Development of Interactive System of Robotic Head

Tran Quang Huy [1] and Nguyen Truong Thinh [2,*]

[1] Department of Mechatronics, Ho Chi Minh City University of Technology and Education, Ho Chi Minh, Vietnam
[2] College of Technology and Design, University of Economics of Ho Chi Minh City—UEH University, Ho Chi Minh, Vietnam; Email: huyquangtran08@gmail.com (T.Q.H.)
*Correspondence: thinhnt@ueh.edu.vn (N.T.T)

***Abstract*—Service robots appear more and more in our life, suitable human-robot interaction concepts are required based on the communication skills and information of database system. In this paper, the interaction of the robot head and the users is described within the contexts in working environment, where the robot and the interactor are deployed. Developing interactive systems for robots, especially service robots, is extremely necessary. This study introduces an interactive system for a robot head. The purpose of the robot head's interactive system is to make the users feel that it is a human head, not a machine or a robot. Its interactive system has many levels and modes. The robot's emotions are also generated by the head with facial movements. Emotions and facial expressions are based on emotional states and human-robot interaction information. The objective of this paper is to create a robot head model capable of interacting with humans naturally thanks to AI combined with facial, eye, and lip gestures, with the aim of improving the efficiency of the human and robot interaction process to assist the elderly person. In the experimental results, we have surveyed the level of satisfaction of the operator with the robot is 78%, while the error level of the robot in this process is only 10%.**

***Keywords*—robotic head, interaction, Human Robot Interaction (HRI), Human Computer Interaction (HCI)**

I. Introduction

In the development of robots, Human-Robot Interaction (HRI) is focused on research, development, and assessment to increase the interoperability between robots and humans. By definition, interaction requires human and robot communication. The study of human behavior and attitudes toward robotics in connection to the physical, technological, and interactive elements of robots is known as HRI, and its goal is to create robots and promote effective interactions between humans and robots like a robot head to assist the elderly person [2]. HRI also acquires the acceptance of people, meets the social and emotional needs of users as well as respects human values. As social robots become more and more multidisciplinary, not only in engineering and computer science, robots now attract more knowledge and resources from the field of human-computer interaction. Then, the term human-robot interaction gradually replaced the previous term "social robot". HRI is a diverse field located between robotics, artificial intelligence, cognitive science, psychology, interactive design [3]. Thanks to the appearance and development of HRI, it is easy for humans to interact, communicate ideas and command to machines and robots [4].

Currently, the field of service robots in the World is developing strongly, but there are still not many models of human-like interactive robot heads to integrate into the humanoid robot system. For the robot head model in this paper, a Vietnamese language data set and a number of language models combined with control methods are created to build the ability to communicate with people like real people. This model contributes to building a robot head system with a friendly shape that can be combined with a complete human robot system. at the same time creating a new language processing system based on advanced language models. This study presents the interactive system of the robot head that allows them to understand and communicate with humans through both speech [5] and facial expressions in reality [6]. In this paper, a robotic head is used to evaluate the interactive functions with users.

II. Configuration of Human-Robot Interaction for Robotic Head

The experimental setup to evaluate the designed HRI is shown in Fig. 1. The goals of those experiments are to enhance and improve the interactions between humans and robots in practice. Communication becomes more authentic as a result.

Figure 1. Robotic head interacting users by HRI.

Manuscript received April 6, 2023; revised May 27, 2023; accepted June 4, 2023; published December 14, 2023.

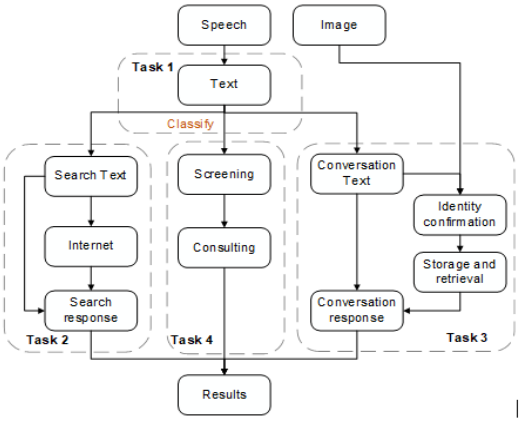

Figure 2. Configuration of the human-robot interaction for robotic head.

The system must also be able to identify symptoms and categorize patient groups according to them [7]. Combining the mentioned three objectives allows the robot to interact and make diagnoses like actual medical professionals, boosting and improving the experience. Fig. 2 depicts the framework of the system for interacting between humans and robots. Four tasks are completed when the user's voice is used as the input. Voice feedback and coding instructions sent down to the system to implement the expressions and change its mouth shape are included in the final output after processing. Three procedures make up Task 1: voice-to-text conversion [8], text classification [9], and emotional recognition [10] in converted text. Processing audio data's structure is difficult. The system performs Task 2 after categorizing the search text in the input data. This involves two processing steps: an Internet search and the initialization of output text. Search for text that requests feedback comprises some knowledge. Due to the high volume of information, a static Recurrent Neural Network model is not able to meet the requirement.

Furthermore, when utilizing static models, it is quite simple to become obsolete with information due to changes in knowledge over time. The system utilizes the internet search technique to resolve this issue. To obtain the required results, a Google search tool is used. After receiving the search results, this data is sent to the process of launching the search response. To generate input for the text response initiation procedure, the searched results are concatenated by the system. An upgraded Recurrent Neural Network (RNN) [11] model is used in this procedure for mapping the inputs to the correct outputs. The outcome is then integrated with the optimization methods to provide a final output.

This system completes Task 3 after categorizing the input data as conversational text. This entails three processing steps: validating the user's identification, saving data, and starting a text response. Robotic communication requires identifying the identity of the interlocutor, just like human communication does. By removing data that is not important for conversation, the system can concentrate on processing information belonging to the interlocutor. After the identification has been verified, the information is sent to the procedure for storing user data.

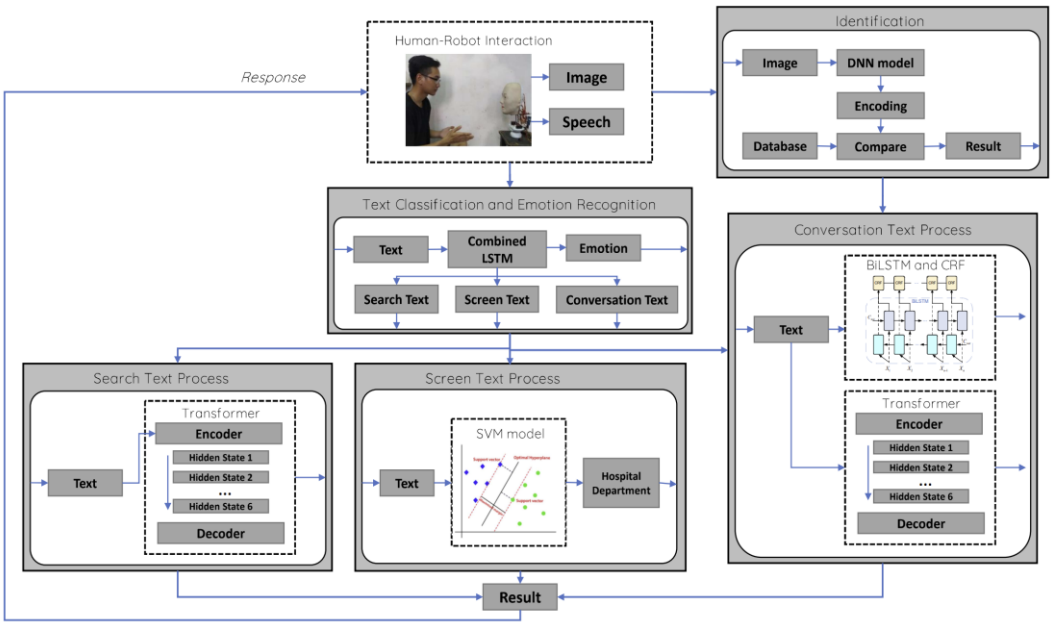

Figure 3. Schematic of Machine Learning models in the study.

Emotional recognition data is also encoded to convey emotion to robots. The system then sends the data to the robot head after synthesizing it. The robot head execute expressive and communicative exchanges through studies with participants after significant processing processes. The interoperability of the robot head system is tested through experiments. Fig. 3 shows the schematic of the interactive control system of the robot head based on the input information of the system that the robotic head gives appropriate responses. The control system is based on the Machine Learning (ML) process with various components such as text classification and emotion recognition,

identification. This control process relies on ML models to provide appropriate responses to the users. The components of an ML-based controller are described in more detail in the following sections.

III. Natural Language Processing for Robotic Head

A supervised learning model [12] is used, "Combined Long Short-Term Memory (LSTM)", Deep learning and artificial intelligence both use the Long Short-Term Memory (LSTM) artificial neural network. The LSTM has feedback connections as compared to standard feedforward neural networks developed on "Bidirectional Long Short-Term Memory (BiLSTM)" [13] is an encoder. Similarly, "Long Short-Term Memory (LSTM)" [14] is a decoder. The RNN model BiLSTM is able to solve issues with inputs and outputs in sequence. This is performed by the capacity to memorize and combine data from previously handled sequence elements. The backward and forward layers of BiLSTM are two hidden layers. A series of LSTM architectures generate each layer. Speech is the primary input signal for the robot-human interaction system for robotic head while interacting with people. The audio format, however, is difficult, lacks visuals, and frequently includes a noise signal. To address this issue and create the stages of data processing easier, signals are switched from audio to text format. The robot-head can respond appropriately to the communication scenario based on emotional identification and then classification. Empathy for the other person and conversation quality are both impacted by emotions. To choose the best processing method for the dialoguer in the next steps, the system uses text classification to determine what the dialoguer requires. The text classification and emotional recognition functions in this part are handled by the AI model.

The system uses deep learning models [15] for categorization after handling the input data. A sequence input data processing model is needed for this classification procedure in order to create predictions. As a result, this procedure uses the combined LSTM model. Once processed, the data will be sent to the encoder block. Components of this block are BiLSTM architectures. The encoder block's duty is to extract some essential data from the input text and save it in the S matrix. Beside that the system regularly modifies the weight matrix values in BiLSTM architecture to gather this data in order to improve efficiency. This procedure is similar to gathering data that aids in the system's understanding of the input text. Next, the Decoder block receives the S storage matrix. This block's components are LSTM architectures. The S matrix and vector null are the primary inputs for the Decoder block's first component (vector zeros). The system keeps changing the weight matrix of LSTM architectures to provide the output that is most similar to the label that was previously assigned in order to be able to anticipate the outcomes for each component. The Decoder block's output is sent on to the Sofmax block. The words that are most likely to appear in this context were determined using the outcome of each Decoder block component. Select the appropriate word to continue refining the result sentence. Robotic-head's system eventually changes the embedding procedure to return the data to its original text format. The combined LSTM model is utilized to complete 2 tasks that classify text and recognize emotions.

IV. Search Text Processing

A transformer is a deep learning model that adopts the mechanism of self-attention [16] and is able to identify the connections between various elements in the input data. Transformer is designed to solve problems that are data in a sequence form such as text translation, question-answer... This is quite similar to the Recurrent Neural Network (RNN) model. However, transformers do not require processing the incoming data in the same order as other RNN models.

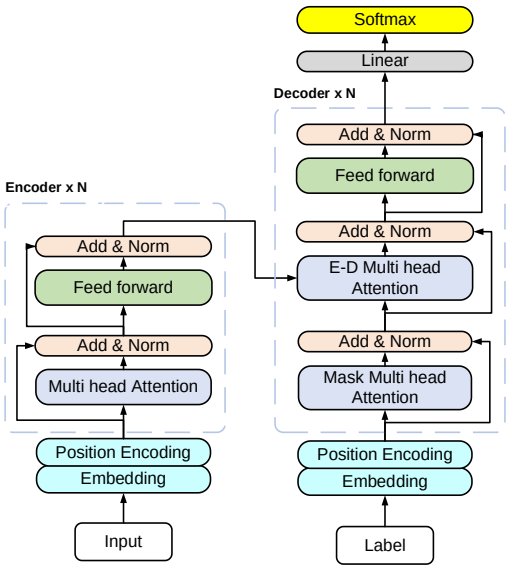

Figure 4. Structure of Transformer model.

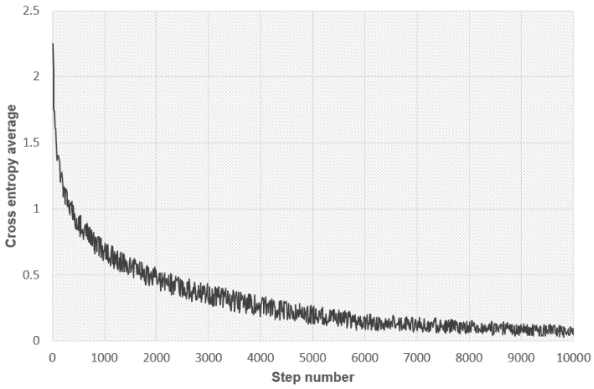

Figure 5. Cross entropy.

When dealing with text or string data, the mechanism of self-attention allows model to comprehend the meaning of the sentence even when analyzing the full sentence at once. Transformers don't have to process the sentence's words in order. Instead, it gives each word a location and establishes how it relates to the other words. The major benefit of Transformer in comparison to most other RNN models is the capacity to conduct numerous calculations in

parallel and shorten processing times. The Transformer model further handles the incoming sentence in two different ways. The structure of the Transformer model is shown in Fig. 4. The encoder of the model receives the data and encodes it into vector S carrying full data to represent the input sentence. Decoder of the model receives vector S and processes it then predicts the output sentence.

Transformer is more complex than the Combined LSTM model that was shown, but the data still needs to be processed sequentially. In consideration of this, combined LSTM is only appropriate for small tasks. The Transformer model is a comfortable option that greatly speeds up data processing for more complex tasks like initializing responses while also improving accuracy. Processing the input data is the first stage. We gather information using the supplied internet search methodology using the 500 questions from various fields that are previously listed. The results of each internet search and question are concatenated. To label the newly generated strings, the solutions are also manually generated using the data mentioned above. To make the model's learning process simpler, the produced answers attempt to utilize the most of the data gathered from the internet. As a result, input data and labels are available for use in model training. The results with the loss of function are cross entropy is shown in Fig. 5. The cross entropy average value is 0.074 and the precision is 90.14%.

V. Screening and Classification of Diseases

The screening and classification model of the disease was built on the Support Vector Machine (SVM) are a type of supervised learning technique that can be applied to classification or regression tasks. The fundamental goal of SVM is to identify a hyperplane that maximally discriminates between the various classes in the training set) (SVM) model [17]. The input is the symptoms of patient and the output is the departments of the hospital that patient needs to go to in next step. Specifically, the information we collect directly from the patients and the results have been screened by the doctors. The dataset used in this study includes 62 symptoms as input and 10 basic hospital departments as outputs. Each symptom is coded as a vector number. The symptom vectors of each case are stacked together to form a matrix. To ensure that the input matrices of the different cases have the same size, each matrix is specified in size 62×62 as the standard size. Where each row represents a symptom. Each symptom is encoded as a hot vector. For symptoms that do not appear in the case are encoded as vector zero. The 10 outputs are encoded as numbers from 0 to 9. The resulting data encoding is shown in Fig. 6. The SVM model used in this study is built on the *Keras* platform. For best results, the model is adjusted manually. Radial Basis Function, Linear, and Poly are the kernels used to compare the results and choose the most suitable method. The linear kernel gives the worst results about 72%. The poly kernel achieves good results at 4 and 5 degrees, and changes with varying input quantity and content. The Radial Basis Function kernel achieves more stable results when varying the amount and content of inputs, it achieves up to 91.21% accuracy on the local dataset. In order for the system to recognize symptoms during user interaction, a symptom dictionary is built. Which includes 62 groups corresponding to the 62 symptoms mentioned above. Each group includes synonyms and regional terms for each symptom. Each input sentence will be checked through this dictionary to detect symptoms appearing in the sentence. To ensure that the system does not confuse the screening process with the normal communication process, when the system perceives that the patient requires a screening, it will ask again to ensure. Avoid cases of misidentification of patient requests.

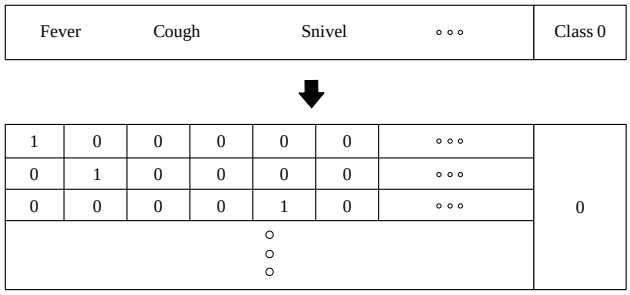

Figure 6. The process of encoding input data of the SVM model.

VI. Experiments and Discussions

The models in the previous sections returned responses in the form of text. In interaction, user often uses his voice to communicate. This section will present the process of converting text to voice so that the system can communicate more like a human. After going through three important processing steps: Text categorization, search text, and conversational text handling the input data generates two results: the voice response and the face's emotions.

The emotions of the dialogue are the result of the text classification step. The response is the result of the remaining two steps. The emotions continue to be encoded and send to the microprocessor of the robot head performing the expression. The purpose of designing according to human structure and ability of motion is to give robotic head the expressions and communicate as much as possible. In this section, the experiments carried out and evaluate the expressions performed by the robotic head. The emotional recognition experiments were conducted with the participation of a group of 100 people. Each volunteer received a survey sheet with six separate rows. Each row consists of seven options including: happy, sad, anger, surprise, disgust, fear, and error. Participants saw the robotic head perform six expressions in a random order. Each time the robot performed an expression, participants were asked to mark the corresponding expression in the survey that they think most appropriate. In case participants were unable to recognize the expression in the making, they are advised to tick the error box. The expressions of the robotic head performed during the experiment are shown in Fig. 7. During the experiment with a large number of participants, to avoid erroneous results or misunderstand the experimental process. The

results collected from the experimental process were tested and eliminated by statistical method. The results of the experiment were aggregated and calculated the standard deviation. Then results with errors three times greater than the standard deviation were eliminated. Finally, the results are compiled and presented in Table I.

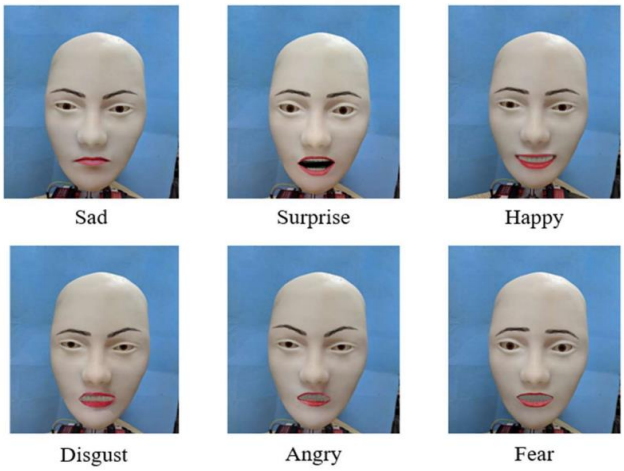

Figure 7. Robotic head interacting to users.

The results of the emotional recognition experiment are shown in Table I. The columns represent the expressions that the robotic head has performed in the experiment. The rows correspond to the choice of the experiment participant. The columns and rows in Table I show the expression of the robotic head and the results of the assessment of volunteers. The results showed that the expressions of the robotic head were appreciated. The expression is recognized with a minimum rate of 81% (Fear). Expressions of happiness, sadness, and anger are the three expressions with the highest recognition rate.

The volunteers did not have any misunderstandings with expressions of happiness, sadness, anger (error is 0.0). This is because these expressions appear with a high frequency on a daily. The remaining expressions have a higher frequency of confusion. In particular, expressions of anger are often chosen by volunteers. Of the 115 options for expressing anger, eight confuse with disgust and nine with fear. In addition, volunteers are sometimes misunderstood about disgust and fear. This can be explained because expressions of disgust and fear appear infrequently and all three of these expressions are negative. The overall assessment expression of robotic head has a recognition rate of up to 89.83%.

TABLE I. EXPERIMENTAL RESULTS OF ROBOT EMOTION RECOGNITION

% match	Surprised	Happy	Disgust	Sad	Anger	Fear
Surprised	88	0	0	0	4	0
Happy	0	97	0	0	0	0
Disgust	1	1	86	1	0	1
Sad	0	0	0	94	1	4
Anger	3	2	8	0	93	9
Fear	7	0	2	5	2	81
Error	1	0	4	0	0	5

The robotic head was created with the main objective of interacting with humans. The user's impression and response would decide how well the interaction performed. Thus, the testing process must be carried out in order to test the robot's interactive system. Two different supervisors carried out the experiment. 50 volunteers were involved, including 3 students, 33 employees, 8 engineers, and the remaining volunteers come from different professions. Men makeup 74% of the total, women 26%, persons over 25 make up 22%, and people under 25 make up 78%. Users were invited to stand or sit 0.5 meters away from the robot (Fig. 1). The robot will converse with each participant for 3–5 min, and there is no restriction on the subject matter. Volunteers are required to say, "Hello robotic head", to begin the interaction. Users were invited to complete a survey after the interaction. Five entries were included in the survey, each of which answered one of five interaction-related questions. What do you know about robots, for example? How do you assess the robotic head's information? Does robotic head frequently make mistakes? Are you happy with how we've been talking? Do you find communicating with robotic head to be simple? A 5-level Likert scale was used to evaluate each of the aforementioned questions (range from 1.0 to 5.0). Level 1 denotes a high negative level, Level 2 a moderate level of negativity, Level 3 is normal, Level 4 a moderate level of positivity, and Level 5 a high level of positivity. For the question, "Do you know about robotics?" participants were required to take one of the scale's five levels. Level 1 and Level 2 point out that the volunteers have no understanding of robots or have not used one, whereas Level 4 and Level 5 demonstrate that the volunteers had contacted with the robotics. Because most volunteers are students or employees who have been exposed to robots, therefore Fig. 8(a) demonstrates that the majority of volunteers have knowledge of or experience with robots at a rate of up to 72% (Level 4 and Level 5). There are 10% of humans who don't understand robots very well (Level 1 and Level 2), the majority of participants select this part are employees, they have minimal possibility of interacting with robots. The volunteers' level of evaluation of the robotic head's information is revealed by the questions, "How do you evaluate the data presented by the robotic head?" There are five options, each one corresponding to one of the Likert scale's five levels: Level 1 and Level 2 have few or no useful information, while Level 4 and Level 5 include useful data. The outcome of the review is shown in Fig. 8(b). The participants claim that the robot cannot respond to questions relating to academic matters because 14% of them (Level 1 and Level 2) believe the information is not or is of little use. The 68% of volunteers for Level 4

or Level 5 on the likert scale show that the data collected from robotic head is valuable, making up more than two-thirds of the total number of volunteers participating. These findings indicate that robotic heads were able to address the majority of the questions posed by humans.

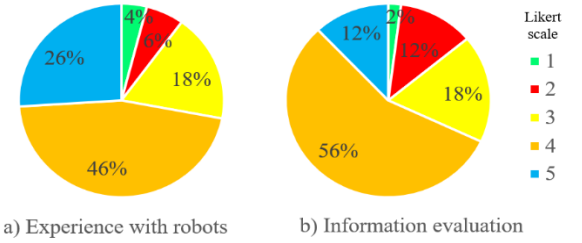

Figure 8. Survey evaluations: a) Experience with robot; b) Information evaluation.

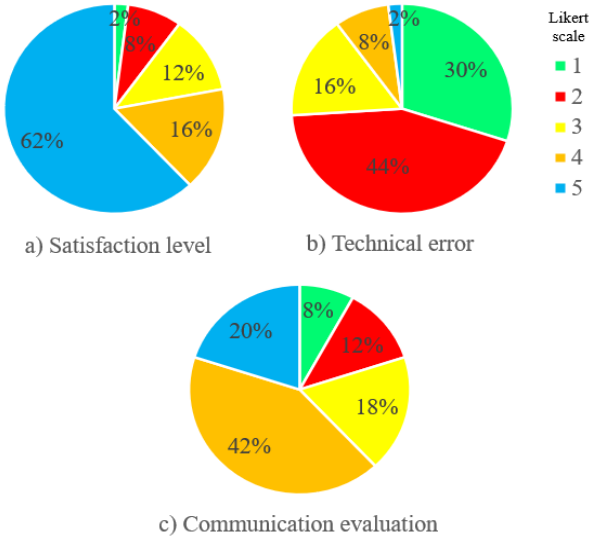

Figure 9. Subjective evaluations: a) Satisfaction level; b) technical error; c) communication evaluation.

The volunteers' satisfaction with their interactions with the robotic head is shown in Fig. 9(a). The response choices to the question "Are you content with this conversation?" will correlate to five likert scales, with Level 1 and Level 2 expressing displeasure and Level 4 and Level 5 expressing pleasure. Only 10% of the volunteers in total expressed dissatisfaction with their interactions with the robot, according to the results (Level 1 and Level 2). The participants gave reasons for their decision, some of them said that the robot could not understand when the interlocutor spoke too quickly or used a distinct regional accent because its voice and intonation were different from those of humans. In contrast to the low levels of Level 1 and Level 2, 78% of volunteers expressed satisfaction when conversing with robotic head (Level 4 and Level 5). The participants claimed the robotic head had a friendly demeanor and gave the other person a sense of security. Participants were requested to take one of five levels for this question "Does robotic head frequently encounter errors?" Level 1 and Level 2 suggest that robots are difficult to make technical faults, while Level 4 and Level 5 indicate that the robot is likely to make technical problems. According to the review's results, which are shown in Fig. 9(b), 10% of volunteers thought robots were prone to technical faults (Level 4 and Level 5). The participants pointed out that they did not think the robot's structure was very sturdy, while others stated that they thought the robot occasionally did not function very well. The proportion of volunteers who believed robots rarely made technical faults reached 74%. (Level 1 and Level 2). This data demonstrates that robotic-head has few technical faults during operation. The question "Do you think it's easy to communicate with a robotic head?" reveals how easy it is for volunteers to communicate with a robotic head. The Likert scale, which has five levels, is used similarly to other questions: Level 1 and level 2 suggest that it is difficult to communicate with robots, whereas Levels 4 and 5 show that it is simple to communicate with robots. The experiment's results are depicted in Fig. 8. When communicating with robots, 20% of volunteers (Level 1 and Level 2) had difficulties. To explain their decision, the volunteers stated that it was difficult for the robot to understand their question if it comprised a few English words, and several candidates claimed that the robot did not hear their inquiries completely. Sixty-two percent of volunteers said it was simple to communicate with robots (Level 4 and Level 5). They claimed that robotic head's voice is very easy to understand and that the robot thoroughly answers all of their questions, with some volunteers even claiming that the robot occasionally tells jokes. Based on the results of the foregoing tests, it is possible to conclude that robotic head fully meets the requirements of human-robot interaction.

VII. Conclusions

The Robot Head Interaction Research has presented the development of emotional expression for the interactive robot head as well as the construction of an emotion recognition model using the voice of the interlocutor as input data. According to the test results, the robot head responds to human facial expressions and verbal commands with good efficiency. The artificial intelligence system will be able to use the interlocutor's speech and picture data as a resource to improve the user's interaction with the Robot once it has been collected and saved.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Tran Quang Huy is the first author—writing original draft, visualization, validation, configured, coding. Nguyen Truong Thinh—methodology, writing, review and editing, project administration. All authors had approved the final version.

Funding

This research is funded by University of Economics Ho Chi Minh City—UEH University, Vietnam.

REFERENCES

[1] T. B. Sheridan, "Human–robot interaction: Status and challenges," *Human Factors*, vol. 58, no. 4, pp. 525–532, 2016.

[2] M. A. Ayub, M. N. A. Patar, and N. A. Zainal, "AZ'humanoid robot head with object and color tracking capabilities," *Int. J. Mech. Eng. Robot. Res*, vol. 9, no. 6, 2020.

[3] M. Zheng, A. Moon, E. A. Croft, and M. Q. H. Meng, "Impacts of robot head gaze on robot-to-human handovers," *International Journal of Social Robotics*, vol. 7, pp. 783–798, 2015.

[4] C. McGinn, "Why do robots need a head? The role of social interfaces on service robots," *International Journal of Social Robotics*, vol. 12, no. 1, pp. 281–295, 2020.

[5] N. T. T. Phong, L. H. T. Nam, and N. T. Thinh, "Vietnamese service robot based on artificial intelligence," *International Journal of Mechanical Engineering and Robotics Research*, vol. 9, no. 5, 701–708, 2020.

[6] N. K. Toan, L. B. L. Thuan, and N. T. Thinh, "Development of Humanoid Robot Head Based on FACS," *International Journal of Mechanical Engineering and Robotics Research*, vol. 11, no. 5, 2022.

[7] O. Engwall and J. Lopes, "Interaction and collaboration in robot-assisted language learning for adults," *Computer Assisted Language Learning*, vol. 35, no. 5–6, pp. 1273–1309, 2022.

[8] A. Trivedi, N. Pant, P. Shah, S. Sonik, and S. Agrawal, "Speech to text and text to speech recognition systems—A review," *IOSR J. Comput. Eng*., vol. 20, no. 2, pp. 36–43, 2018.

[9] K. Kowsari, J. Meimandi, M. Heidarysafa, S. Mendu, L. Barnes, and D. Brown, "Text classification algorithms: A survey," *Information*, vol. 10, no. 4, 150, 2019.

[10] E. Pranav, S. Kamal, C. S. Chandran, and M. H. Supriya, "Facial emotion recognition using deep convolutional neural network," in *Proc. 2020 6th International Conference on Advanced Computing and Communication SYSTEMS (ICACCS)*, IEEE, 2020, pp. 317–320.

[11] J. Duque-Domingo, J. Gómez-García-Bermejo, and E. Zalama, "Gaze control of a robotic head for realistic interaction with humans," *Frontiers in Neurorobotics*, vol. 14, no. 34, 2020.

[12] V. Nasteski, "An overview of the supervised machine learning methods," *Horizons*, vol. 4, pp. 51–62. 2017.

[13] G. Liu and J. Guo, "Bidirectional LSTM with attention mechanism and convolutional layer for text classification," *Neurocomputing*, vol. 337, pp. 325–338, 2019.

[14] Y. Yu, X. Si, C. Hu, and J. Zhang, "A review of recurrent neural networks: LSTM cells and network architectures," *Neural Computation*, vol. 31, no. 7, pp. 1235–1270, 2019.

[15] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[16] Y. Wang, J. Zhang, M. Kan, S. Shan, and X. Chen, "Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation," in *Proc. the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 12275–12284.

[17] S. Suthaharan, "Support vector machine," in *Machine learning Models and Algorithms for Big Data Classification*, Boston, MA.: Springer, 2016, pp. 207–235.