ADVANCING HUMAN-ROBOT INTERACTION: DEEP LEARNING-BASED EMOTION AND GESTURE RECOGNITION FOR IVASTBOT

Ha Thi Kim Duyen
Faculty of Electronic Engineering
Hanoi University of Industry
Hanoi, Vietnam
ha.duyen@haui.edu.vn

Tien Ngo Manh
Institute of Physics
Vietnam Academy of Science and Technology
Hanoi, Vietnam
nmtien@iop.vast.vn

Dang Cam Thach
Faculty of Electronic Engineering
Hanoi University of Industry
Hanoi, Vietnam
Dangcamthach@haui.edu.vn

Duy Ngo Manh
Faculty of Electrical and Electronic Engineering
Phenikaa University
Hanoi, Vietnam
ngomanhduy098@gmail.com

Quang Doan Khoi
Faculty of Electronic Engineering
Hanoi University of Industry
Hanoi, Vietnam
vodien1102ks1@gmail.com

Hiep Tran Nguyen
Thanh Dong University
Hai Duong, Vietnam
hiepnt@thanhdong.edu.vn

ABSTRACT

In this paper, we introduce a novel approach to enhance the capabilities of the humanoid robot IVastBot by in- tegrating various software components. This integration enables IVastBot to effectively recognize and respond to a wide array of human gestures and behaviors. Through the utilization of the open-source MediaPipe Pose library and LSTM networks, IVastBot becomes proficient in generating contextually appropri- ate responses. Furthermore, we incorporate emotion recognition into the system using Convolutional Neural Networks (CNN). The entire recognition module seamlessly integrates into the Robot Operating System (ROS) architecture, resulting in efficient execution. Consequently, IVastBot achieves the ability to execute adaptive actions in response to human gestures and emotions, sig- nificantly enriching the intuitiveness and engagement of human- robot interactions.

Keyword: *Human-robot interaction (HRI); Machine Learn- ing (ML); Deep Learning (DL); Robot Operating System (ROS).*

1. INTRODUCTION

In recent years, the realm of robotics has experienced remarkable strides, leading to the emergence of social robots that hold the potential to redefine human-machine interaction (HRI). These robots, enriched with artificial intelligence (AI) capabilities, are poised to establish profound and transforma- tive connections with humans, reshaping our relationship with technology.

Social robots, characterized by their anthropomorphic at- tributes and

interactive functionalities, present a compelling opportunity to bridge the gap between humans and machines. Through adeptly perceiving and responding to human emo- tions and gestures, these robots enable more seamless, natural, and intuitive interactions. This, in turn, elevates user engage- ment, fosters trust and ultimately enhances the overall user experience. The integration of deep learning models further amplifies this potential, given their exceptional performance across a spectrum of AI tasks.

By harnessing the synergy of AI, emotion recognition, and gesture recognition, social robots are positioned to usher in a new era of interactive technology. As this paper delves into the exploration of these advancements, we aim to un- ravel the intricacies of this multidisciplinary field, providing insights into the mechanisms that drive empathetic and effec- tive human-robot interactions. Paper (2), (3) has introduced new demands for this interaction, aiming for the robot's capability to perceive human facial expressions, comprehend, and appropriately respond to emotions (4), (5). In recent years, emotionally intelligent robots have garnered significant attention. Despite notable progress, this domain is still nascent, with only a handful of intelligent service systems successfully utilizing emotion recognition technology.

For instance, one of the pioneering social robots, Kismet, was developed at MIT by Dr. Cynthia Breazeal and her research team in the 1990s. Kismet possesses the ability to recognize humans and objects, as well as simulate various emotional expressions (1). The robot Jibo, designed with a swiveling head and a communication screen, employs voice recognition technology and perceptual abilities to interact with humans. Pepper, a humanoid social robot developed by Softbank, surpasses its predecessors by incorporating two arms and legs for mobility, enabling more dynamic interactions with its environment. Pepper can sense and respond to human emotions through sensors embedded in its body (1).

The OpenPose algorithm (25,26) is characterized by its ability to detect 2D body key points of multiple individuals within a single frame, with configurable key point options. Notably, it maintains real-time invariance concerning the num- ber of individuals in the frame. Additionally, it offers real- time 3D keypoint tracking for an individual. In terms of its network architecture, OpenPose initially consists of a CNN network with two branches. The first branch is responsible for generating keypoint heatmaps, while the second branch connects these key points to form human poses using the Part Affinity Fields (PAFs) algorithm.

OpenPose can accept various inputs, including images, videos, webcams, and IP cameras. The output includes the original image alongside the visualization of key points in popular formats such as PNG, JPG, and AVI, or the storage of key points in JSON, XML, and YML file formats. Once the key point weights are obtained, the data is fed into a network system for training to produce corresponding behavioral and gesture results.

Numerous authors have employed the combination of Open- Pose and LSTM / RNN methods for Human Activity Recog- nition (HAR) (20-24). However, when applying these ap- proaches to a specific robot with the capability to recognize the actions of surrounding individuals, the closest practices involve integrating these algorithms into the central control system of the robot, often utilizing ROS (27-31). Notably, there is a scarcity of applications that integrate these methods into humanoid robot systems.

This paper introduces an innovative approach to enhance the capabilities of the humanoid robot IVastBot through software integration, enabling it to recognize both gestures and behav- iors exhibited by interacting humans. This integration empow- ers IVastBot to respond with contextually appropriate actions. To achieve this, we leverage the potential of the open-source MediaPipe Pose library and LSTM networks. Additionally, we incorporate emotion recognition using Convolutional Neural Networks (CNNs) into the system. The entire recognition module is seamlessly embedded within the Robot Operating System (ROS) architecture, executing on the high-performance Nvidia Jetson TX2 processing unit. As a result, IVastBot gains the ability to execute adaptive actions in response to both human gestures and emotions, fostering more intuitive and engaging human-robot interactions.

2. ARCHITECTURE OVERVIEW OF THE INTELLIGENT HUMANOID ROBOT IVASBOT

2.1. Hardware Structure

The intelligent humanoid robot, IVASTBot, effectively achieves autonomous movement utilizing an Omnidirectional wheel configuration (12), coupled with an advanced deep reinforcement learning algorithm for obstacle avoidance, en- compassing both dynamic and static impediments. Its core processing power stems from the specialized Nvidia Jetson TX2, a high-performance computing platform, with the entire system orchestrated through the Robot Operating System (ROS).

IVASTBot's physical embodiment features a dexterous dual- arm arrangement, boasting three joints per arm, while the robot's head incorporates a two-joint design, facilitating in- tricate

bodily interactions with users. Notably, IVASTBot demonstrates a remarkable capability in deciphering multi- faceted emotional cues and human facial expressions using the Astra camera. Subsequently, it engages users by presenting interfaces and fostering emotional connections through an integrated LCD screen (10).

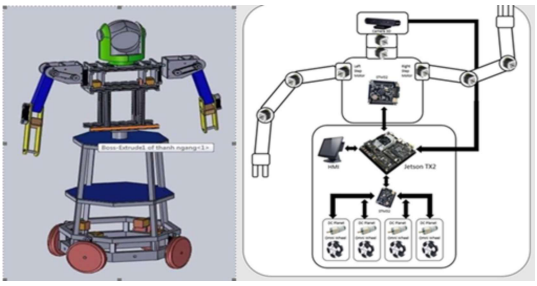

Fig. 1. *IVASTBot Hardware structure*

2.2. System Architecture

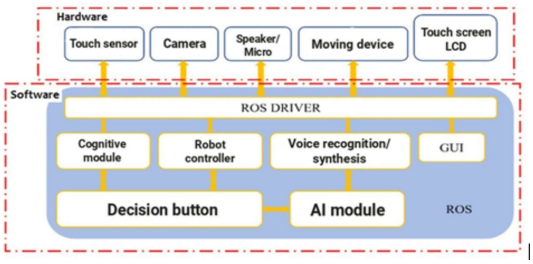

Fig. 2. *System architecture of IVASTBot*

Within the software framework, an array of modules can be identified, each serving distinct functions critical for efficient system operation. These modules encompass data processing for camera inputs, object recognition and retrieval from the scene, voice recognition and synthesis, gesture recognition and synthesis, artificial intelligence components, and robot control modules for actuating devices. The output decision node effectively amalgamates data sourced from the various sensors, culminating in the decisive determination of the subsequent course of action. The ROS control layer plays a pivotal role in seamlessly interfacing all sensors and transmission devices with ROS, while the graphical user interface (GUI) provides an intuitive visual interaction through the LCD control panel.

3. EMOTION RECOGNITION

3.1. Input Images Preprocessing

The input images captured by the camera undergo a trans- formation process, converting them from RGB to multi-level grayscale while ensuring gray-level balancing. This optimiza- tion enhances the facial recognition system's quality.

The Haar-cascade method (13) is employed for facial detec- tion and region-of-interest extraction. This technique utilizes a sliding window approach across the image, extracting Haar- like features (based on the Histogram of Oriented Gradients) at each window position. These features are incorporated into the Adaboost model with a cascading mechanism (4) to accurately classify and eliminate unrelated regions.

This method exhibits exceptional computational speed due to the utilization of Haar-like feature extraction, surpassing other techniques. The cascading Adaboost model

effectively eliminates non-facial candidate regions from the early stages, focusing on fewer candidates in subsequent stages. Compared to the facial detection tool in the OpenCV library, the Haar- cascade facial detection technique, based on the Viola-Jones algorithm and implemented in the Dlib library, demonstrates superior performance even under challenging and constrained image conditions (15).

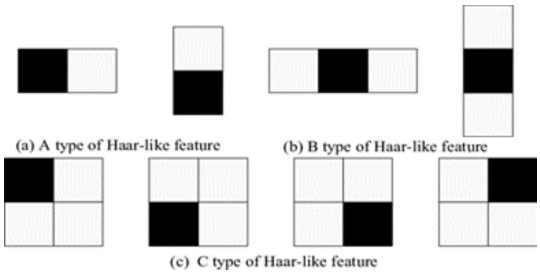

Fig. 3. *Haar-like features.*

3.2. Model

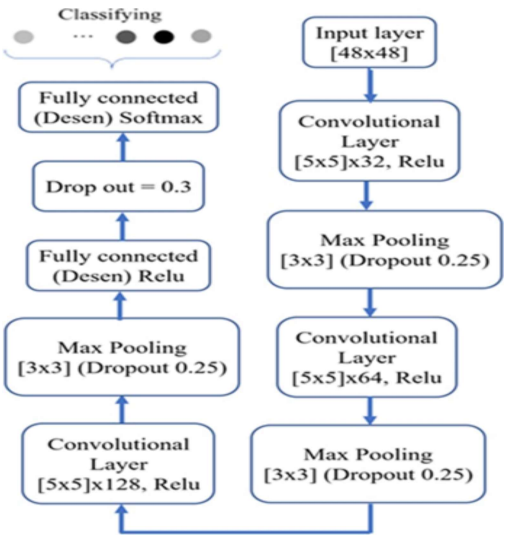

Fig. 4. *Model Architecture*

In this study, we propose a convolution neural network architecture shown in Figure 4. The input image is a grayscale image of 48x48 pixels. The model comprises 3 convolutional layers, 3 pooling (also known as neuron aggregation) layers, and 2 fully connected neuron layers for classification (Dense layers) with 674,823 parameters to be adjusted. The activation function employed in all 3 convolutional layers is the ReLU (1) (Rectified Linear Unit) non-linear function to extract char- acteristic values of the image. To mitigate overfitting during neural network training, this model incorporates the Dropout technique after each Pooling layer, rendering the model less sensitive to specific weights within the network. The model's dropout rates for each layer are set using heuristic methods based on experimentation (16).

$$f(z_1) = \max(0, z_i) \qquad (1)$$

In the final layer, the activation function utilized is SoftMax, enabling the distribution of output probabilities across a total of 7 different classes. The loss function employed is the categorical cross-entropy function (2), combined with the Momentum optimization function with a learning rate of 0.001 to optimize the model's parameters.

$$L(y, \hat{y}) = -\sum_{i=1}^{N} y_i \log(\hat{y}_i) \qquad (2)$$

3.3. Trainning and Evaluating Model

The model utilizes the FERC-2013 dataset available on Kaggle for training purposes. The dataset comprises a total

of 35,887 grayscale images, each with a dimension of 48x48 pixels.

Fig. 5. *FERC-2013 Dataset*

In the context of this study, an essential strategy for en- hancing the generalization capability of neural networks is the careful allocation of data for training. To mitigate the risk of overfitting, 80% of the available data from the FERC- 2013 dataset was allocated for model training. The initial preprocessing step involved classifying and arranging images within the FERC-2013 dataset.

The dataset contained image representations of seven dis- tinct emotional categories. These images were converted into binary matrices with dimensions of 48x48 pixels, thereby enabling compatibility with the selected neural network ar- chitecture.

The model was trained for a total of 100 epochs and achieved an approximate accuracy of 67%, aligning with the initial objective of developing a simple application. The accuracy results showed minimal deviation when compared to the testing accuracy.

In the confusion matrix (Figure 7), emotions such as Surprise, Happy, and Neutral exhibited the highest accuracy percentages (above 70%) due to their prominent facial expres- sions and the substantial amount of training data available. However, emotions like Disgusted, Sad, Fearful, and Angry had lower accuracy percentages (below 60%) as they are more

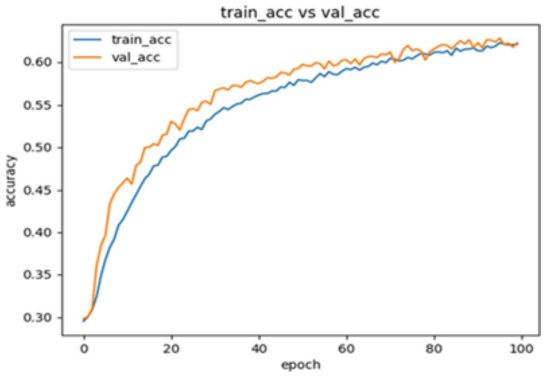

Fig. 6. *Training and Testing Accuracy Result*

challenging to differentiate. The prediction matrix indicated that a smaller amount of training data for certain emotions resulted in lower accuracy, particularly for the ”Disgusted” emotion. Overall, the model’s performance is satisfactory for a simple application, but further improvements may be needed to enhance accuracy, especially for emotions with limited training data.

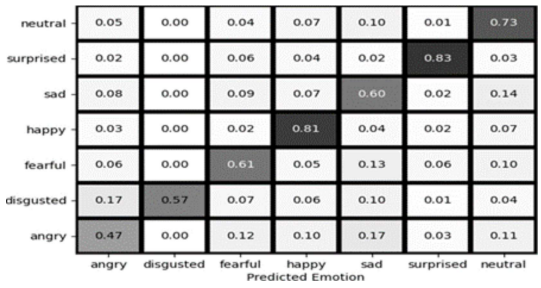

Fig. 7. *Confusion Matrix*

4. GESTURE RECOGNITION

4.1. MediaPipe Pose and Long Short-Term Memory

MediaPipe is an open-source framework by Google for building real-time multimedia processing pipelines. It simpli- fies multimedia app development with pre-trained models for tasks like image processing and pose estimation. ”MediaPipe Pose” is a key component using Google’s BlazePose model to track body poses in videos. It works on mobile devices, desk- tops, and web environments. The process involves detecting the person, predicting landmarks, and using a renderer. The pose model predicts 33 body key points.

In our video frame model, we use an LSTM-based deep Recurrent Neural Network (RNN). LSTM is an RNN archi- tecture for sequential data, incorporating feedback connections and gates for information flow control. It handles sequences like speech or video by using cells, input, output, and forget gates.

4.2. Proposed Solution

First, to create the training dataset, the approach involves using a sliding window technique with a width of n step time steps applied to the data sequence. With each step of the sliding window, a data point is created along with a corre-sponding output label value based on the object’s approach direction (Figure 8). As the data consists of consecutive time-based frames, the sliding step size is chosen to be 1 in this case.

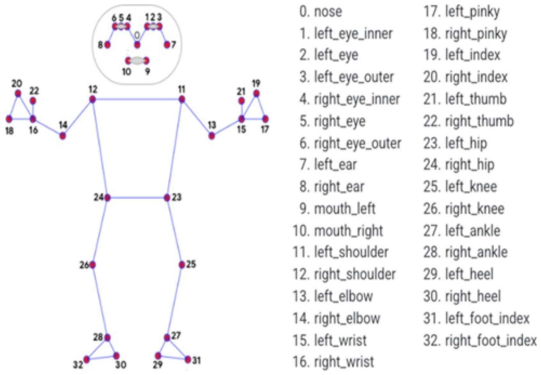

Fig. 8. *33 body pose key points*

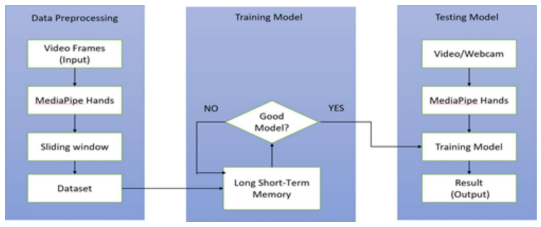

Fig. 9. *Confusion Matrix*

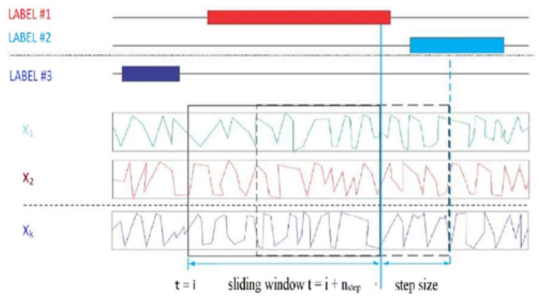

Fig. 10. *Creating Dataset*

The values within each step of the sliding window are recorded in the X file, while the corresponding label values are recorded in the Y file. The data is divided into two sets, namely the training set and the test set, with a split ratio of approximately 80% for the training set and 20% for the test set. These two sets are completely independent, with no overlapping data points. This ensures

objectivity when evaluating the model after the training process.

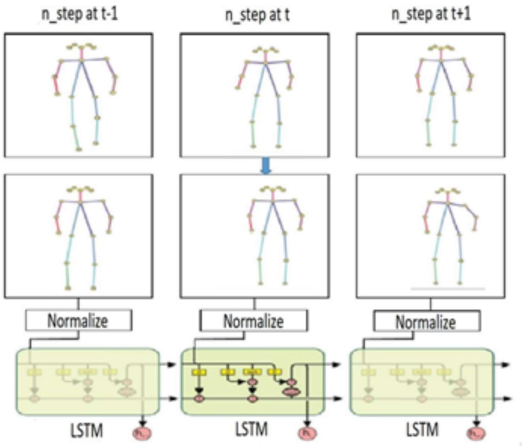

Fig. 11. *Model Training Process*

The posture of an individual contains information about their intention for interaction. We employ LSTM networks to observe the posture of a person over n- time steps to represent the interaction between humans and robots (Figure 9). Here, we construct a simple LSTM network to train sample data files using the available hardware. However, to diversify the training environment, we also utilize the existing MediaPipe Pose LSTM network and train data from the internet, obtaining data files for further use.

The human posture is extracted from MediaPipe. A posture comprises the 2D coordinates of j key points on the body, in- cluding 33 key points. Each primary point has two coordinates. Therefore, at each time step, we have a coordinate vector.

$x = [x1, x2, ..., xk]$ with $xi \in R, i = 1, ..., k, \quad k = 2j$

The input of the LSTM network is an X matrix of dimensions n steps x k, while the output represents the cases of human-robot interaction intentions, denoted as n case, and presented as a one-hot vector. After 100 epochs of training, model achieving an accuracy result at an approximate threshold of 90% is considered relatively good.

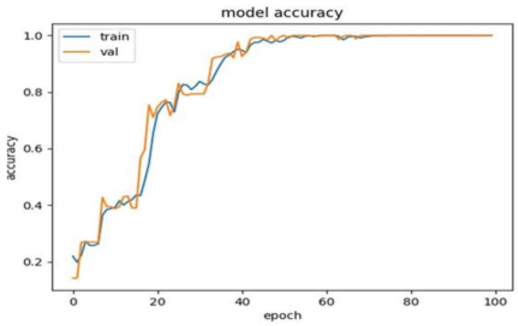

Fig. 12. *LSTM Model Training Result*

5. SIMULATING AND EXPERIMENT RESULT

To imbue a robot with the capacity to perceive human emo- tions and gestures, we propose an integrated approach. Real- time facial image capture is realized through C++ program- ming interfacing with Astra camera. The feature extraction and classification processes are executed within the Robot Operating System (ROS), optimizing interaction efficiency. Additionally, an LCD screen serves as an operational interface, displaying live images captured by the robot's embedded Astra camera. This interface showcases emotion and gesture recognition outcomes, allowing seamless human-robot com- munication.

This multifaceted framework empowers the robot to respond adeptly to both emotions and gestures, fostering more intuitive and engaging interactions. Our research endeavors encompass the practical implementation and efficacy assessment of this approach. By amalgamating feature extraction, classification algorithms, and real-time facial image acquisition using C++ with Astra camera, all integrated within ROS, we strive to create a sophisticated emotion and gesture recognition system. The LCD augments user experience by presenting live facial images alongside emotion and gesture recognition results. This synthesis enables a seamless exchange of information between humans and robots, propelling the realm of human-robot interaction toward greater sophistication and responsiveness.

Demo Video Link: https://youtu.be/NnI-r4oxvl8.

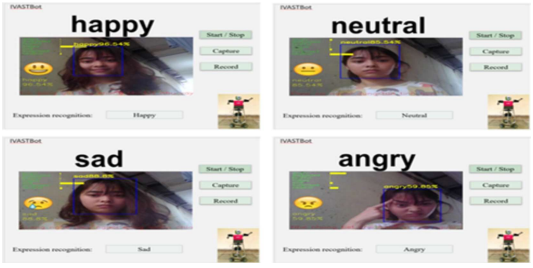

Fig. 13. *LSTM model Accuracy Result*

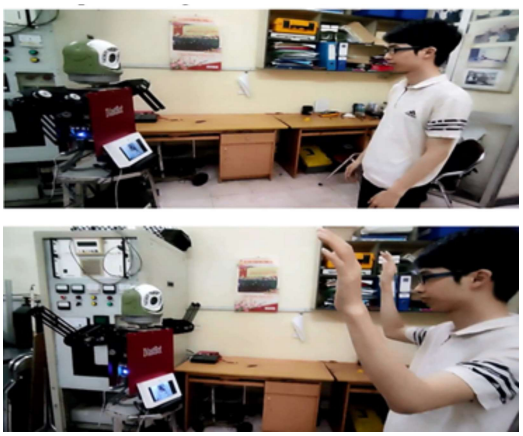

Fig.14. *IVASTBot Gesture Recognition Result*

CONCLUSION

In this research, we have proposed and implemented an integrated software system for the humanoid robot IVastBot, aimed at enhancing intelligent interaction with humans. We employed the open-source MediaPipe Pose library in com- bination with LSTM networks to recognize and comprehend human gestures and behaviors, enabling the robot to perform appropriate interactive actions and creating a more immersive and lifelike experience.

Furthermore, we integrated an intelligent algorithm for human facial emotion recognition using a convolutional neural network (CNN) model along with traditional features. This system empowers the robot to perceive and understand human emotions through facial expressions. The robot is capable of reflecting corresponding emotions through simple expressions, thus fostering a

multidimensional and natural interaction ex- perience.

The results of our experimentation have demonstrated the feasibility and effectiveness of the developed system. IVastBot robot not only excels in identifying and responding to gestures and behaviors of humans but also discerns and expresses human emotions through facial analysis. This advancement elevates the robot's intelligence, adaptability, and friendliness in human communication, unveiling potential applications in the realm of social robotics and human-machine interaction.

REFERENCES

[1] Ramkumar Gandhinathan and Lentin Joseph, "ROS Robotics Project" Packt Publishing Ltd, 1181219, 2019.

[2] K. Qian, J. Niu, and H. Yang, "Developing a gesture-based remote human- robot interaction system using Kinect," Int. J. Smart Home, vol. 7, no. 4, pp. 203-208, Jul. 2013.

[3] Z. K. Wang, K. Mulling, M. P. Deisenroth, H. B. Amor, D. Vogt, B. Scho¨lkopf, and J. Peters, "Probabilistic movement modeling for intention

¨ inference in human-robot interaction," Int. J. Robot. Res., vol. 32, no. 7, pp. 841-858, Apr. 2013

[4] M. Awais and D. Henrich, "Human-robot interaction in an unknown hu- man intention scenario," in Proc. 11th Int. Conf. Frontiers of Information Technology, Washington, DC, USA, 2013, pp. 89-94.

[5] [5]. L. Zhang, M. Jiang, D. Farid, and M. A. Hossain, "Intelligent facial emotion recognition and semantic-based topic detection for a humanoid robot," Exp. Syst. Appl., vol. 40, no. 13, pp. 5160-5168, Oct. 2013.

[6] [Kam17] Patrik Kamencay, Miroslav Benco, Tomas Mizdos, and Roman Radil, "A New Method for Face Recognition Using Convolutional Neural Network", Digital Image Processing and Computer Graphics, Vol. 15, No. 4, pp.663-672, 2017.

[7] A. Shima and F. Azar, "Convolutional Neural Networks for Facial Expression Recognition," arXiv:1704.06756v1 [cs.CV], 2017.

[8] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv: pp. 1409-1556v6 [cs.CV], 2015.

[9] Ngô Mạnh Tiến, Nguyễn Mạnh Cường, Hà Thị Kim Duyên, Bùi Quang Tuấn, Trần Bá Hiến, Nguyễn Minh Đông, Đỗ Quang Hiệp, "Xây dựng Hệ thống Bản đồ hóa SLAM và ứng dụng điều hướng cho Robot đa hướng sử dụng bánh xe Mecanum dựa trên hệ điều hành ROS,"Hội thảo quốc gia lần thứ XXIV: Một số vấn đề chọn lọc của Công nghệ thông tin và truyền thông – Thái Nguyên, 13-14/12/2021.

[10] Nguyễn Thị Duyên, Ngô Mạnh Tiến, Hà Thị Kim Duyên, Bùi Quang Tuấn, Trần Bá Hiến, Nguyễn Minh Đông, Đỗ Quang Hiệp, "Xây dựng hệ điều hướng trên bản đồ, định vị SLAM cho Robot tự hành trong nhà kính nông nghiệp dựa trên hệ điều hành ROS,"Hội nghị - Triển lãm quốc tế lần thứ 4 về Điều khiển và Tự động hoá (VCCA), 2021.

[11] L. Xie, C. Scheifele, W. Xu, and K. A. Stol, ”Heavy-duty omnidi- rectional Mecanum-wheeled Robot for autonomous navigation: System development and simulation realization,” IEEE International Conference on Mechatronics (ICM), pp. 256-261, 2015.

[12] M. A. Abuzneid, A. Mahmood, “Enhance Human Face Recognition Using LBPH Descriptor, Multi-KNN, and BPNN”, IEEE Access, Vol.6, pp.20642-20651, 2018.

[13] Li Cuimei, Qi Zhiliang, Jia Nan and Wu Jianhua, “Human face detection algorithm via Haar cascade classifier combined with three additional clas- sifiers”, IEEE 13th International Conference on Electronic Measurement & Instruments, pp.43-487, 2017.

[14] Ekberjan Derman and Albert Ali Salah, “Continuous Real-Time Vehicle Driver Authentication Using Convolutional Neural Network Based Face Recognition”, 13th IEEE International Conference on Automatic Face & Gesture Recognition, 2018.

[15] Dương Thăng Long, Bùi Thế Hùng, “Một phương pháp nhận dạng khuôn mặt dựa trên mạng nơ ron tích chập”, Tạp chí Khoa học – Viện Đại học Mở Hà Nội (08/2019)1-20.

[16] Rudenko, A., et al., Human motion trajectory prediction: A survey. The International Journal of Robotics Research, 2020. 39(8): p. 895-935.

[17] K. Qian, J. Niu, and H. Yang, “Developing a gesture-based remote human-robot interaction system using Kinect,” Int. J. Smart Home, vol. 7, no. 4, pp. 203-208, Jul. 2013.

[18] Boon Giin Lee, Su Min Lee, “Smart Wearable Hand Device for Sign Language Interpretation System with Sensors Fusion”,2017.

[19] Wei Fang, Yewen Ding, Feihong Zhang, and Jack Sheng, “Gesture Recognition Based on CNN and DCGAN for Calculation and Text Output”, 2019.

[20] Paulo Trigueiros, Ferando Ribeiro, Luis Paulo Reis, “Vision Based Portuguese Sign Language Recognition System”, New Perspectives in Information System and Technologies, Volume 1, Advances in Intelligent System and Computing, Springer International Publishing Switzerland, 2014.

[21] F. M. Noori, B. Wallace, M. Z. Uddin, and J. Torresen, ”A robust human activity recognition approach using open pose, motion features, and deep recurrent neural

network," in Scandinavian conference on image analysis, 2019: Springer, pp. 299-310.

[22] C. Sawant, "Human activity recognition with openpose and Long Short- Term Memory on real time images," EasyChair, 2516-2314, 2020.

[23] M. Z. Uddin and J. Torresen, "A deep learning-based human activity recognition in darkness," in 2018 Colour and Visual Computing Sympo- sium (CVCS), 2018: IEEE, pp. 1-5.

[24] G. Hidalgo, Z. Cao, T. Simon, S.-E. Wei, H. Joo, andY. Sheikh, "OpenPose library," https://github.com/CMU-Perceptual-Computing-

Lab/openpose

[25] Cao, Z., et al., OpenPose: real-time multi-person 2D pose estimation using Part Affinity Fields. 2019. pp. 172-186.

[26] Hochreiter, S. and J.J.N.c. Schmidhuber, Long short-term memory. 1997. pp. 1735-1780

[27] Ramkumar Gandhinathan and Lentin Joseph, "ROS Robotics Project" Packt Publishing Ltd, 1181219, 2019.

[28] Alessandra Rossi, Maria Di Maro, Antonio Origlia, Agostino Palmiero, and Silvia Rossi, "A ROS Architecture for Personalised HRI with a Bartender Social Robot", 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2022).

[29] Ignazio Infantino, Agnese Augello, Umberto Maniscalto, Giovanni Pi- lato, Pietro Storniolo, Filippo Vella, "A Cognitive Architecture for Social Robots", 2018 IEEE 4th International Forum on Research and Technology for Society and Industry (RTSI).

[30] Youssef Mohamed, Se´verin Lemaignan, "ROS for Human-Robot Inter- action", 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 27 Dec 2020.

[31] Canadian Journal on Electrical and Electronics Engineering Vol.3, No 6, July 2012.

[32] Chaudhary, G., & Ohri, J. (2016), "3-DOF Parallel manipulator control using PID controller". 2016 IEEE 1st International Conference on Power Electronics, Intelligent Control and Energy Systems (ICPEICES).

[33] Kebria, P. M., Al-wais Saba, Abdi, H., & Nahavandi, S. (2016), "Kinematic and dynamic modeling of UR5 manipulator". 2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC).