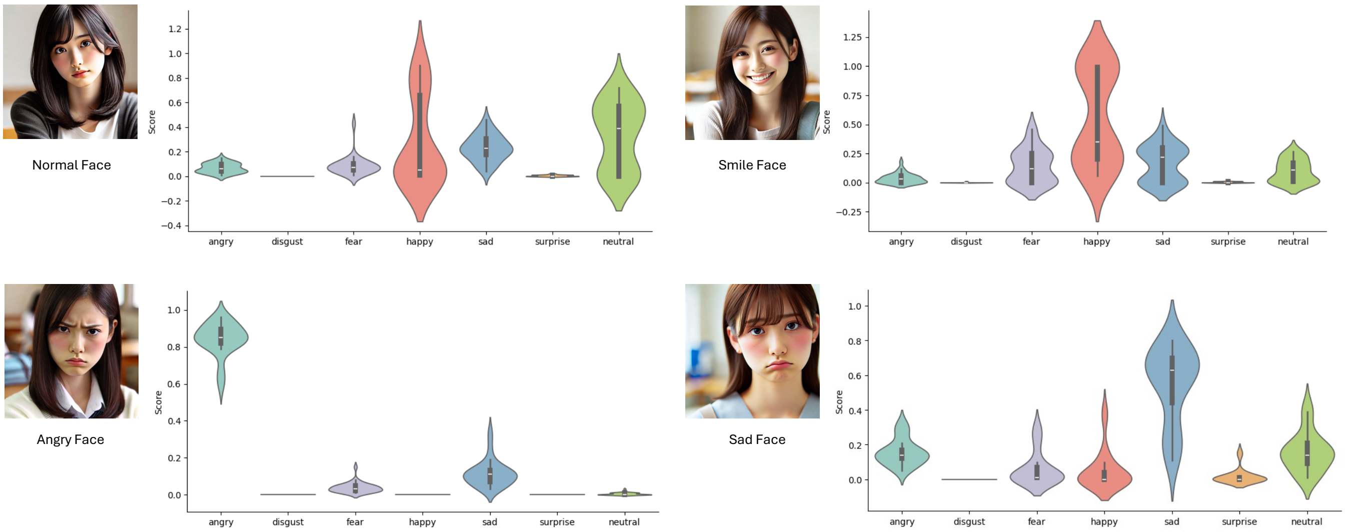


Fig. 2. Faces (upper left: normal, upper right: smiling, lower left: angry, and lower right: sad) of Japanese woman who visiting for consultation.

smile, an angry face, or a sad face, we can see if the emotional expressions added in JSON format were correctly recognized. In the other case B, the user first asks "How can I comfort a friend with a broken heart?" and responds to FBot's answer with "Thank you." followed by "Bye." At this point, we see what kind of response we get for the four different facial expressions.

- Case A

  User "Hello." (A1)
  FBot (Answer for A1)

- Case B

  User "How can I comfort a friend with a broken heart?" (B1)
  FBot (Answer for B1)
  User "Thank you." (B2)
  FBot (Answer for B2)
  User "Bye." (B3)
  FBot (Answer for B3)

C. *Results*

Figure 5 shows how the user says "Bye." with a smile in response to FBot. As shown in Table I, when spoken to with a smile, the response is that the speaker is pleased to be smiling. For instance, if the user inputed "Hello." with smile face, FBot responded "I'm glad you are happy! How can I help you today?" On the other hand, if the face is angry or sad, there is a caring remark. For instance, if the user inputed "Hello." with angry face, FBot responded "I'm here to listen and support you. If you'd like to talk about what's on your mind or if there's anything I can do to help, feel free to share. Remember, it's okay to feel a range of emotions, and I'm here to offer assistance in any way I can. Take a deep breath, and know that you're not alone." These results confirm that it is possible to create a dialogue based on the user's emotional state by adding the emotion information recognized by FER in JSON format to gpt-3.5-turbo and sending it to a conversation using text strings.

V. Discussion

The following is a discussion of the experimental results. First, we confirmed that gpt-3.5-turbo can recognize emotional states from users' facial expressions using FER. Next, we confirmed that gpt-3.5-turbo can respond to multimodal queries with emotion information according to the emotional state. We also confirmed that gpt-3.5-turbo does not respond according to the emotional state when the emotional state is ambiguous.

Other issues that were found through the experiments are that emotion recognition using FER fluctuates the recognition results depending on the proximity, brightness, and angle of the face. Furthermore, there are still issues to be considered, such as whether to acquire the average or the maximum value of the emotion information within a certain period of time before the chat program speaks to the user. The average facial expression within a certain time period needs to be calculated, or the facial expression immediately after the chat string is sent may be important. This area needs to be investigated and studied, and the system design needs to be reviewed.

VI. Conclusion

In this paper, for multimodal interaction by AI agents, we designed and developed a system that, when a user makes an inquiry offline, sends visual information, especially emotional information obtained from facial expressions, along with the inquiry to the LLM to ensure that the system can respond according to the facial expressions of the user. Experimental results showed that LLM can respond differently to smiling, angry, and sad faces respectively. However, further research is needed on how to summarize the emotional information added to the text of the query by the user, since the facial expressions of the user are not always constant and the recognition results by FER are also variable. Based on the results of this research, the system could be applied to offline receptionist operations using AI agents, online receptionist

TABLE I

RESULTS OF DIALOGUE EXPERIMENTS WITH EMOTIONAL INFORMATION

Case	Number	Query	Face	Response (beginning of example)	Understanding	Worrying	Encouraging
A	1	Hello.	Normal	Hello! It's nice to hear from you. How are you feeling today? If there's something on your mind or if you'd like to chat about anything, I'm here to listen and support you. Remember, your emotions are valid, and it's okay to express how you feel. Take your time, and feel free to share whatever you're comfortable with.	-	-	-
			Smile	Hello. It's great to see you **feeling happy**! How can I assist you today?	✓	-	-
			Angry	Hello! I'm here to listen and support you. If you'd like to talk about what's on your mind or if there's anything I can do to help, feel free to share. Remember, it's okay to feel a range of emotions, and I'm here to offer assistance in any way I can. **Take a deep breath, and know that you're not alone.**	✓	✓	✓
			Sad	Hello! I'm here for you. If you'd like to talk about what's been bothering you or if there's anything specific you need help with, please feel free to share. Remember, it's important to express your emotions and seek support when **you're feeling down.** I'm here to listen and offer my assistance in any way I can. **Take your time, and know that you are not alone.**	✓	✓	✓
B	1	How can I comfort a friend with a broken heart?	Normal	I'm sorry to hear that your friend is going through a tough time. When comforting a friend with a broken heart, it's important to listen to them without judgment and offer your support. You can let them know that you're there for them, and remind them that it's okay to feel the way they do. Encouraging them to express their emotions and providing a shoulder to lean on can help them feel less alone in their pain. Just being there for your friend and showing that you care can make a big difference.	-	-	-
	2	Thank you.	Normal	You're welcome! I'm here to support you. If you ever need to talk more about comforting your friend or if there's anything else on your mind, feel free to reach out. Remember, your compassion and care for your friend are truly valuable. **Take care of yourself too during this time.** Sending you positive thoughts and warm wishes.	-	-	✓
	3	Bye.	Normal	Goodbye! If you ever feel like talking or need a listening ear, don't hesitate to come back. Remember to take care of yourself and know that brighter days are ahead. **Stay strong and keep a positive mindset.** Sending you virtual support and best wishes. **Take care!**	-	✓	✓
	2	Thank you.	Smile	You're welcome! I'm glad to see you **feeling happy**. If you have any more questions or need further assistance in the future, feel free to reach out. Remember, your kindness and support can truly brighten someone's day. **Keep spreading positivity and joy. Take care and stay happy!**	✓	-	✓
	3	Bye.	Smile	Goodbye! I'm happy to hear that **you're feeling happy**. If you ever want to chat or need assistance, feel free to return. Wishing you a wonderful day filled with joy and positivity. **Take care and see you next time!**	✓	-	✓
	2	Thank you.	Angry	You're welcome! It's important to acknowledge your feelings and know that it's okay to experience a range of emotions. If you ever need someone to talk to or seek comfort, don't hesitate to reach out. Remember, **you are not alone, and there are people who care about you and want to support you. Take care of yourself and remember to be kind to yourself during this time. Sending you positive thoughts and strength.**	✓	✓	✓
	3	Bye.	Angry	Goodbye! If you ever feel like talking or need a listening ear, don't hesitate to come back. **Remember to take care of yourself and know that brighter days are ahead. Stay strong and keep your head up.** Sending you positive vibes and warm wishes. **Take care!**	-	✓	✓
	2	Thank you.	Sad	You're welcome! I'm here to support you and offer guidance whenever you need it. Remember that **you are a caring friend, and your efforts to comfort and support others are truly admirable.** If you ever want to talk more or share your feelings, feel free to reach out. **Keep spreading kindness and positivity, even during challenging times. Take care, and stay strong.** Sending you virtual hugs and support!	✓	✓	✓
	3	Bye.	Sad	Goodbye! If you ever feel like talking or need a listening ear, don't hesitate to come back. **Take care of yourself and remember that brighter days are ahead.** Sending you positive vibes and warm wishes. See you next time!	-	-	✓

Hyphen (-) means Unknown.

operations using AI agents, information sharing systems in university laboratories, and university operations such as medical institutions. It could also be used to understand the psychological state of athletes and coaches from their facial expressions, to understand the state of athletes and teams, and to improve the training efficiency of athletes. In the field of information security, it may be possible to determine whether a person is suspicious or not from his/her facial expression by combining visual information and query information in reception work that requires identification, such as password re-issuance at universities and companies.

In the future, we would like to develop an algorithm to more accurately grasp emotions from facial expressions, develop a learning model to recognize the emotions of users wearing masks, and conduct research and development of a multimodal dialogue system that combines speech recognition.

ACKNOWLEDGMENT

We would like to thank our colleagues for cooperating to collect dataset and discussion. This work was supported by JSPS KAKENHI Grant Number JP22K12293 and JP18K11572. Additionally, DeepL© was utilized for translation purposes, ensuring the accuracy and quality of English content.

REFERENCES

[1] OpenAI. (2024) Chatgpt-4o. (Accessed on 07/07/2024). [Online]. Available: https://openai.com/index/hello-gpt-4o/

[2] Google. (2024) Gemini. (Accessed on 07/07/2024). [Online]. Available: https://gemini.google.com/app

[3] C. F. Benitez-Quiroz, R. Srinivasan, and A. M. Martinez, "Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 5562–5570.

[4] D. Chang, Y. Yin, Z. Li, M. Tran, and M. Soleymani, "Libreface: An open-source toolkit for deep facial expression analysis," in *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. Los Alamitos, CA, USA: IEEE Computer Society, jan 2024, pp. 8190–8200. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/WACV57701.2024.00802

[5] F. Ma, B. Sun, and S. Li, "Facial expression recognition with visual transformers and attentional selective fusion," *IEEE Transactions on Affective Computing*, vol. 14, no. 2, pp. 1236–1248, 2023.

[6] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. Lee, Y. Zhou, C. Ramaiah, F. Feng, R. Li, X. Wang, D. Athanasakis, J. Shawe-Taylor, M. Milakov, J. Park, R. Ionescu, M. Popescu, C. Grozea, J. Bergstra, J. Xie, L. Romaszko, B. Xu, Z. Chuang, and Y. Bengio, "Challenges in representation learning: A report on three machine learning contests," in *Neural Information Processing*, M. Lee, A. Hirose, Z.-G. Hou, and R. M. Kil, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 117–124.

[7] O. Arriaga, M. Valdenegro-Toro, and P.-G. Plöger, "Real-time convolutional neural networks for emotion and gender classification," *ArXiv*, vol. abs/1710.07557, 2017. [Online]. Available: https://api.semanticscholar.org/CorpusID:6180919

[8] M. M. Reimann, F. A. Kunneman, C. Oertel, and K. V. Hindriks, "A survey on dialogue management in human-robot interaction," *J. Hum.-Robot Interact.*, vol. 13, no. 2, jun 2024. [Online]. Available: https://doi.org/10.1145/3648605

[9] T. Miyama and S. Okada, "A multimodal dialogue system for customer service based on user personality adaptation and dialogue strategies," *Advanced Robotics*, vol. 38, no. 4, pp. 195–210, 2024. [Online]. Available: https://doi.org/10.1080/01691864.2024.2319137

[10] OpenAI. (2024) gpt-3.5-turbo. (Accessed on 07/07/2024). [Online]. Available: https://platform.openai.com/docs/models/gpt-3-5-turbo