Gesture Recognition using FastDTW and Deep Learning Methods in the MSRC-12 and the NTU RGB+D Databases

Júlia Schubert Peixoto, Anselmo Cukla, Marco Antonio de Souza Leite Cuadros, Daniel Welfer and Daniel Fernando Tello Gamarra

***Abstract*—This work explores the use of three deep learning methods for gesture recognition: Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) using Fast Dinamic Time Warping (FastDTW). The gestures were captured by kinect sersors, two skeleton-based databases are used: Microsoft Research Cambridge-12 (MSRC-12) and NTU RGB+D. Also, the FastDTW technique was also employed to standardize the input size of the data. With MSRC-12 database was achieved an accuracy rate of 82,36% in test set with the CNN, with the LSTM was achieved an accuracy rate of 87,30% also in the test set, and in GRU the accuracy achieved in the test set was 89,34%. With NTU RGB+D database two evaluation methods were used: Cross-View and Cross-Subject. In the test set with Cross-View evaluation was obtained an accuracy rate of 63,53%, 55,14% and 61,00%, with CNN, LSTM and GRU respectively. And with Cross-Subject evaluation method it was achieved an accuracy rate of 66,19%, 64,43% and 60,17% in the test set on CNN, LSTM and GRU, respectively.**

***Index Terms*—Deep learning, convolutional neural networks, long short-term memory, gated recurrent unit, gesture recognition, FastDTW**

I. Introduction

Using computer vision could be beneficial to many applications as American Sign Language(ASL) recognition or control of a mobile robot with differente objectives. Gesture Recognition can be used in Smart Homes in order to control devices inside a residence through Internet of Things (IoT). Gesture detection can be used to turn on a light or even running an app on a Smart TV. Furthermore, the low cost sensors developed for video games can be helpfull in order to make simpler to realize gesture recognition tasks. Kinect Sensor, developed for Xbox, is an example of this type of sensor. It can record human joint positions in videos and allow us to use computer vision to recognize gestures performed by humans.

Microsoft Research Cambridge-12 dataset known as the MSRC-12 is integrated of Kinect images and includes 12 different gestures captured from different persons [1]. NTU RGB+D is also a dataset captured by Kinect sensor it constains 60 gestures perfomed by different subjects [2].

In order to recognize gestures it is necessary to analize the movement captured in videos. Each video is divided in segments called frames. Since each movement can have a different execution time and each person can perfomed this gesture with different execution time, each gesture video can have different number of frames. FastDTW algorithm is a dynamic time warping algorithm ([3]) that can be used in order to make that the image frames in a sequence could have the same size.

Neural Networks is a technique that has been used in different fields, and Gesture recognition is a problem of pattern recognition in which neural networks could contribute with its solution. There are many types of neural networks architectures that can be used for gesture recognition. Convolutional neural networks are being widely used in computer vision [4] [5]. And recurrent architectures can also be used to capture a sequence patterns.

This paper presents an application of three deep learning methods for skeleton-based gesture recognition, two databases are used: MSRC-12 and NTU RGB+D and three neural networks models are proposed: CNN, LSTM and GRU. The paper is divided in six sections. After a brief introduction, the second section describes some related works are presented. The third section describes the Theoretical Background of this paper. The fourth section presents the Methodology. The fifth section presents the Experimental Results. Finally, the last section summarizes the conclusions of the article.

II. Related Work

Some previous works will be reviewed. Pfitscher et al [6] and [7] used a convolutional neural network for the MSRC-12 gesture recognition and controlled a mobile robot. Two approaches for training were used: combined training and individual training. The proposed neural network presents an accuracy of 86,67% in combined training and 90,78% in individual training. Futhermore, the network was trained to identify gestures in images that come from a Kinect sensor and achieved an accuracy of 72,08% with combined training and 81,25% in individual training. This work was focused in the MSRC-12 and the evaluation of two different forms of training and just use a CNN.

Peixoto et al also applied a CNN [8] and a RNN [9] for the MSRC-12 dataset and proposes three techniques to preprocess the information derived from the joint coordinates captured with the kinect sensor: 3D coordinates method, Subtraction method and Normalization method. Two types of training are also used: combined and individual training.

Meng et al [10] applies an hierarchical dropped convolutional neural network (hd-CNN) on the MSRC-12 and NTU

Marco A. Hernandez-Nochebuena is with LINCE Lab, Instituto Politécnico Nacional Mexico e-mail:email@utfpr.edu.br.

RGB+D. The hdd-CNN on the MSRC-12 obtained an accuracy of 94,54%. Appling the same network on the NTU RGB+D with Cross-Subject and Cross-View achieved 84,33% and 92,21% accuracies, respectively. This work, different from ours only used a different CNN for both datasets.

Tu et al [11] employed a two-stream 3D convolutional neural network on two different datasets, the SmartHome and the NTU RGB+D. The experiments with the SmartHome datase delivered 79,38% of accuracy, and the experiments with the NTU RGB+D with Cross-Subject and Cross-View evaluation methods resulted in accuracies of 66,85% and 72,58%, respectively .This work compared to the work developed here just uses one dataset.

Lai et al [12] used a convolutional neural network combined with a recurrent neural network for hand gesture recognition, for this task the skeleton information derived from the kinect and the depth data were also used. The combined network architecture was able to identify gestures with an accuracy of 85,46% in the Dynamic Hand Gestures 14/28 dataset. This work different from our proposal uses a different network product of the combination of 2 networks and the dataset employed was the hand gesture dataset.

Lee et al [13] implemented a recurrent neural network to find patterns in the gait recognition system of persons using 3D coordinates joint, the network's accuracy was 97% .This work different from the work developed in this paper is just focused in one type of network with a different dataset.

Molchanov et al [14] employed a recurrent 3D convolutional neural network to recognize hand gestures. A new dataset was proposed and an accuracy of 83,8% was obtained. Also, in this work just one kind of network is explored with a different dataset from the one that is proposed in this article.

Song et al described in [15] richly activate multi-stream graph convolutional neural network (RA-GCN) for the classification of gestures derived from skeleton data. Testing the model in the NTU RGB+D dataset an accuracy of 85,8% with Cross-Subject evaluation was obtained and 93% with Cross-View evaluation.Also, this work explores one dataset and just one network architecture.

Ha et al [16] proposed a deep neural network using capsule networks for video gesture recognition and achieved 98,5% of accuracy with UCF101 and 95,3% of accuracy on HMDB51. Hao et al [17] proposes a hypergraph neural network for skeleton based action recognition using NTU RGB+D with Cross-Subject evaluation was achieved an accuracy of 89,5% ans with Cross-View evaluation 95,7%. The Kinectis skeleton dataset was also used and the accuracy achieved for Top-1 was 37,7% and for Top-5 was 60%.

The main contributions of this paper compared with others related in the scientific literature are: (i) the FAstDTW technique is applied to deal with the problem of different image sizes; (ii) three different neural networks architectures for skeleton-based gesture recognition are explored in the paper: a convolutional neural network, a long short-term memory and a gated recurrent unit. (iii)Two different databases are used the NTU RGB+D and the MSRC-12 in order to evaluate the architecture proposed in this work.

III. Theoretical Background

In this sections will be reviewed some important contents. First, will be presented the FastDTW algorithm, and after the CNN, LSTM and GRU.

A. FastDTW

Different gestures have different execution time and different subjects can perform gestures with different execution time. Then different number of frames are generated for every gesture.

In order to have a fixed input size were used a time warping algorithm as proposed by Salvador et al in [3]. FastDTW algorithm is based on the DTW algorithm and makes possible to work with bigger time series and databases. This algorithm determine similarities and a correspondent region between two time series.

The FastDTW algorithm uses a multilevel graph bisection algorithm, which will split a graph and get smaller graphs as possible. This multilevel approach is used to find an optimal solution for each small graph and makes the algorithm linear in time and space, as shown by Salvador [3].

B. Convolutional Neural Network

Convolutional neural networks has its origin in the applications of neural networks to image processing [18], it is characterized for a convolution of the input image and the filters of the layer, this interaction is done through the convolutions operation.Convolutional neural networks have been applied in different fields such as in mobile robotics for object detection [19] or through the use of some layers of the network such in deep reinforcement learning for the control of the actions in a mobile robot [20].

Different layers are employed in a convolutional neural network such as a pooling layer, a fully connected layer and a softmax function is used in the final layer. Fig. 1 shows a basic CNN architecture with its layers.

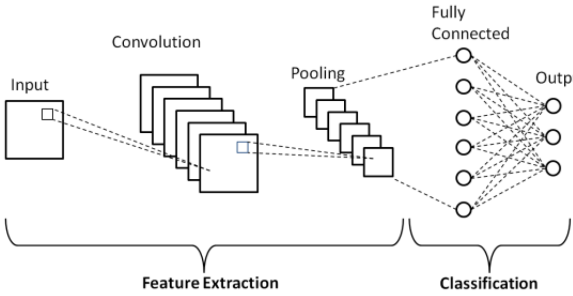

Fig. 1. Basic CNN architecture [21].

C. Long Short-Term Memory

According to Hochreiter et al [22], LSTM structure is a method to tackle the problem of vanishing gradients, it is constituted by gate units and memory cells. The Long Short Term Memory network has in its structure the following gates: forget gate, input gate and output gate. Figure 2 depicts the structure of a LSTM neuron with its gates and activation functions.

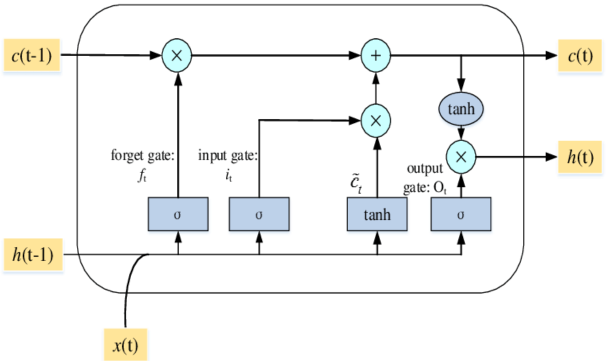

Fig. 2. LSTM neuron [23].

D. Gated Recurrent Unit

Gated recurrent unit was introduced by Cho et al [24]. This neural network achitecture was proposed to ovecome the vanishing gradient problem that occurs in recurrent neural networks. GRUs also have gate units, like the LSTM, that decides which information to keep and which information to discard at each time step. Figure 3 presents a Gated Recurrent Unit diagram with the gates and activation functions.

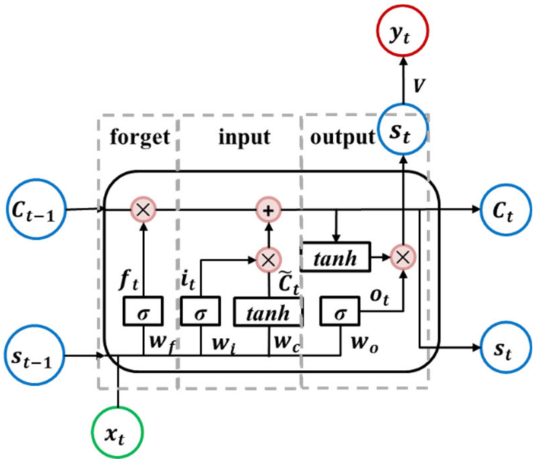

Fig. 3. Gated Recurrent Unit diagram [25].

IV. Materials and Methodos

This section describes the data preprocessing of the MSRC-12 and NTU RGB+D dataset that will be necessary to use with the neural network , CNN proposed model, LSTM proposed model, GRU proposed model and the software architecture.

A. The MSRC-12 Dataset

The MSRC-12 dataset consists in a human body joint movement sequence represented by the skeleton of Kinect sensor proposed by Fothergill et al [26]. This database archives has the position of 20 joints acquired by a Kinect version 1. The dataset has 12 different actions, with 594 sequences, about six hours of movement frames performed by 30 subjects, resulting in 6244 gesture samples.

Video, image and text were used in order to show the gestures to the participants. In some cases were presented texts with images, or videos with texts, so the participants can realize the movements combinations in a simpler way, without any sofisticated previous knowledge.

The gestures are divided in two categories: iconic gestures that has conection between the gesture and the reference and metaphorical gestures that represented abstract concepts. Figure 4 presents a gesture sequence of MSRC-12.

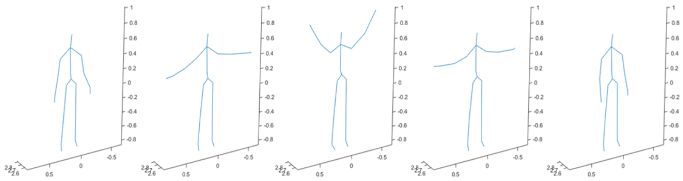

Fig. 4. MSRC-12 example gesture.

The gestures used for training our models are shown in Table I.

TABLE I
MSRC-12 Gestures

Gestures	
Crouch or hide	Start music/Raise volume (of music)
Shoot a pistol	Navigate to next menu
Throw an object	Wind up the music
Change weapon	Take a bow to end music session
Kick	Protest the music
Put on night vision goggles	Move up the time song

B. The NTU RGB+D Dataset

The NTU RGB+D Dataset is a large-scale dataset for human actions recognition. This dataset is divided into two structures: NTU RGB+D, that contains 60 action classes and 56,880 video samples, and NTU RGB+D 120 that contains 120 action classes and 114,480 video samples. Both datasets were captured with the Microsoft Kinect V2 and each sample contains RGB videos, depth maps, 3D skeletal data and IR videos. 3D skeletal data capture 3D coordinates of 25 human body joints. Each gesture is performed by 40 people with ages between 10 e 35 years.

All gestures were captured by three Kinect sensors positioned in the same height but in different horizontal angles, one of them is positioned in -45º, another one is positioned in 0º and the last one is positioned in 45º. Furthermore, 17 camera setups are proposed where height and distance are changed. Sharoudy et al [2], proposed two standard evaluation methods: Cross-View and Cross-Subject. Cross-View method consists in using all samples of camera 1 for testing and the samples of cameras 2 and 3 for training. Cross-Subject method consists of dividing the 40 subjects in two groups of 20 subjects and using one of the groups for training and another one for testing. Figure 5 presents a gesture of NTU RGB+D.

The gestures used for training our models are shown in Table II.

C. CNN

The convolutional neural network architectured described by Pfitscher [6] will be used. It has one input layer, three

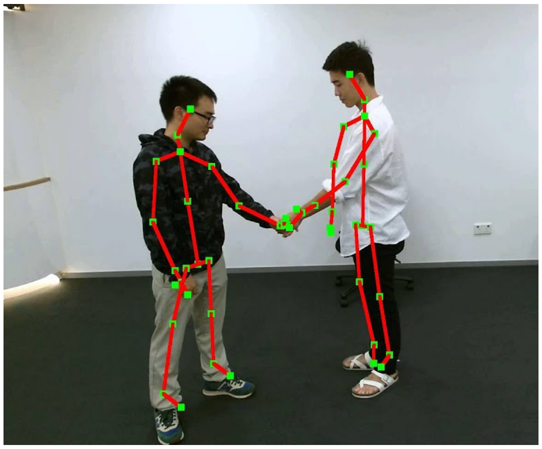

Fig. 5. NTU RGB+D example gesture [2].

TABLE II
NTU RGB+D GESTURES

Gestures	
Drink water	Eat meal
Brush teeth	Brush hair
Drop	Pick up
Throw	Sit down
Stand up	Clapping
Writing	Reading

convolution layers, three pooling layers, one fully connected layer and a dropout operation.

The proposed model contains three convolution layers with rectified linear unit activation function (ReLu). Three pooling layers, two with 3x3 kernel and one with 2x2 kernel. Some tests were performed with the hyperbolic tangent as activation function in the convolution layers but the results were not so good as with ReLu function, and for this reason, ReLu was used.

The fully connected layer has 1024 neurons and 9472 units and the activation function is also ReLu. At last, a dropout layer is added with 0,4 rate to avoid overfitting. The last layer is a dense layer and uses as an activation function the softmax. For the convolutional neural network training the optimizer selected was the Adam and the loss function selected was Sparse Categorical Cross Entropy.

D. LSTM

The recurrent neural network proposed contains an input layer, two LSTM layers each one with 128 neurons, and the second one with a dropout rate of 0,1. The last layer is a dense layer with softmax activation function. The LSTM network training used the the RMSProp optimizer and the loss function selected was the Sparse Categorical Cross Entropy.

E. GRU

The proposed GRU has one input layer, two GRU layers, both with 128 neurons, the last one with a dropout rate of 0.1, one 1D convolution layer with the ReLu activation function and a dense layer with a softmax activation function.

For GRU network training was also used RMSProp as the optimizer and the loss function selected was Sparse Categorical Cross Entropy.

F. Software Architecture

The data preprocesing was implemented in the environment of a Intel Core i3-4005U 1.70GHz 4GB/Windows10 Home-x64, using Python3.6. The neural networks training was implemented in the free version of Google Colaboratory, using a Graphic Processing Unit (GPU). Once the free version was used the GPU could be a Nvidia K80, T4, P4 or P100.

V. EXPERIMENTAL RESULTS

In this section will be presented the experimental results with MSRC-12 and NTU RGB+D datasets using a Convolutional Neural Network, LSTM network and GRU network.

For training the three proposed neural networks the datasets were splited in training data, evaluation data and test data. With this data separated was possible to obtain the training accuracy, evaluation accuracy and test accuracy for each deep learning method proposed.

CNN, LSTM and GRU proposed architectures were trained for 30 epochs with 10 samples by step.

A. Preprocessing Using the FastDTW

The CNN neural networks need that all the images should have the same size, if the images do not have the same size, the neural network will not be able to process the images and will fail, so an important issue to work with CNN neural networks is having images with the same size, for constructing the image, the article employed a method that consists in making an array with the x,y and z joint coordinates forming a matrix, this method for constructing an image is very simple compared with others that exist in the literature, that require more computational processing, as for example the method of Gramian Angular Matrix that converts a signal into an image but is computational heavier.

The MSRC-12 preprocessing has four steps, First, the FastDTW algorithm is applied in the data in order to let all gestures with the same shape, same number of frames, the output of this step creates a matrix with 667×80 rows that are related to the number of frames and columns and condensed the information of each coordinate (x, y, z) of each one of the twenty joints,plus a separation between them. Then, the sci-kit learn algorithm MaxAbsScaler is used to normalize the data between -1 and 1. Finally, the data is splitted in the training set, evaluation set and test set.

The NTU RGB+D preprocessing has six steps. First, the gesture selection was executed, once the database has 60 gestures and some of them are performed by more than one subject, these gestures need to be discarded as soon we are using the same neural networks used for MSRC-12 and wanted our data to be similar to the MSRC-12. From 60 gestures remain 49 with only one skeleton in the scene. We need to reduce our dataset since the software architecture cannot process the 49 gestures and each gesture has approximately

900 files. So were used only 12 gestures for training the neural networks. In the second step of preprocessing some transformations were realized. Once the NTU RGB+D was captured by the Kinect v2 and it captures 5 extra body joints that Kinect v1, used for capture the MSRC-12, does not capture. So we needed to remove these 5 extra joints.

In the third step the one hot encoded labels where created, since we are using a supervised training technique this step is very important. In the next step we used FastDTW algorithm to let all the gestures with the same number of frames. After this step all gestures had 299 frames. Then, all data was normalized by a sci-kit learn algorithm: MaxAbsScaler. Finally, was realized the data split in training data, evaluation data and test data.

B. MSRC-12

For training the proposed models 4192 samples were used, 1032 samples were used to evaluation and 1032 samples were used for the prediction test. Table III shows the accuracy obtained for training, evaluation and test in the MSRC-12 dataset with two deep learning methods: LSTM network and CNN.

TABLE III
ACCURACY IN MSRC-12

Network	Training	Evaluation	Test
CNN	95,81%	81,98%	82,36%
LSTM	88,69%	85,47%	87,30%
GRU	93,40%	89,44%	89,14%

Table III shows that the CNN training accuracy was 95,81% and in the LSTM the training accuracy was 88,69%. In the prediction test the CNN presented an accuracy of 82,36%, and the LSTM presented an accuracy of 87,30% in the test set. The best results obtained were in GRU that presented 89,14% in the prediction test.

Figure 6 presents the accuracy obtained in GRU network, that presented the best results in MSRC-12.

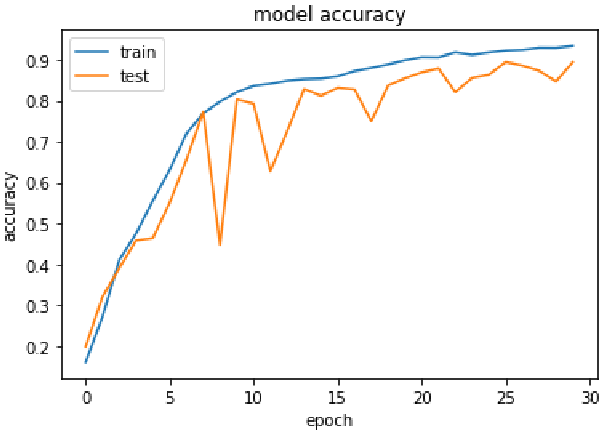

Fig. 6. Accuracy on GRU network.

In Figure 6 the blue line represents the training accuracy, and the orange line represents the evaluation accuracy. The training accuracy achieves values above 90% of accuracy and the evaluation accuracy achieves values above 85%.

Table IV presents the accuracy and F1 score for each gesture.

TABLE IV
MSRC-12 SCORES FOR EACH GESTURE

	Accuracy (F1 score)		
Gesture	CNN	LSTM	GRU
Start system	80,72% (84%)	80,72% (82%)	80,72% (87%)
Duck	88,89% (83%)	87,65% (89%)	91,36% (90%)
Push Right	83,78% (78%)	77,02% (86%)	91,89% (87%)
Goggles	87,64% (86%)	59,55% (69%)	89,89% (87%)
Wind it up	79,80% (81%)	87,50% (83%)	83,65% (86%)
Shoot	85,54% (88%)	90,36% (81%)	84,33% (88%)
Bow	80,00% (87%)	83,00% (87%)	90,00% (91%)
Throw	92,59% (82%)	91,01% (93%)	94,38% (90%)
Had enough	85,88% (90%)	83,53% (90%)	96,47% (96%)
Change weapon	92,59% (91%)	97,53% (86%)	92,59% (91%)
Beat both	79,74% (82%)	87,34% (85%)	89,87% (89%)
Kick	86,90% (91%)	92,86% (87%)	89,29% (92%)

The gesture Change Weapon presents the best accuracy (97,53%) with LSTM model. And the gesture Had Enough presents the best F1 score (96%) with GRU model as shown in Table V.

C. NTU RGB+D

For training the CNN, the LSTM and the GRU network with 12 gestures with Cross-View Evaluation method 7486 samples were used, 1894 samples were used for evaluation and 1895 samples were used for test. For training the three neural networks proposed with Cross-Subject Evaluation method 7991 samples were used, 1642 samples were used for evaluation and 1642 samples were used for test. The accuracy obtained for all networks is shown in Table IV.

TABLE V
ACCURACY IN NTU RGB+D WITH 12 GESTURES

Network	Training	Evaluation	Test
CNN Cross-View	82,29%	64,36%	63,53%
LSTM Cross-View	74,18%	56,97%	55,14%
GRU Cross-View	73,40%	58,03%	61,00%
CNN Cross-Subject	86,51%	67,05%	66,19%
LSTM Cross-Subject	77,88%	65,47%	64,43%
GRU Cross-Subject	75,71%	58,31%	60,17%

Figure 7 presents the accuracy obtained in the CNN with cross-subject evaluation, that presented the best results in NTU RGB+D. In Figure 7 the blue line represents the training accuracy, and the orange line represents the evaluation accuracy. The training accuracy achieves values above 80% of accuracy and the evaluation accuracy achieves values above 60%.

Table VI presents the accuracy and F1 score for each gesture with Cross-View Validation.

The gesture Pick Up presents the best accuracy (98,16%) with CNN model. And the gesture Stand Up presents the best F1 score (97%) with GRU model as shown in Table VI.

Table VII presents the accuracy and F1 score for each gesture.

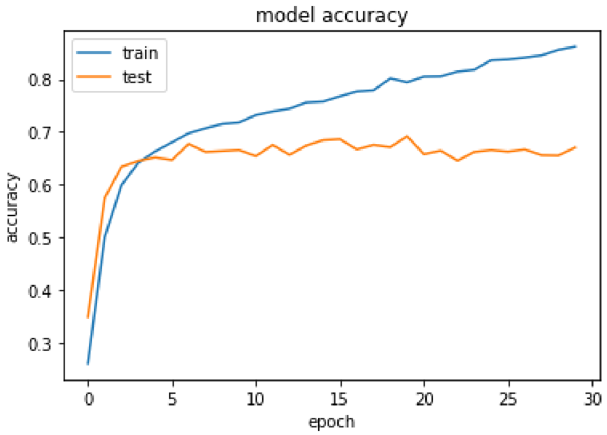

Fig. 7. Accuracy on CNN.

TABLE VI
NTU RGB+D Cross-View scores for each gesture

	Accuracy (F1 score)		
Gesture	CNN	LSTM	GRU
Drink water	26,79% (36%)	32,03% (40%)	14,38% (24%)
Eat meal	43,79% (47%)	55,03% (62%)	35,5% (48%)
Brush teeth	40,64% (42%)	63,87% (59%)	11,61% (19%)
Brush hair	77,58% (49%)	60,00% (53%)	64,24% (43%)
Drop	80,25% (77%)	59,87% (58%)	71,97% (74%)
Pick Up	98,16% (96%)	93,86% (94%)	96,93% (96%)
Throw	84,85% (75%)	87,27% (75%)	93,94% (68%)
Sit down	87,50% (92%)	83,93% (89%)	91,07% (93%)
Stand up	95,10% (95%)	85,31% (75%)	97,20% (97%)
Clapping	31,13% (44%)	68,87% (63%)	47,68% (54%)
Reading	48,68% (43%)	50,66% (50%)	46,71% (41%)
Writing	22,08% (31%)	24,67% (35%)	59,74% (51%)

The gesture Pick Up presents the best accuracy (98,41%) with CNN model. And the gesture Pick Up presents the best F1 score (95%) with GRU model as shown in Table VII.

D. Discussion

Convolutional Neural Networks and Recurrent Neural Networks have been used recently for gesture recognition, in this article it is done a more complete study of the problem of gesture recognition not only using one neural network, we have used three different architectures of neural networks, one convolutional and two recurrent ones. Although, we have used two different datasets instead of one in order two generalize our approach. Also, the preprocessing method used based in the conversion of the information of the joint coordinates in images and its transformation of images of the same sizes using the technique of FastDTW. Then the article is a holistic study, because of these three aspects, different network architectures, different datasets and a different methodology for preprocessing the data, and all these aspects composed the architecture proposed in this paper

VI. Conclusions

TABLE VII
Scores for each gesture

	Accuracy (F1 score)		
Gesture	CNN	LSTM	GRU
Drink water	54,61% (49%)	59,57% (55%)	65,96% (59%)
Eat meal	31,25% (37%)	56,25% (58%)	36,11% (40%)
Brush teeth	35,34% (44%)	51,13% (57%)	55,64% (58%)
Brush hair	52,27% (53%)	74,24% (69%)	71,97% (55%)
Drop	66,42% (74%)	79,56% (80%)	33,58% (49%)
Pick Up	98,41% (93%)	93,65% (95%)	95,24% (94%)
Throw	65,80% (70%)	65,16% (71%)	78,71% (77%)
Sit down	89,51% (94%)	89,52% (92%)	90,32% (94%)
Stand up	94,37% (92%)	90,84% (93%)	84,51% (91%)
Clapping	67,97% (55%)	48,44% (57%)	53,12% (58%)
Reading	59,85% (47%)	46,21% (36%)	37,88% (39%)
Writing	37,84% (44%)	41,89% (42%)	64,86% (53%)

This article contains a successful implementation of deep learning and the FastTDTW technique for skeleton-based gesture recognition, different datasets and different neural networks architectures were used in the experiments. In the MSRC-12 dataset good experimental results were achieved with an accuracy of 82,36% in the test set with the CNN model, 87,36% with the LSTM model and 89,14% with the GRU model. In the NTU RGB+D with Cross-View evaluation we achieved an accuracy rate of 63,53% in the teste set with the CNN, 55,14% with LSTM and 61,00% with GRU. In the NTU RGB+D dataset with Cross-Subject evaluation we achieved an accuracy rate of 66,19% in the CNN, 64,43% in the LSTM and 60,17% in the test set with the GRU. The recurrent models proposed presented a better accuracy in the MSRC-12 dataset compared with the convolutional model. The best result in the test set of MSRC-12 is with the GRU model. In NTU RGB+D the convolutional model proposed presents the best results in Cross-View evaluation and in Cross-Subject evaluation. The LSTM presented better accuracy than GRU in Cross-Subject evaluation method, and the GRU model presented better accuracy in Cross-View evaluation method. As we can see the results on MSRC-12 were better than the results presented in the NTU RGB+D this happens because NTU RGB+D is a most challenging database. The codes of our paper can be found in this github repository.

In future work a different way of preprocessing the joint information and other deep learning architectures can be explored for comparisson with the results described in this paper. Also other more modern devices can be used as the Azure Kinect [27] or the wearnotch [28].

References

[1] S. Fothergill, H. M. Mentis, P. Kohli, and S. Nowozin, "Instructing people for training gestural interactive systems," *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2012.

[2] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, "Ntu rgb+d: A large scale dataset for 3d human activity analysis," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1010–1019, 2016.

[3] S. Salvador and P. Chan, "Fastdtw: Toward accurate dynamic time warping in linear time and space," 2004.

[4] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings* (Y. Bengio and Y. LeCun, eds.), 2015.

[5] C. Szegedy, Wei Liu, Yangqing Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–9, June 2015.

[6] M. Pfitscher, D. Welfer, E. J. D. Nascimento, M. A. D. S. L. Cuadros, and D. F. T. Gamarra, "Article users activity gesture recognition on kinect sensor using convolutional neural networks and fastdtw for controlling movements of a mobile robot," *Inteligencia Artif.*, vol. 22, no. 63, pp. 121–134, 2019.

[7] M. Pfitscher, D. Welfer, M. A. de Souza Leite Cuadros, and D. F. T. Gamarra, "Activity gesture recognition on kinect sensor using convolutional neural networks and fastdtw for the msrc-12 dataset," in *Intelligent Systems Design and Applications*, pp. 230–239, Springer International Publishing, 2020.

[8] J. S. Peixoto, M. Pfitscher, M. A. de Souza Leite Cuadros, D. Welfer, and D. F. T. Gamarra, "Comparison of different processing methods of joint coordinates features for gesture recognition with a cnn in the msrc-12 database," in *Intelligent Systems Design and Applications*, (Cham), pp. 590–599, Springer International Publishing, 2021.

[9] J. S. Peixoto, A. R. Cukla, D. Welfer, and D. F. T. Gamarra, "Comparison of different processing methods of joint coordinates features for gesture recognition with a rnn in the msrc-12," in *Intelligent Systems Design and Applications*, (Cham), pp. 498–507, Springer International Publishing, 2022.

[10] F. Meng, H. Liu, Y. Liang, M. Liu, and W. Liu, "Hierarchical dropped convolutional neural network for speed insensitive human action recognition," in *2018 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6, 2018.

[11] J. Tu, M. Liu, and H. Liu, "Skeleton-based human action recognition using spatial temporal 3d convolutional neural networks," in *2018 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6, 2018.

[12] K. Lai and S. N. Yanushkevich, "Cnn+rnn depth and skeleton based dynamic hand gesture recognition," in *2018 24th International Conference on Pattern Recognition (ICPR)*, pp. 3451–3456, 2018.

[13] D.-W. Lee, K. Jun, S. Lee, J.-K. Ko, and M. S. Kim, "Abnormal gait recognition using 3d joint information of multiple kinects system and rnn-lstm," in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 542–545, 2019.

[14] P. Molchanov, X. Yang, S. Gupta, K. Kim, S. Tyree, and J. Kautz, "Online detection and classification of dynamic hand gestures with recurrent 3d convolutional neural networks," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4207–4215, 2016.

[15] Y.-F. Song, Z. Zhang, and L. Wang, "Richly activated graph convolutional network for action recognition with incomplete skeletons," in *2019 IEEE International Conference on Image Processing (ICIP)*, pp. 1–5, 2019.

[16] M.-H. Ha and O. T.-C. Chen, "Deep neural networks using capsule networks and skeleton-based attentions for action recognition," *IEEE Access*, vol. 9, pp. 6164–6178, 2021.

[17] X. Hao, J. Li, Y. Guo, T. Jiang, and M. Yu, "Hypergraph neural network for skeleton-based action recognition," *IEEE Transactions on Image Processing*, vol. 30, pp. 2263–2275, 2021.

[18] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems* (F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, eds.), vol. 25, Curran Associates, Inc., 2012.

[19] D. Henke dos Reis, D. Welfer, M. A. de Souza Leite Cuadros, and D. F. Tello Gamarra, "Object recognition software using rgbd kinect images and the yolo algorithm for mobile robot navigation," in *Intelligent Systems Design and Applications*, (Cham), pp. 255–263, Springer International Publishing, 2020.

[20] J. Costa de Jesus, V. Kich, A. Kolling, R. Grando, M. Cuadros, and D. F. Tello Gamarra, "Soft actor-critic for navigation of mobile robots," *Journal of Intelligent & Robotic Systems*, vol. 102, 06 2021.

[21] V. Phung and E. Rhee, "A deep learning approach for classification of cloud image patches on small datasets," *Journal of Information and Communication Convergence Engineering*, vol. 16, pp. 173–178, 01 2018.

[22] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[23] X. Yuan, L. Li, and Y. Wang, "Nonlinear dynamic soft sensor modeling with supervised long short-term memory network," *IEEE Transactions on Industrial Informatics*, vol. PP, pp. 1–1, 02 2019.

[24] K. Cho, B. van Merriënboer, C. Gulcehre, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," 06 2014.

[25] H. Zhao, Z. Chen, H. Jiang, W. Jing, L. Sun, and M. Feng, "Evaluation of three deep learning models for early crop classification using sentinel-1a imagery time series-a case study in zhanjiang, china," *Remote Sensing*, vol. 11, p. 2673, 11 2019.

[26] S. Fothergill, H. Mentis, P. Kohli, and S. Nowozin, "Instructing people for training gestural interactive systems," *Conference on Human Factors in Computing Systems - Proceedings*, 05 2012.

[27] Q. Wu, "Research on human body detection and tracking algorithm based on kinect," in *2021 International Conference on Electronic Information Technology and Smart Agriculture (ICEITSA)*, pp. 7–10, 2021.

[28] M. TR, C. ME, Metsis, N. AHH, and R. CC., "A smartwatch-based fall detection system using deep learning," *Sensors*, vol. 10, 2018.

Júlia Schubert Peixoto has a degree in Control and Automation Engineering from the Universidade Federal de Santa Maria in Brazil (2021). She is currently a Data Scientist at Luizalabs. Her current research interests include robotics, deep learning, computer vision and natural language processing. ORCID https://orcid.org/0000-0001-5150-9745

Anselmo Rafael Cukla has a degree in Electrical Engineering from the Universidad Nacional de Misiones in Argentina (2010). He completed Master and PhD in Mechanical Engineering in the Federal University of Rio Grande do Sul (UFRGS), Brazil (2012 and 2016). He also is PhD in Electrical and Computers Engineering in the FCT (Faculty of Science and Technology) of the UNINOVA (Portugal). He is currently a professor in the Electrical Engineering course at the Federal University of Santa Maria, RS, Brazil. His research interests include automations, industrial robotics, robotics mobile, optimization algorithms and evolutionary systems. ORCID https://orcid.org/0000-0002-5313-4593.

Marco Antonio de Souza Leite Cuadros received his BSc. Degree in electrical engineering from Universidad Nacional del Centro del Peru (UNCP) in 1998, , and his MSc in electrical Engineering from Universidade Federal do Espirito Santo (UFES) in 2004, and PhD degree in electrical engineering from Universidade Federal do Espírito Santo in 2011.Currently, he is professor of the Instituto Federal do Espeirito Santo (IFES) in Vitoria-Espirito Santo- Brazil. ORCID https://orcid.org/0000-0003-4191-1794

Daniel Welfer Professor at the Department of Applied Computing at Federal University of Santa Maria (UFSM). Currently, he is also a permanent professor at the Graduate Program in Computer Science (PPGC). He works with medical image processing and analysis, software engineering, deep learning and mobile programming. ORCID https://orcid.org/0000-0003-1560-423X

Daniel Fernando Tello Gamarra received his BSc. Degree in mechanical engineering from Universidad Nacional del Centro del Peru (UNCP), in Huancayo, Peru (1999), and his MSc in electrical Engineering from Universidade Federal do Espirito Santo (UFES), in Vitoria, Espirito Santo, Brazil(2004) and PhD degree in Biomedical Robotics from Scuola Superiore Santa Anna, Pisa, Italy (2009). He is Professor at the Universidade Federal de Santa Maria(UFSM), in the Department of Control and Automation in Santa Maria, Rio Grande do Sul, Brazil. His current research interest centers on robotics, computational vision and machine learning.ORCID https://orcid.org/0000-0002-4714-7849