International Journal of Engineering Research and Technology. ISSN 0974-3154, Volume 14, Number 8 (2021), pp. 829-834
© International Research Publication House. http://www.irphouse.com

A Scheme Based on Convolutional Networks for the Estimation of Emotional States in Service Robots from Body Postures

Fernando Martínez Santa[1], Holman Montiel Ariza[2] and Fredy H. Martínez S.[3]

Associates Professors, Facultad Tecnológica, Universidad Distrital Francisco José de Caldas, Bogotá D.C, Colombia.

[1]*ORCID:* 0000-0003-2895-3084, [2]*ORCID:* 0000-0002-6077-3510
[3]*ORCID:* 0000-0002-7258-3909

Abstract

The months of social isolation that most societies around the world have suffered have left many lessons, including the need to develop strategies for the remote care of lonely, sick, elderly, and children in the education process. One of the most promising strategies is structured around service robots. These robotic platforms are designed to support the human being, which is why aspects such as interaction, morphology, and autonomy take on special importance. The Nao robot from Aldebaran Robotics is one of the most famous service robots and has a specific academic version for research and education. It is a platform on which interaction and manipulation strategies can be explored in human-to-human environments. Notably, one of the keys to human-machine interaction lies in the machine's ability to identify a person's emotional state and interact accordingly. The robots use different strategies to recognize these states, including facial parameters, tone of voice or body postures. In this article, we propose a scheme for the automatic identification of emotional states from the user's body posture in front of the robot. We use as a robotic platform the robot Nao V5, and as a strategy for the design of the identification model, we utilize a convolutional neural network trained for six specific emotional states in real operating environments (indoor environments similar to a home). In the laboratory tests, we obtained an accuracy above 80%, which allows us to validate the success of the strategy.

Keywords: *Deep learning, emotion recognition, gesture, human-robot interaction, posture, service robotics.*
.

1. INTRODUCTION

The success of a service robot in carrying out its task (caring for people, for example) lies largely in the level of integration of the machine with the human being. This is one of the reasons for which the service robots have a friendly morphology for the human being, many times anthropomorphic, with the intention that the human being feels comfortable with the robot (Martínez, Acero, & Castiblanco, 2015). The integration can be programmed in the robot of two forms different, some gestures can be defined a priori to the way of code that the robot it is capable of to decode when to interpret the signal (Medina-Catzin, Martin-Gonzalez, Brito-Loeza, & Uc-Cetina, 2017), but also is possible to program strategies of identification of gestures and postures characteristic of the being human, which exhibit under certain conditions (Cho & Jeong, 2017). In Human-Robot Collaboration (HRC) theory this is defined as structured forms of communication between members (between the robot and the user), but in the case of a service robot, it is expected that the robot will adapt and interpret the current conditions of its user (Callens, Van Der Have, Rossom, De Schutter, & Aertbelien, 2020; Mohammadi, Rezayati, van de Venn, & Karimpour, 2020). In the first case, it is sought a general strategy of control in which it is achieved to control the behavior of the robot when it is shown the gestures, in the second case it is sought a behavior more complex of the robot in which the robot defines its actions with the user according to what it can identify autonomously in his movements.

The robot can be programmed to identify human gestures, which are in fact communication mechanisms. For example, the robot can identify gestures made with hands and fingers (Martínez, Betancourt, & Arbulú, 2020), facial expressions (Martinez, Hernández, & Rendón, 2020), head movements (Olade, Fleming, & Liang, 2020; Yunardi, Dina, Agustin, & Firdaus, 2020), and body postures (Obaid et al., 2016). The objective is to interpret and to ponder autonomously and in time real such movements through some algorithm fed with some sense of this information. However, for effective integration, these gestures must be interpreted with a high rate of precision and respond in coherence with response times similar to human times (real-time operation) (H. Liu & Wang, 2018).

The processes of identification, tracking, and sensing of the event are carried out by the control unit from the raw information provided by the sensors. There are different strategies and technologies for the detection of these events, however, they can all be grouped into two categories, image-based and non-image-based sensor systems (H. Liu & Wang, 2018). In the first group is the strategy more similar to the biological scheme, i.e., one that tries to replicate the information captured by the human eye. In principle, it is of supposing that to interact with humans the robots should have similar optical capacities, which means a stereoscopic system that captures luminous images produced by the reflection of the robot user and its environment. Even so, depending on the capacity of the robot, exist different strategies for sensing based on images.

The first strategy of simple structure uses a camera and a processing algorithm to detect specific markers in the user's body (Fang, Zheng, & Wu, 2017; Moreno & Páez, 2017). This strategy, however, is expensive, intrusive, and complex to use (Takano, Ishikawa, & Nakamura, 2015). Also, with a single camera, it is impossible to identify depth-related information, which is why schemes with two cameras (stereoscopic system) were developed (Elhayek et al., 2017). With two cameras it is possible to reconstruct a three-dimensional system and identify more robustly the user's gestures. Some stereoscopic systems also use markers, but the norm today is to use digital image processing to detect specific parameters from which to estimate the user's posture, which is why these schemes are complex in terms of processing and calibration. Today there are systems on the market based on a single camera (or sensor) capable of measuring the depth in the scene (Tan et al., 2020). Depth sensors generally have no calibration or lighting problems and produce three-dimensional depth information much easier to process than images from a stereo system. Sensors such as Microsoft Kinect 2 allow implementing gesture recognition systems relatively easily and economically (B. Wang, Li, Lang, & Wang, 2020). Besides, the fact that it is a commonly used entertainment device has facilitated its wide diffusion (Kassim et al., 2020).

Most gesture recognition strategies use images. However, some new strategies use MEMS (Micro Electromechanical Systems) and other non-image-based schemes. An example of this type of sensor is the Myo armband that uses electromyo-gram sensors to read the electrical activity of the arm and wirelessly transmit this information (Barona et al., 2020; Martínez, Jacinto, & Montiel, 2019). Other strategies use gloves equipped with accelerometers and gyroscopes (Li et al., 2020; Pan et al., 2020), and even non-wearable elements such as Google's Project Soli that uses radiofrequency signals (S. Wang, Song, Lien, Poupyrev, & Hilliges, 2016), or MIT's WiTrack and RF-Capture system that uses radiofrequency signals reflected by the human body (Z. Wang, Xiao, Ye, Wang, & Yang, 2017).

Gesture recognition is performed from the sensed data. This recognition can be separated into two stages, in the parameterization of characteristics extracted from the raw data, and in the categorization of the gesture (Machado, Luísa Gomes, Gamboa, Paixão, & Costa, 2015). These stages are performed by applying filters, machine learning algorithms, and skeleton models (Liang, Chen, Wu, Yan, & Huang, 2020). However, the strategy depends on the type of data collected by the sensors. In terms of performance, identification strategies based on depth data are less computationally costly and with a higher level of accuracy than those based on images, whether monocular or stereoscopic (Barrero, Robayo, & Jacinto, 2015; H. Liu & Wang, 2018). Most skeletal models simplify the structure of the human body and use the concept of depth (Do, Kim, Yang, & Lee, 2020; X. Liu et al., 2020; Sapiński, Kamińska, Pelikant, & Anbarjafari, 2019).

2. MATERIALS AND METHODS

A The main means by which human beings' express emotions is the face, which is why most automatic emotion identification schemes process images of the user's face. However, in the real interaction between human beings it is not always possible to have access to the subject's face, he may be looking in another direction, or his face may be partially hidden with accessories such as glasses or medical face mask. This is the reason why secondary support schemes are implemented that seek other types of information, such as the user's voice or posture. Through the user's body language, it is also possible to establish his emotional state, information that can be paralleled with other schemes to increase reliability. Body language tends to be specific to each individual, and the variability increases, even more, when regional, cultural, age, and gender factors are considered. For this reason, this identification scheme should be used in parallel with others based on other parameters (face, voice, etc.), and the algorithm should always be trained within the social and cultural group for which the service robot is intended to be developed.

The social group selected in our research corresponds to family groups in homes with parents, children, and elderly in middle and upper-class homes in an urban Latin American city. Under this social delimitation, it is possible to characterize the possible users of the robot, and therefore define the group of actors that will represent the emotional states that will train our classification model. With these characteristics, actors were selected in the age ranges from 8 to 15 years old for the group of children (both sexes), in the age ranges from 25 to 40 years old for the group of adults (both sexes) and over 60 years old for the group of elderly people (both sexes).

The emotional states selected to conform the database were defined according to the six basic emotions defined by Ekman and Freisen (Michalik & Kucharska, 2020):

1. Anger
2. Disgust
3. Fear
4. Happiness
5. Sadness
6. Surprise

These names and numbers were also used as labels for each category. The actors were instructed to represent these emotional states naturally. From there we captured a total of 150 images for each of the categories, in total 900 images shaped the final database (see Figure 1). No discrimination by race, gender, or age was made. The distance between the actor and the camera was also not kept constant. The intention is that the identification model detects characteristics common to each emotional state independently of these factors.

Many of the human posture identification schemes consider only the upper body, keep the distance from the actor constant, and instruct the actors to perform the movements in a certain way. In this research, the user's entire body is considered, and on some occasions, the actors lie on the floor to represent the emotional state. Besides, the actor was given complete freedom in terms of location and movements to perform (some were even expressed verbally during the performances). These characteristics of the images guarantee less bias concerning the expected results and provide more information to the classification model.

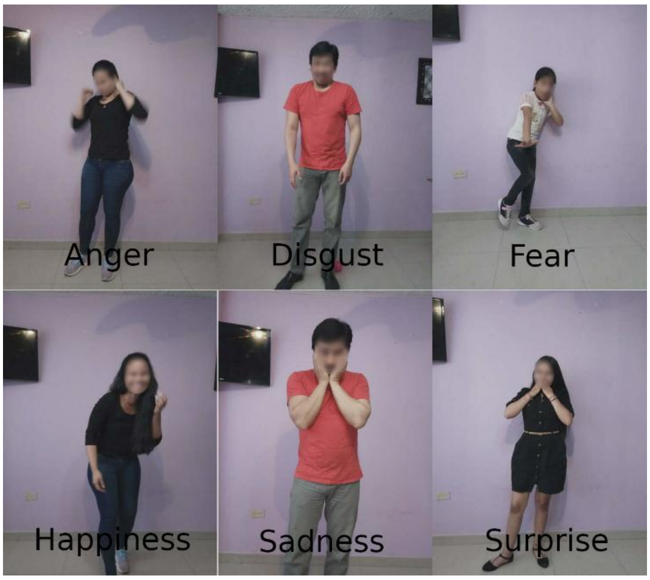


Fig. 1. Sample of the images used for the training of the model. The six categories used in the classification were: 1 anger, 2 disgust, 3 fear, 4 happiness, 5 sadness, and 6 surprise

The gesture recognition process can be characterized as a six-step system: identification of the instant in which the gesture occurs, user tracking, sensing the gesture, extracting characteristics from the sensing information, categorizing the gesture, and defining the response from the identified gesture (see Figure 2). In this research, we focus on the fourth and fifth stages of this process.

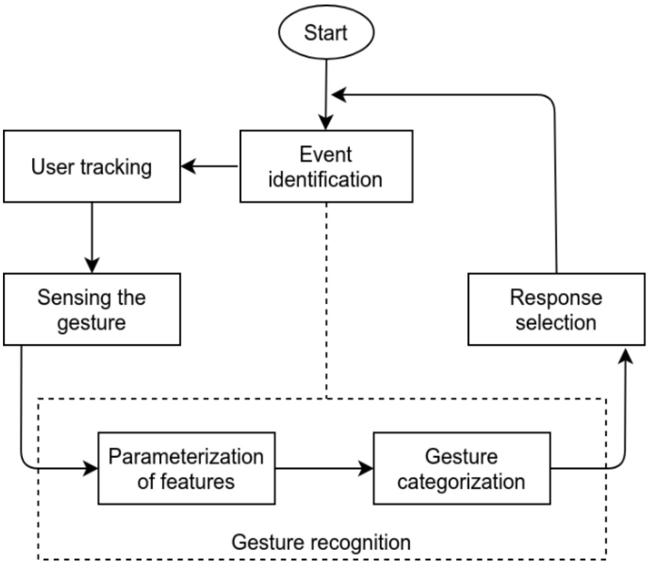


Fig. 2. Gesture recognition scheme for service robot

For our processes of parameterization of features and gesture categorization, we propose the use of a convolutional neural network. The neural model is trained with 80% of our database and then validated with the remaining 20%. In the output layer, we use a softmax function, or normalized exponential function, which performs a categorical distribution by selecting one of the outputs as the most likely category to which the input image belongs. As convolutional topology, we use the DenseNet (Densely Connected Convolutional Network), specifically a DenseNet121 network. This convolutional network was selected because it provides the highest performance with the least number of trainable parameters. The architecture of this deep network is based on simplifying the connection parameters between layers. Instead of adding the output characteristics of a layer with the input ones, it concatenates them (add dimension, but not add values), which dramatically reduces the number of parameters required by the network without decreasing its performance (avoids learning the same characteristic multiple times). In DenseNet there are blocks in which additional connections are included between the layers near the inputs and the layers near the output (Figure 3).

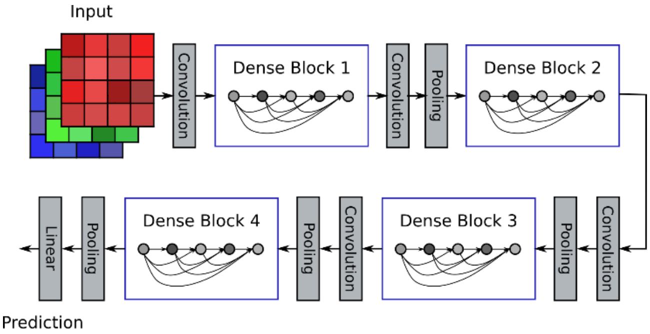


Fig. 3. Architecture of DenseNet121

3. RESULTS AND DISCUSSION

For the construction of the model, we use Keras 2.4.3 as a high-level API for TensorFlow 2.3.0. The database images were randomly mixed in the list and resized to a size of 256x256 pixels. During model fit testing, it was observed that such a reduction in size and aspect ratio did not significantly affect the model's performance, but it did significantly reduce the memory requirements for training. The pre-processing of the images was done with OpenCV 4.1.2, and all the code was developed in Python 3.

The images were normalized in the range of 0 to 1, and the labels were coded in one-hot format. The number of nodes in the input layer was defined according to the final size of the images, 256x256x3 = 196,608. The model was adjusted until a good behavior with the validation data was obtained, and no pre-trained weights were used. A total of six nodes were defined in the output layer, the number of output classes in the model. As loss function, we used categorical cross-entropy, and the optimization was performed with the stochastic gradient descent (SGD). During the training, accuracy, and MSE (Mean Squared Error) metrics were calculated in each cycle for both validation and training data. The depth of the network was 121 layers, with a total of 7,043,654 parameters, of which 6,960,006 correspond to trainable parameters. The final model was trained during 40 epochs, at the end a 100% accuracy for the training data and 76.1% for the validation data was achieved. The overall behavior of this training is shown in Figure 4.

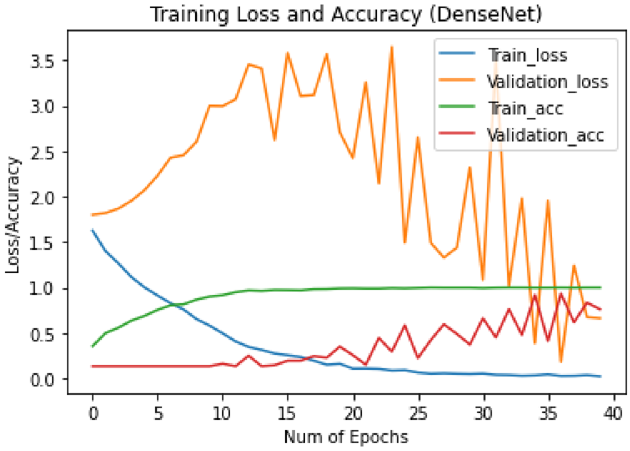
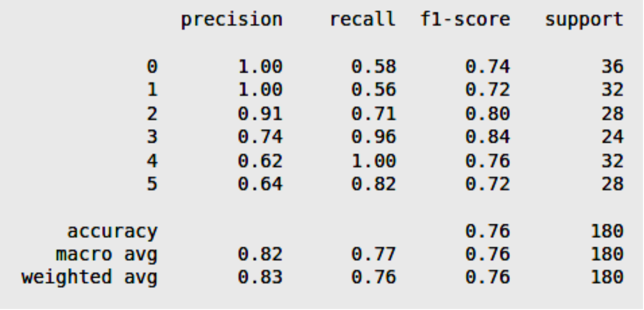


Fig. 4. Model behavior with training data and validation data

The error in the training data practically disappears after 30 epochs, at the same time it is reduced for the validation data, although the trend, in this case, is not continuous. The same behavior is observed in the precision of the model, after 30 epochs the precision of the training data is 100%, while the validation data is increasing continuously from epoch 20 to reach an average value of 82%. The tuned model was evaluated in each category with the metrics precision, recall, and f1-score (see Figure 5). For the first two categories (Anger and Disgust) 100% precision was achieved with the validation data, the lowest value was presented in the Sadness category which reached only a value of 62%. However, this last category achieved a value of 100% in the recall, which indicates that the gestures classified in this category correspond to the Sadness category. The recall in the Anger and Disgust categories was the lowest (58% and 56%), which means that despite classifying the correct images in this category, it also included images that do not correspond to it. The F1-score value combines these two metrics, the combined performance confirms the previous analysis, 74% and 72 for the Anger and Disgust categories, and 76% for the Sadness category.



```
              precision    recall  f1-score   support

           0       1.00      0.58      0.74        36
           1       1.00      0.56      0.72        32
           2       0.91      0.71      0.80        28
           3       0.74      0.96      0.84        24
           4       0.62      1.00      0.76        32
           5       0.64      0.82      0.72        28

    accuracy                           0.76       180
   macro avg       0.82      0.77      0.76       180
weighted avg       0.83      0.76      0.76       180
```

Fig. 5. Behavior of model metrics for unknown data (validation data)

The results of the precision, recall, and F1-score metrics can be seen graphically in the confusion matrix (see Figure 6). The lighter colors are assigned to the higher values, reflecting that the diagonal of the matrix has excellent behavior, particularly for the fifth category. Again, the number of false categorizations looks relatively high for the first two categories, even though all the elements in the category are correctly categorized. The ROC curve is another graphical representation of the model's performance, but in this case, we see the sensitivity in each category to the specificity of the classifier (Figures 7 and 8). True positives versus false positives again are high for all categories (see Figure 7), for the first two categories (see Figure 8).

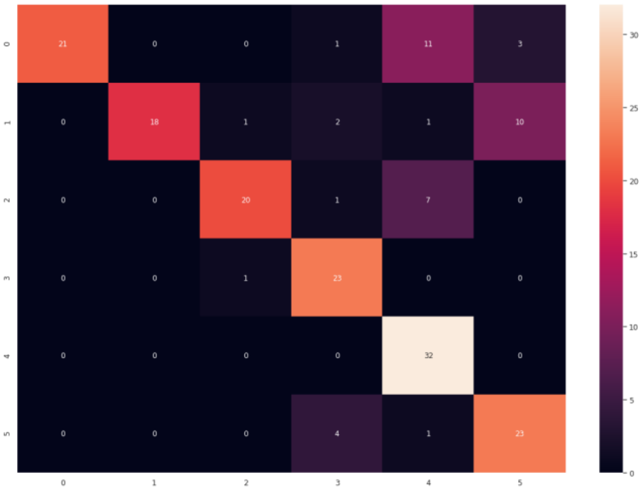


Fig.6. Confusion matrix (error matrix)

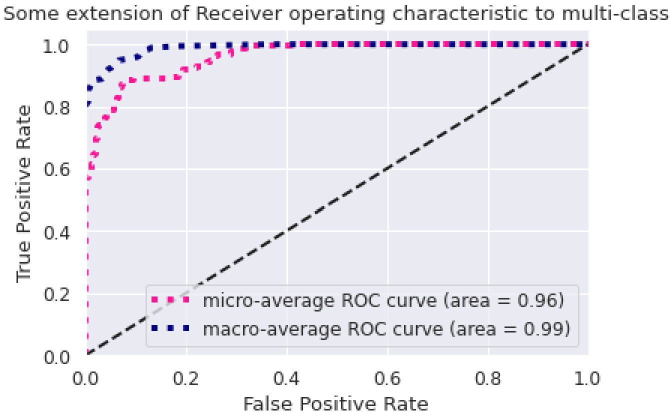


Fig. 7. ROC curve

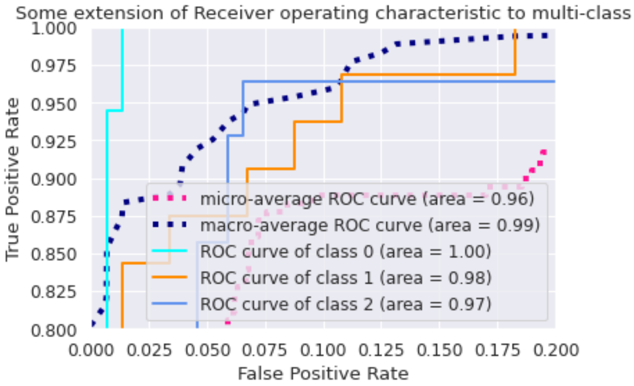


Fig. 8. ROC curve (zoom of the upper left corner)

4. CONCLUSIONS

This paper presented a strategy based on the DenseNet (Densely Connected Convolutional Network) for the identification of six specific emotional states based on the body posture of the respective users; as a loss function, categorical cross-entity was used, and optimization was performed with stochastic gradient decline (SGD). The main objective of the research was focused on evaluating a functional model of classification of high performance and reduced size against the training parameters of the neuronal network, this in order to be implemented in service robots and various applications in real environments that allow improving the human-robot interaction. A total of 900 images from the test database were used for the various emotional states with a size of 256x256 pixels. The resulting classification model was trained during 40 periods, achieving at the end a 100% accuracy for the training data and a percentage higher than 80% for the validation data (average value), these results demonstrate the effectiveness of the proposed model.

In our research, we have not yet included a study of the possible effects on the performance of the model, such as the sex of the user, age, height or distance between the robot and the user, or other accessory elements such as the use of hats, glasses, or medical face mask. These are aspects that deserve further analysis and will be addressed as future focuses of our research.

5. ACKNOWLEDGMENTS

This work was supported by the Universidad Distrital Francisco José de Caldas, in part through CIDC, and partly by the Facultad Tecnológica. The views expressed in this paper are not necessarily endorsed by Universidad Distrital. The authors thank the research group ARMOS for supporting the development of the code.

REFERENCES

Barona, L., Valdivieso, A., Vimos, V., Zea, J., Vásconez, J., Álvarez, M., et al. (2020). An Energy-Based Method for Orientation Correction of EMG Bracelet Sensors in Hand Gesture Recognition Systems. Sensors (Basel, Switzerland), 20(21).

Barrero, A., Robayo, M., & Jacinto, E. (2015). Algoritmo de navegación a bordo en ambientes controlados a partir de procesamiento de imágenes. Tekhnê, 12(2), 23–34.

Callens, T., Van Der Have, T., Rossom, S., De Schutter, J., & Aertbelien, E. (2020). A Framework for Recognition and Prediction of Human Motions in Human-Robot Collaboration Using Probabilistic Motion Models. IEEE Robotics and Automation Letters, 5(4), 5151–5158.

Cho, M., & Jeong, Y. (2017). Human gesture recognition performance evaluation for service robots. In 2017 19th International Conference on Advanced Communication Technology (ICACT) (pp. 847–851).

Do, N., Kim, S., Yang, H., & Lee, G. (2020). Robust hand shape features for dynamic hand gesture recognition using multi-level feature LSTM. Applied Sciences (Switzerland), 10(18).

Elhayek, A., De Aguiar, E., Jain, A., Thompson, J., Pishchulin, L., Andriluka, M., et al. (2017). MARCOnI - ConvNet-Based MARker-less motion capture in outdoor and indoor scenes. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(3), 501–514.

Fang, W., Zheng, L., & Wu, X. (2017). Multi-sensor based real-time 6-DoF pose tracking for wearable augmented reality. Computers in Industry, 92-93, 91–103.

Kassim, M., Mohd, M., Tomari, M., Suriani, N., Zakaria, W., & Sari, S. (2020). A non-invasive and non-wearable food intake monitoring system based on depth sensor. Bulletin of Electrical Engineering and Informatics, 9(6), 2342–2349.

Li, Y., Zheng, C., Liu, S., Huang, L., Fang, T., Li, J., et al. (2020). Smart Glove Integrated with Tunable MWNTs/PDMS Fibers Made of a One-Step Extrusion Method for Finger Dexterity, Gesture, and Temperature Recognition. ACS Applied Materials and Interfaces, 12(21), 23764–23773.

Liang, S., Chen, J., Wu, Y., Yan, S., & Huang, J. (2020). Recognition of Subtle Gestures by 2-Channel sEMG Using Parameter Estimation Classifiers Based on Probability Density. IEEE Access, 8, 169835–169850.

Liu, H., & Wang, L. (2018). Gesture recognition for human-robot collaboration: A review. International Journal of Industrial Ergonomics, 68, 355–367.

Liu, X., Shi, H., Hong, X., Chen, H., Tao, D., & Zhao, G. (2020). 3D Skeletal Gesture Recognition via Hidden States Exploration. IEEE Transactions on Image Processing, 29, 4583–4597.

Machado, I. P., Luísa Gomes, A., Gamboa, H., Paixão, V., & Costa, R. M. (2015, March). Human activity data discovery from triaxial accelerometer sensor: Non-supervised learning sensitivity to feature extraction parametrization. Information Processing & Management, 51(2), 204–214.

Martínez, F., Acero, D., & Castiblanco, M. (2015). Robótica autónoma: Acercamiento a algunos problemas centrales. Universidad Distrital Francisco José de Caldas.

Martínez, F., Betancourt, F., & Arbulú, M. (2020). A gesture recognition system for the Colombian sign language based on convolutional neural networks. Bulletin of Electrical Engineering and Informatics, 9(5), 2082–2089.

Martinez, F., Hernández, C., & Rendón, A. (2020). Identifier of human emotions based on convolutional neural network for assistant robot. TELKOMNIKA (Telecommunication Computing Electronics and Control), 18(3), 1499–1504.

Martínez, F., Jacinto, E., & Montiel, H. (2019). Neuronal Environmental Pattern Recognizer: Optical-by-Distance LSTM Model for Recognition of Navigation Patterns in Unknown Environments. In Y. Tan & Y. Shi (Eds.), Data Mining and Big Data (pp. 220–227). Singapore: Springer.

Medina-Catzin, J., Martin-Gonzalez, A., Brito-Loeza, C., & Uc-Cetina, V. (2017). Body Gestures Recognition System to Control a Service Robot. International Journal of Information Technology and Computer Science, 9(1), 69–76.

Michalik, K., & Kucharska, K. (2020). Implementation of Deep Neural Networks in Facial Emotion Perception in Patients Suffering from Depressive Disorder: Promising Tool in the Diagnostic Process and Treatment Evaluation. In K. Arai, S. Kapoor, & R. Bhatia (Eds.), Intelligent Computing (pp. 174–184). Cham: Springer International Publishing.

Mohammadi, F., Rezayati, M., van de Venn, H., & Karimpour, H. (2020). A Mixed-Perception Approach for Safe Human-Robot Collaboration in Industrial Automation. Sensors (Basel, Switzerland), 20(21).

Moreno, A., & Páez, D. (2017). Performance evaluation of ROS on the Raspberry Pi platform as OS for small robots. Tekhnê, 14(1), 61–72.

Obaid, M., Sandoval, E., Złotowski, J., Moltchanova, E., Basedow, C., & Bartneck, C. (2016). Stop! That is close enough. How body postures influence human-robot proximity. In 2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN) (pp. 354–361).

Olade, I., Fleming, C., & Liang, H. (2020). Biomove: Biometric user identification from human kinesiological movements for virtual reality systems. Sensors (Switzerland), 20(10).

Pan, J., Luo, Y., Li, Y., Tham, C.-K., Heng, C.-H., & Thean, A.-Y. (2020). A Wireless Multi-Channel Capacitive Sensor System for Efficient Glove-Based Gesture Recognition with AI at the Edge. IEEE Transactions on Circuits and Systems II: Express Briefs, 67(9), 1624–1628.

Sapiński, T., Kamińska, D., Pelikant, A., & Anbarjafari, G. (2019). Emotion recognition from skeletal movements. Entropy, 21(7).

Takano, W., Ishikawa, J., & Nakamura, Y. (2015). Using a human action database to recognize actions in monocular image sequences: Recovering human whole-body configurations. Advanced Robotics, 29(12), 771–784.

Tan, C., Sun, Y., Li, G., Jiang, G., Chen, D., & Liu, H. (2020). Research on gesture recognition of smart data fusion features in the IoT. Neural Computing and Applications, 32(22), 16917–16929.

Wang, B., Li, Y., Lang, H., & Wang, Y. (2020). Hand gesture recognition and motion estimation using the Kinect sensor. Mechatronic Systems and Control, 48(1), 17–24.

Wang, S., Song, J., Lien, J., Poupyrev, I., & Hilliges, O. (2016). Interacting with Soli: Exploring Fine-Grained Dynamic Gesture Recognition in the Radio-Frequency Spectrum. In Proceedings of the 29th Annual Symposium on User Interface Software and Technology (pp. 851–860). New York, NY, USA: Association for Computing Machinery.

Wang, Z., Xiao, F., Ye, N., Wang, R., & Yang, P. (2017). A See-through-Wall System for Device-Free Human Motion Sensing Based on Battery-Free RFID. ACM Transactions on Embedded Computing Systems, 17(1), 6:1–6:21.

Yunardi, R., Dina, N., Agustin, E., & Firdaus, A. (2020). Visual and Gyroscope Sensor for Head Movement Controller System on Meal-Assistance Application. Majlesi Journal of Electrical Engineering, 14(3), 39–44.