International Journal of Applied Engineering Research ISSN 0973-4562 Volume 13, Number 22 (2018) pp. 15872-15877

Mouth States Recognition System Focused on Feeding task

Javier Orlando Pinzón-Arenas[1], Robinson Jiménez-Moreno[2], Astrid Rubiano-Fonseca[3]

Department of Mechatronics Engineering, Nueva Granada Military University, Bogotá, Colombia.

Abstract

This paper presents the implementation of a Long-Short Term Memory (LSTM) network oriented at recognizing states of the mouth focused on feeding assistance, which are "Chewing" and "Waiting". In order to take into account the variations of chewing that may occur in different users, it is decided to create 2 sub-states called "Closed" and "Intermittent between open and closed", belonging to the general state "Chewing". To develop this work, a database was created with a total of 1731 sequences taken from different users, which is divided for training and validation. Each sequence contains 50 time steps with 6 distance relationships of the mouth that were acquired through 12 characteristic points obtained through the OpenFace software. With this database, 6 variations of the proposed network are trained, obtaining a final network with an accuracy of 99.3% in the recognition of the two main states. In addition, a comparison is made with a first approximation done using a combined algorithm of Viola-Jones and morphological operations as a technique for extracting features from the mouth, demonstrating the robustness of the complete system developed in this work against changes in environment and face rotation.

Keywords. LSTM Network, Mouth States, Sequence-to-label network, face landmarks.

INTRODUCTION

The recognition of movements of the human body plays an important role in the human-machine interaction, because by means of these, the robot can follow orders, manage its workspace or provide a service to a person. The movements that a person can do to perform this interaction can be static, i.e. invariant in time or that only depend on being captured only once, as are certain gestures of the hand [1] or expressions of the face [2]. Or they can be dynamic, in other words, depending on a sequence of steps to be completely identified, for example compound gestures of sign language [3] or movement of the body to know its trajectory or behavior [4, 5].

Various techniques have been developed to do the recognition of human actions. Within the static ones, the work shown in [6] is presented, where a convolutional neural network with a DAG architecture is used to perform hand signal or gesture recognition to control a mobile manipulator. In [7], a fuzzy support vector machine model is implemented to recognize 7 types of expressions in different subjects, where static data of the facial features extracted by means of the wavelet transform are used. On the other hand, for the recognition of dynamic actions, works as the one presented in [8], by means of depth movement maps and support vector machines, sequences of movements of the whole body are recognized. Another example is given in [9], where several human actions are recognized using video sequences of people executing each movement, they are entered into a convolutional network specialized in learning spatio-temporal features, achieving results above 85% with 101 categories.

Another of the techniques implemented for the recognition of actions are artificial neural networks called Long-short term memory (LSTM) networks [10], that are a special type of recurrent neural network with the ability to learn sequences of a large number of time steps, having as network input arrangements of data of the movements made or behavior of people. An example of application of this network can be seen in [11], where a Spatio-Temporal LSTM is used to recognize human actions in a 3D environment using data obtained by means of a human skeleton algorithm, achieving accuracies greater than 93%. Similarly, in [12] this type of network is used in combination with a convolutional neuronal network to recognize 7 emotions by means of face expressions through video sequences. However, although some works have recently been developed using this technique to recognize mouth gestures, such as the work presented in [13] oriented to the visual recognition of speech, the LSTM has not been used to recognize states of the mouth that focus on movements related to food, i.e. know if a person is chewing food or not, so that it can be applied to a robotic assistance system.

Taking into account the above, this work proposes to make use of an LSTM network to recognize 2 general states necessary in the task of assisted feeding, which are "chewing" and "waiting". However, when analyzing the way chewing of different people, it was opted to divide the state of "chewing" into 2 sub-categories for training purposes, called: "Closed", when the person tends to chew with his mouth closed, and "Intermittent", which is when the person chews by opening and closing the mouth, causing this state to have variations in its characteristics. The future purpose of this implementation is to be used in a robotic assistance system for feeding people.

This paper is divided into 4 sections, where section 2 presents the building of the database to be used and the proposed architecture for the implementation of the system. In section 3,

the results obtained from the training of the network and its comparison with the first approximation elaborated in a previous work are given. Finally, in section 4, the conclusions reached and future work are given.

METHODS AND MATERIALS

First, states to recognize the mouth are established, named as "chewing" and "waiting". However, since the state "chewing" can have long times and different forms of execution, as established in [14], it is decided to divide this state into 2 subcategories, called "Closed" and "Intermittent", where the last one is when the person chews with his mouth open. Likewise, "waiting" is renamed "Open".

In order to develop this work, once the states to be used have been established, a dataset must be built that contains the necessary data to be able to recognize these states of the mouth, and then propose the parameters of the network that is going to be used to carry out the recognition. Each of these steps is discussed below.

Dataset

To make the dataset, certain features of the mouth should be extracted in order to know the opening of this. In a first approximation, it was decided to perform an image processing using the algorithm of Viola-Jones and morphological operations, with which 2 types of data were obtained: opening ratio of the mouth and the height of the mouth in pixels (px) [14], calculated by means of the landmarks shown in Figure 2a. However, this processing depends on a system controlled in face positioning and light environment. For this reason, an extraction system was chosen that would allow a better characterization of the mouth and be robust with respect to facial movements and changes in the work environment. This system is called OpenFace [15], developed by Tadas Baltrušaitis, which has the ability to locate different landmarks on the entire face (including the mouth) through a neural network called MTCNN, as shown in Figure 1.

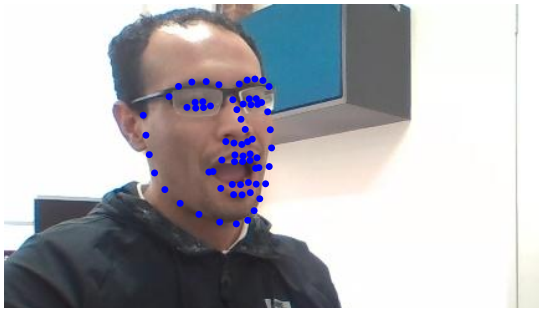
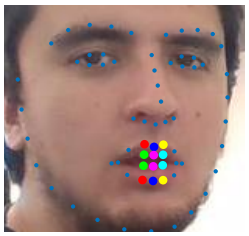


Figure 1. OpenFace landmarks example.

With the landmarks obtained, 12 points belonging to the mouth are taken, where 6 belong to the upper lip and 6 to the lower lip. A comparison between the two methods (the one used in the first approach and the one used in this work) is shown in Figure 2, being able to observe the landmarks obtained in each one for the same example image.

(a)



(b)

Figure 2. Landmarks obtained using a) Viola-Jones + Morphologic operations algorithm and b) OpenFace.

To capture the environment, an Intel® RealSenseTM SR300 camera is used, using a resolution of 1280x720 px. Also, the camera is located in a fixed way on the work table, at a distance of approximately 570 mm from the user, with a variation of ± 100 mm, since the users are not always located in the same part and the camera could also cover the bowl where the food that the user will receive is located, as shown in Figure 3. The data that make up the dataset with respect to the points obtained previously, are the internal and external distance between the upper and lower lip, in 3 different locations, having a total of 6 data related to each distance shown in Figure 2b, where each distance relationship is given with the points of the same color. This is done in order to avoid confusion due to face rotations, since when there is a big rotation of the face (greater than 45% with respect to the front of the camera), the distance of the reference points located in the direction of rotation may be affected, while those located on the opposite side would not have this problem, helping to maintain the distance relationship in at least a couple of points. Since the camera is fixed, pixels are used as distance measure. Taking this into account, video recordings from 30 users performing each of the required states are taken, with time length between 10 and 12 seconds.

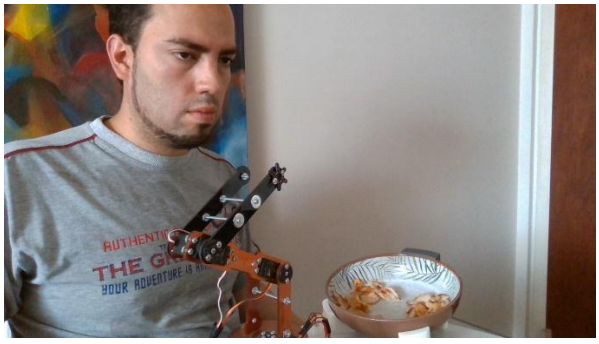


Figure 3. Workspace of the feeding robot assistant.

For the recognition of the state of the mouth, a sequence of movements of the mouth is required, because if each individual image is analyzed, when a person chews with his mouth open, this would be recognized as "Waiting" at certain moments of movement. This sequence consists of obtaining the data over time. To avoid using fixed sampling times, dynamic sequences are used that depend on the time of acquisition of the image and subsequent saving. For this, an Alienware R3 laptop with the characteristics shown in Table 1 is used. Additionally, to parameterize the amount of data obtained, a limit of 50 time steps (frames) in total per sequence is set. With the aforementioned, sequences between 1.3 and 2.2 seconds are obtained, of 50 frames each, with a total of 6 data per frame. In total, of the captured videos, 1731 sequences are obtained, of which 1435 are destined for training and 296 for validation. An example of each of the sequences can be observed in Figure 4, where it can be seen that the opening and closing of the mouth in the intermittency state achieves at least 2 complete cycles. Within the figure, it can be observed that the behavior of the external (Ext.D) and internal (Int.D) distances are similar to each other, however, due to the rotation of the face, they tend to have small variations in their measurement, which can be seen by comparing distances 1 against distances 3. Likewise, it is possible to see that the person to whom these example samples belong, when performing the intermittency state, does not tend to completely close the mouth, since the shortest distance is far from the distance measured in "Closed", therefore, if this subcategory had not been created, the sample could be confused with the status of "Open" or "Waiting".

Table 1. Hardware Specifications

Processor	i7-7700HQ
Generation	7th Generation Intel® Core™
Num. Cores	4
Frequency	2,80 GHz
RAM	32 GB DDR4
GPU	NVIDIA® GeForce® GTX 1070
GPU Memory	8 GB GDDR5
Multiprocessor Count	16
CUDA Cores	2048
Clock Rate	1695 MHz

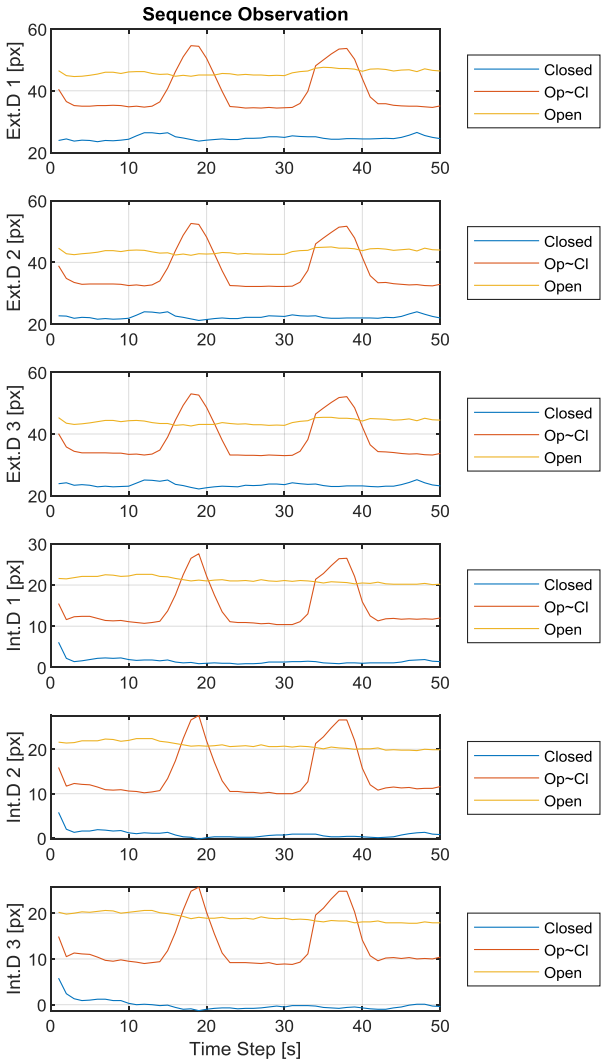


Figure 4. Sequence observations example.

Architecture proposed

With the dataset and number of categories to be recognized, it is proceeded to propose the architecture of the neural network to be used. Since each sequence has a large number of time steps, a network that does not have Long-Term dependencies should be used, so it is proposed to use a LSTM network. Also, as the purpose is to categorize a sequence in its entirety and not each of its steps, it is defined as a sequence-to-label network. Additionally, 6 features or data are set as input to the network, and 3 categories are output. The proposed architecture is shown in Figure 5.

RESULTS

Training results

An important parameter, which is necessary to be set, is the number of hidden units that the network should have. To define this parameter, several training sessions are performed, varying only the number of units. For this, it was decided to vary from 50 units (minimum amount required, since the sequence has a

maximum of 50 time steps) up to 100 units. Each variation is trained with a learning rate of 10-4 for 600 epochs, obtaining the results of training and validation of Table 2.

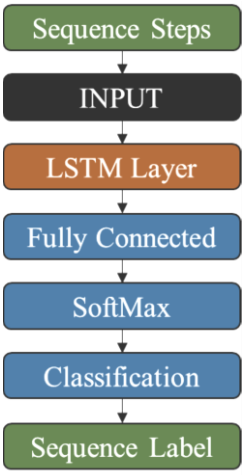


Figure 5. Architecture Proposed

Table 2. Training and validation accuracy of each variation of the network

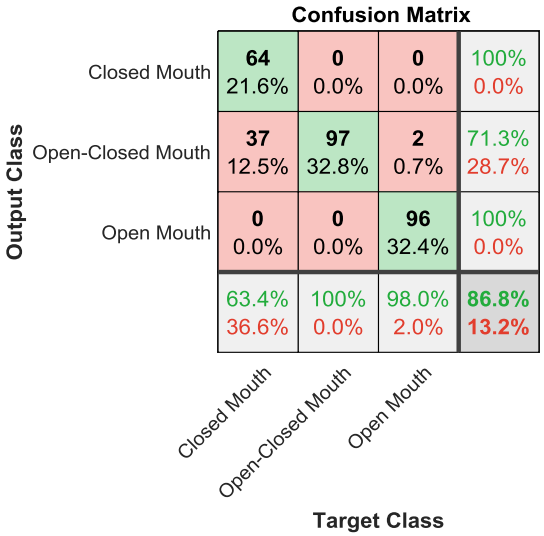
Network	Number of Hidden Units	Train		Test	
		3 Cat	*2 Cat*	*3 Cat*	*2 Cat*
Net_1	50	98,7%	99,3%	81,8%	99,0%
Net_2	60	97,7%	99,5%	**86,8%**	99,0%
Net_3	70	98,8%	99,8%	**86,8%**	**99,3%**
Net_4	80	99,2%	99,9%	82,8%	98,3%
Net_5	90	98,8%	99,8%	83,1%	99,0%
Net_6	100	96,8%	97,8%	82,1%	98,6%

As can be seen in the previous table, there are the final results of each training and validation for the recognition of 3 categories and for the union between the two subcategories of the "Chewing" state. In general terms, the networks achieved an accuracy of more than 97% within the training, however, the validation is below 90% for 3 categories. Nevertheless, as it is required to classify the two main categories, the overall accuracy with the merging surpass 98%, and, among all the networks, the one with the best performance is Net_3, with 70 hidden units, which achieved a 99.3% accuracy, therefore, this network is selected as the best architecture for this work.

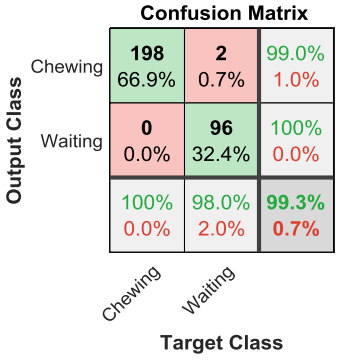
COMPARISON WITH PREVIOUS WORK

To do a more detailed analysis of Net_3 performance, a confusion matrix is made with the validation for 3 and 2 categories to see their behavior for each state, in addition to a comparison between it and the one performed in the previous work (called Net_A).

In Figure 6, it can be observed the confusion matrix of the Net_3, where for the classification of the 3 categories with which it was trained, it had its greatest confusion in the recognition of the closed mouth, classifying 36.6% of this category as “Intermittent” between closed and open (chewing with the mouth open), and may be caused by some actions of users when they performed that state, unconsciously opening their mouth at some point during the recording or moving their lips quickly. However, when merging between the two subcategories of the "Chewing" category, the network does not present mistakes in it, and confuses only 2 sequences of "Waiting" as “Chewing”.



(a)



(b)

Figure 6. Confusion matrix for a) 3 categories and b) the 2 main categories.

Comparing this network with the one implemented in the first approach, it can be seen that, although equal learning rates were used and the Net_A had to use 2 characteristics and only 15 time steps, the Net_A required a computational cost much higher than the Net_3, having to train it for 4500 epochs, as shown in Figure 7. On the other hand, even having a controlled

environment, the validation accuracy of the Net_A was lower than the Net_3, as shown in Table 3, showing the capacity of the Net_3 and the processing system used for the implementation of a system capable of recognizing the states of the mouth.

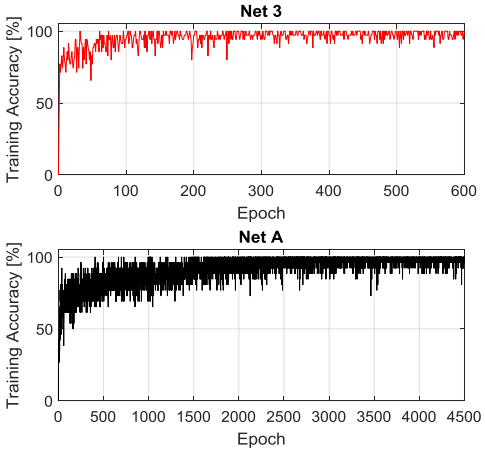


Figure 7. Training Behaviors.

Table 3. Performance of network 3 and network A.

	Net_3	**Net_A**
3 Categories	86.8%	84.8%
2 Categories	99,3%	97.9%

CONCLUSIONS AND FUTURE WORK

The implementation of a system capable of recognizing states of the mouth when chewing or waiting to receive food requires identifying the different ways in which people can perform these actions, for this reason, it is required to subdivide as much as possible the variations that may exist, as was done in this work, dividing the chewing state into two sub-states. Also, to achieve better recognition, it is necessary to obtain a sequence of movements with enough time steps to characterize it adequately, i.e. that within a sequence it is possible to have at least 2 repetitions of a state, however, this makes it necessary for a network capable of learning many steps back, so the LSTM network is adapted to the application. With this, it is possible to obtain a system of recognition of states of the mouth focused to feeding assistance, with a general accuracy of 99.3%.

This results not only depend on the training carried out on the network, but on the type of processing and possible features to be acquired from the mouth, since in the current work, it was possible to strengthen the processing system, through the use of the OpenFace application, acquiring a greater number of feature points and making the system insensitive to variations in the environment, compared to the previous work done, where only 2 points of the mouth were obtained and needed a very controlled environment. This not only allowed to observe the synergy of the two implemented processes (processing and recognition), but the capacity and reliability of the complete system.

As future work, this system is expected to be coupled with an assistance robot in a way that allows it to increase the autonomy of the robot when executing the task of granting food to users, for this reason, the robustness of the work presented in this paper is required.

REFERENCES

[1] Rautaray, S. S., & Agrawal, A. (2015). Vision based hand gesture recognition for human computer interaction: a survey. *Artificial Intelligence Review*, *43*(1), pp. 1-54. DOI: https://doi.org/10.1007/s10462-012-9356-9

[2] Bettadapura, V. (2012). Face expression recognition and analysis: the state of the art. *arXiv preprint arXiv:1203.6722*.

[3] Kurakin, A., Zhang, Z., & Liu, Z. (2012, August). A real time system for dynamic hand gesture recognition with a depth sensor. In *Signal Processing Conference (EUSIPCO), 2012 Proceedings of the 20th European,* pp. 1975-1979. IEEE.

[4] Dıaz, E. M. (2002). Theory of planned behavior and pedestrians' intentions to violate traffic regulations. *Transportation Research Part F: Traffic Psychology and Behaviour*, *5*(3), pp. 169-175. DOI: https://doi.org/10.1016/S1369-8478(02)00015-3

[5] Vemulapalli, R., Arrate, F., & Chellappa, R. (2014). Human action recognition by representing 3d skeletons as points in a lie group. In *Proceedings of the IEEE conference on computer vision and pattern recognition,* pp. 588-595. DOI: https://doi.org/10.1109/CVPR.2014.82

[6] Arenas, J. O. P., Moreno, R. J., & Beleño, R. D. H. (2018). Convolutional Neural Network with a DAG Architecture for Control of a Robotic Arm by Means of Hand Gestures. *Contemporary Engineering Sciences*, *11*(12), pp. 547-557. DOI: https://doi.org/10.12988/ces.2018.8241

[7] Zhang, Y. D., Yang, Z. J., Lu, H. M., Zhou, X. X., Phillips, P., Liu, Q. M., & Wang, S. H. (2016). Facial emotion recognition based on biorthogonal wavelet entropy, fuzzy support vector machine, and stratified cross validation. *IEEE Access*, *4*, pp. 8375-8385. DOI: https://doi.org/10.1109/ACCESS.2016.2628407

[8] Chen, C., Liu, K., & Kehtarnavaz, N. (2016). Real-time human action recognition based on depth motion maps. *Journal of real-time image processing*, *12*(1), pp. 155-163. DOI: https://doi.org/10.1007/s11554-013-0370-1

[9] Sun, L., Jia, K., Yeung, D. Y., & Shi, B. E. (2015). Human action recognition using factorized spatio-temporal convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision,* pp. 4597-4605. DOI: https://doi.org/10.1109/ICCV.2015.522

[10] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), pp. 1735-1780.

[11] Liu, J., Shahroudy, A., Xu, D., & Wang, G. (2016, October). Spatio-temporal lstm with trust gates for 3d human action recognition. In *European Conference on Computer Vision,* pp. 816-833. Springer, Cham. DOI: https://doi.org/10.1007/978-3-319-46487-9_50

[12] Fan, Y., Lu, X., Li, D., & Liu, Y. (2016, October). Video-based emotion recognition using CNN-RNN and C3D hybrid networks. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction,* pp. 445-450. ACM. DOI: https://doi.org/10.1145/2993148.2997632

[13] Petridis, S., Li, Z., & Pantic, M. (2017, March). End-to-end visual speech recognition with LSTMs. In *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on,* pp. 2592-2596. IEEE. DOI: https://doi.org/10.1109/ICASSP.2017.7952625

[14] Pinzon-Arenas, J. O., Jimenez-Moreno, R., & Rubiano-Fonseca, A. (2018). Mouth States using LSTM Neural Network. In *2018 IEEE International Conference on Automation (ICA-ACCA)*. pp. 1-5.

[15] Baltrusaitis, T., Zadeh, A., Lim, Y. C., & Morency, L. P. (2018, May). OpenFace 2.0: Facial Behavior Analysis Toolkit. In *Automatic Face & Gesture Recognition (FG 2018), 2018 13th IEEE International Conference on,* pp. 59-66. IEEE. DOI: https://doi.org/10.1109/FG.2018.00019