

HAND GESTURES RECOGNITION USING TIME DELAY NETWORKS

Irina Mocanu¹
Tatiana Cristea²

ABSTRACT

This paper proposes a system for body gestures recognition using the coordinates of the body skeletal returned by a Kinect sensor who are processed in order to compute a set of angles. Gesture recognition is achieved using a Time Delay Neural Network, implemented in two ways: with delay layer, and with delay synapse. The resulting system was trained on a set of gestures, such as: hands up and elbows bent, lifting arms, lowering arms, round, greeting gesture and pointing assertion. Each gesture was repeated at least 5 times with different speed. The accuracy of the proposed method is approximately 80%.

Keywords: gesture recognition, time delay networks, Kinect sensor, intelligent systems;

1. INTRODUCTION

The development of a series of devices has become an essential part of our daily life. The ease of use of this equipment has become an intensely studied in current research in the field of human interaction with the surrounding environment [1]. A method that facilitates communication with the human environment is the interpretation of his gestures. Thus there is no need for direct interaction with the devices but it is enough a gesture made from any position / location to trigger one or more actions (for example turning light, TV, alarm trigger).

This paper proposes a system for body gestures recognition using the coordinates of the skeletal returned by a Kinect device. The coordinates of the skeletal provided by the Kinect are processed in order to compute a set of angles for gesture recognition. Angles are calculated based on their relevance to detect gestures: the arm - forearm, neck - shoulders, arms - shoulders. Next step is made to identify the actual gesture based on the information processed previously. In [2], hand movements are analysed using time delay neural networks with RGB images as inputs in order to recognize hand movements. Based on that result we apply time delay neural networks (TDNN) with Kinect skeletal as input in order to recognize hand gestures. Time Delay Neural Network is implemented in two ways: with delay layer, and second with delay synapse.

The rest of the paper is organised as follows. Section 2 presents some existing methods for gesture recognition. The gesture recognition method with Time Delay Neural Network is

¹ Lecturer, University POLITEHNICA of Bucharest, Spaliul Independentei No. 313, Bucharest 060042, Romania, *irina.mocanu@cs.pub.ro

² Master Student, Artificial Intelligence, VU University Amsterdam, De Boelelaan 11051081 HV Amsterdam, The Netherlands, tatiana.cristea@gmail.com

described in Section 3. Experimental results, conclusion and future developments are presents in Section 4 and 5.

2. RELATED WORKS

Methods for human gestures analyses can be classified in two classes: static and dynamic detection methods, as it is described in [3]. This classification occurs due to various issues that need to treat algorithms used to identify gestures from each category. The static gestures can be identified using decision tree as in [4]. The dynamic gestures involve several aspects:

- it is necessary to be able to represent a temporal dependency between actions;
- the learned patterns must be invariant in time and
- actions must not require alignment over time.

In [5] motion patterns are learned from the extracted trajectories from videos using a time-delay neural network. The experiments described in the paper demonstrate recognition of 40 hand gestures of the American Sign Language. Experimental results show that hand gestures can be extracted and identified with high recognition rate. In paper [6] is described a model for recognizing general activities using Hidden Markov Models (HMM) and stochastic parsing. These activities are first identified as a series of low-level primitive actions, represented using the HMM and then they are recognized by string parsing using a context-free grammar. In [7] human activities are recognized by matching temporal templates with saved instances of images for known activity. Variable Length Markov Models to model human behavior in order to include temporal dependencies is used in [8]. The above approaches require image processing to extract relevant data to detect human body postures. This process involves some limitations, such as depth information which is difficult to extract only from one image. Thus, in [9] is adopted a system based on four synchronized cameras to recognize human postures.

3. GESTURE RECOGNITION

Based on the necessary aspects that we must consider in recognizing gestures we present a method for gesture recognition using Time Delay Neural Networks (TDNN) [10]. TDNN will be implemented in two ways: with delay layer, and second with delay synapse. In both cases TDNN has an input level, an output level and two hidden states. The differences between the two types of the TDNN are at the construction level of the network unit. The basic unit used in many neural networks computes the weighted sum of its inputs and then passes this sum through a nonlinear function, most commonly a threshold or sigmoid function. In a TDNN a basic unit is modified by introducing delays $D1$ to DN , as in Figure 1, as described in [10].

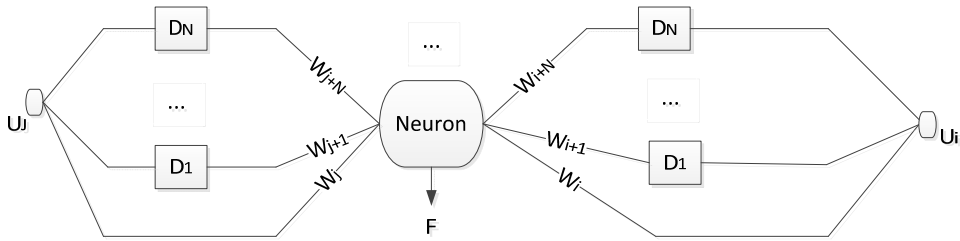


Figure 1. A Time Delay Neural Network unit

The J inputs of such a unit now will be multiplied by several weights, one for each delay and one for the not delayed input.

The Recognition System

We detect human gestures by processing the human skeletal. In order to detect the human skeletal we use a Kinect sensor. The information provided by Kinect is processed and we obtain the relevant data which will be used to train the TDNN. After the training phase we test the learned model and determine the error rate. The system architecture is presented in Figure 2 and is composed of three main modules:

- *the preprocessing module* - receives the human skeletal from the Kinect, computes a feature vector for each skeletal formed by a set of relevant angles;
- *the training module* – trains the TDNN; the input data is composed by the feature vectors computes by the processing module; The parameters of the trained TDNN are saved in the TDNN Parametres database.
- *the classification module* – classifies a new gesture; it receives a feature vector of a new human skeletal from the preprocessing module and classifies it using the learned parameters of the neural network from the TDNN Parameters database.

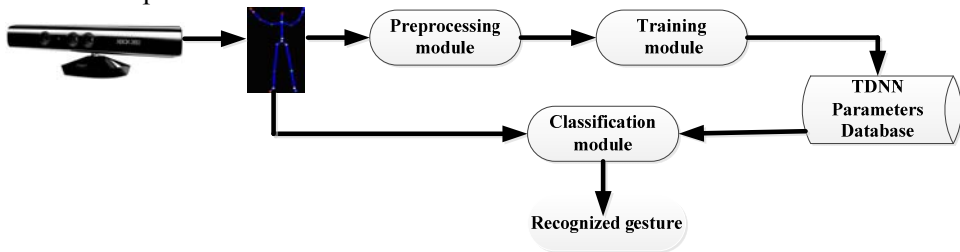


Figure 2. The system architecture

The Kinect sensor produced by Microsoft (Microsoft research, 2011) uses the PrimeSense's 3D sensing technology [11] and is composed by three main parts: an infrared laser projector, infrared camera, and a RGB color camera. The depth projector simply floods the surrounding space with Infra Red laser beams creating a depth field that can be seen only by the IR camera. Due to infrared's insensitivity to ambient light, the sensor can be used in any lighting conditions. The hardware is the basis for creating an image that the processor can interpret, the software behind the sensor is what makes everything possible. Using

statistics, probability, and testing for different natural human movements the SDK (software development kit) is able to track the movements of 20 main joints on a human body, as described in Figure 3. This software is how the sensor can differentiate a human from other objects (robots and other devices) that happens to be in front of the IR projector or different humans that share the same space. The sensor has the capabilities of tracking up to six different persons at a time, but as of now the software can only track up to two active operators.

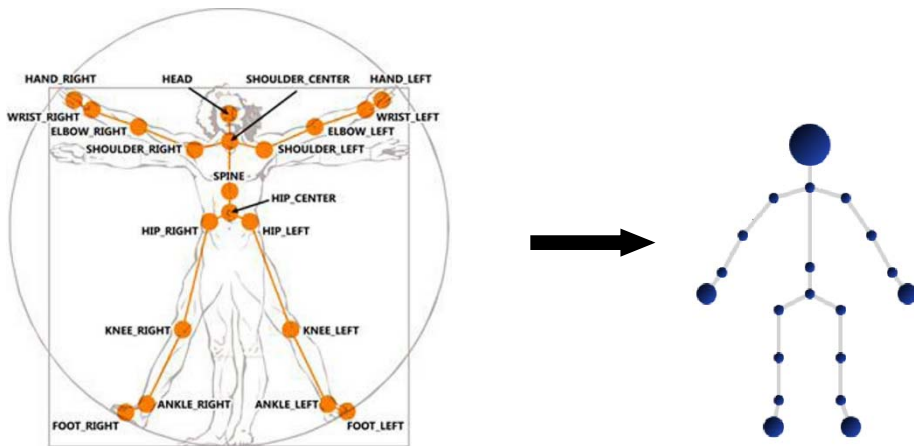


Figure 3. The skeletal computed by Kinect: (a) general view, b) a skeletal example

The processing module receives the skeletal provided by Kinect (the 20 joint points) and computes a feature vector for each skeletal. The feature vector is composed of 5 angles computed using only eight joint points from the Kinect skeletal: ShoulderLeft, ElbowLeft, WristLeft, ShoulderRight, ElbowRight, WristRight, ShoulderCenter and Head: (α , β , γ , δ , θ), as shown in Figure 4.

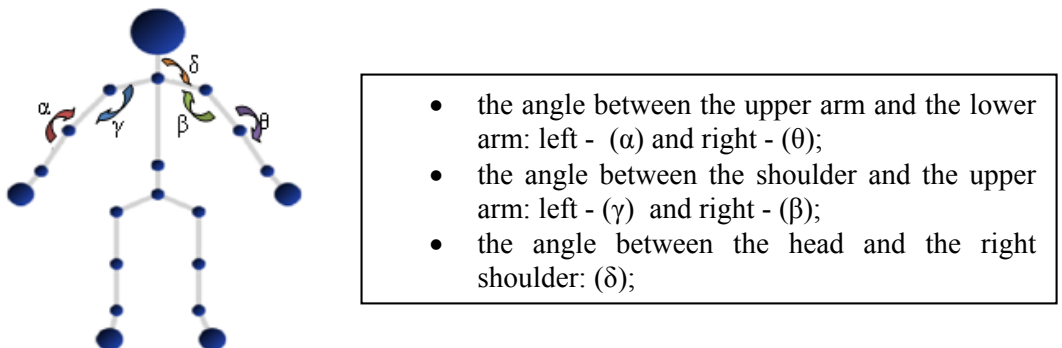


Figure 4. The feature vector for a human skeletal

The feature vector will be passed as input to the training module or to the classifying module. Both the training module and the classifier module use a neural network with a sigmoid activating function, who accepts input data in domain $[0, 1]$. Thus the feature vector will be normalized.

The training module uses a TDNN and must configure the architecture of the: number of synapses, the number of levels and the size of the time window. After that the network is trained and the weights associated with the synapses are calculated. The weights are computed using the back-propagation algorithm. It must be constructed two types of networks:

- TDNN with delay synapse and
- TDNN with delay layers.

The network parameters: the type of the network, the number of synapses, the number of nodes, the number of levels, the weights, the temporal window and the mean square error are saved in the TDNN Parameters Database.

The classifying module will classify a new gestured using a trained network with parameters selected from the TDNN Parameters Database. The result of the classifying module is a vector with the detected gestures of the supervised person.

4. EXPERIMENTAL RESULTS

The gesture recognition system was implemented in Java using the Java Object Oriented Neural Network (JOONE) library [12]. JOONE is a free neural network framework to create, train and test artificial neural networks. The aim is to create a powerful environment both for enthusiastic and professional users, based on the newest Java technologies. JOONE is composed by a central engine that is the fulcrum of all applications that are developed with JOONE.

For testing we use 3500 frames with gestures. Each gesture was repeated at least 5 times with different speed. We evaluate the two types of TDNN networks: with delay synapse and with delay layer, with 10.000 epochs for training, varying the parameters of the network: the number of neurons from the hidden layers, dimension of the temporal window, and finding the optimal network configuration for a good accuracy of gesture recognition.

We obtain the accuracy of the proposed method approximately 80% and the parameters of the optimal TDNN are presented below.

Gesture Recognition with Type Delay Neural Networks with Delay Synapse

The obtained results are presented in Figure 5. The optimal configuration for the TDNN is obtained with the following parameters: the number of neurons in the hidden levels: 3 and 4 and the size of the temporal window: 4, 6 and 9. The error is high if we vary the number of neurons from the hidden levels while maintaining constant size for the temporal windows. Also if the number of neurons from the hidden layers is constant and we vary the size of the temporal windows the global error is high. These facts are possible based on how a gesture is formed: the gesture is outlined in the first temporal window and it is accentuated in the second window and finally it is classified.

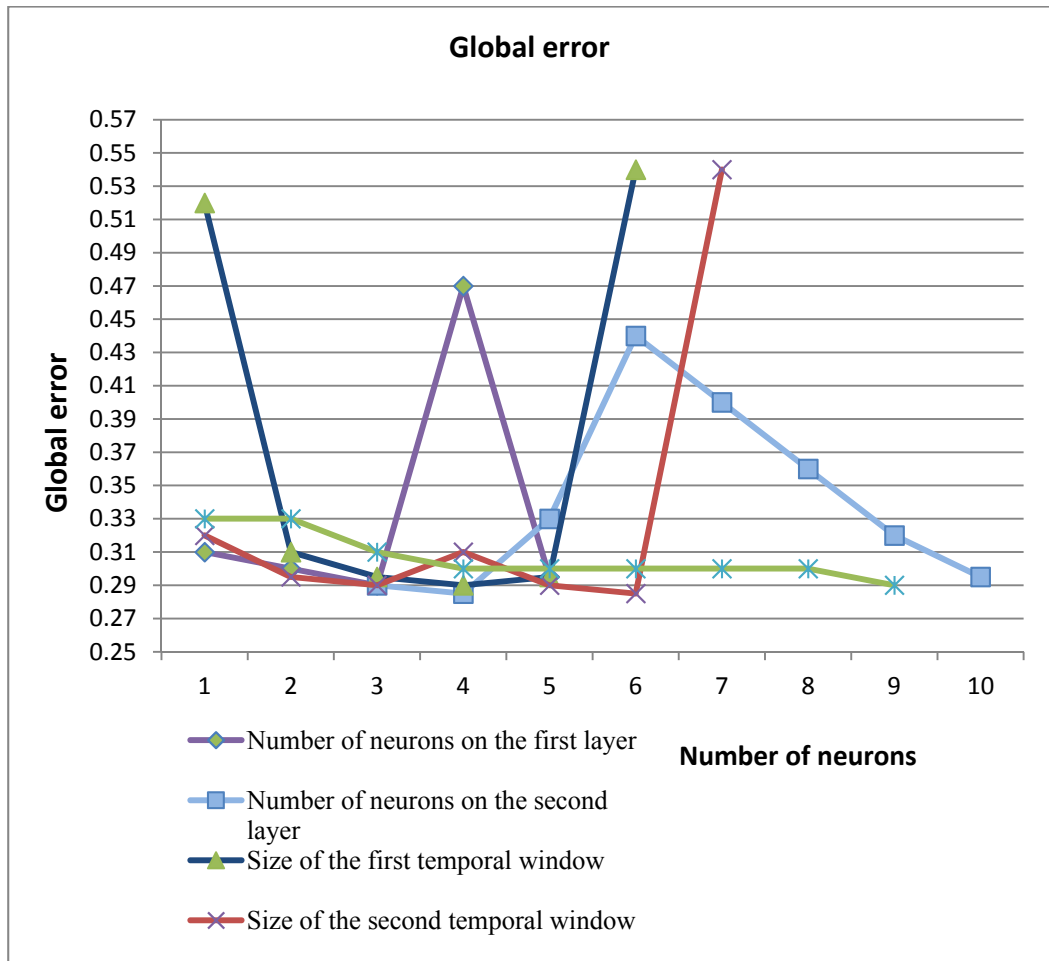


Figure 5. Global error depending of the number of neurons

Gesture Recognition with Type Delay Neural Networks with Delay Layer

The obtained results are presented in Figure 6 a) and b). The optimal configuration for the TDNN is obtained with the following parameters: the number of neurons in the hidden levels: 60 and 60 and the size of the temporal window: 28, 6 and 3. The global error is increasing if we vary the size of the temporal window or the number of the neurons from the hidden layers.

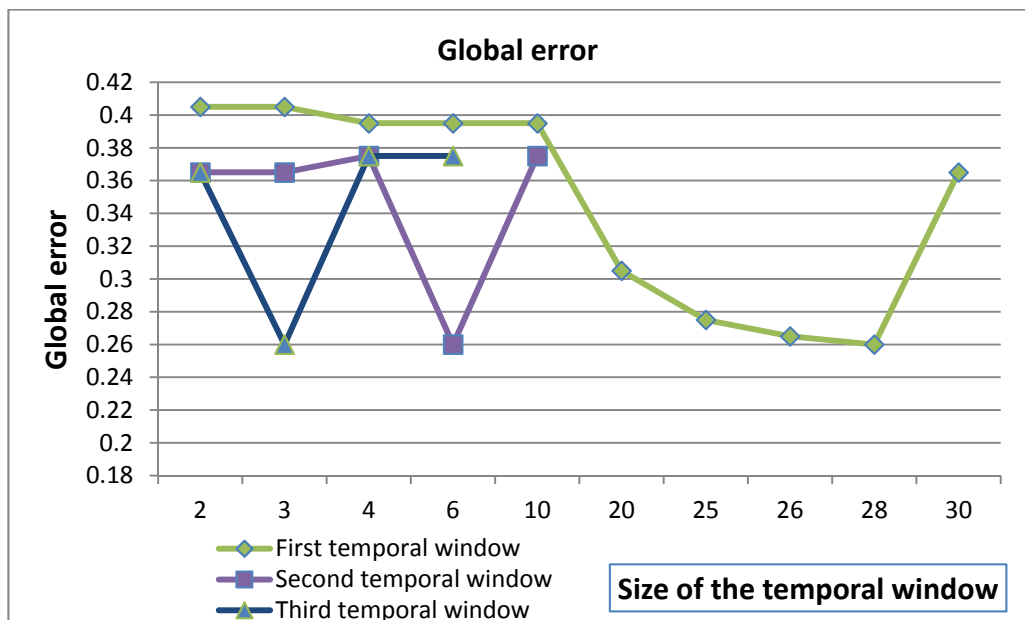


Figure 6. (a) Global error depending of the size of the temporal window

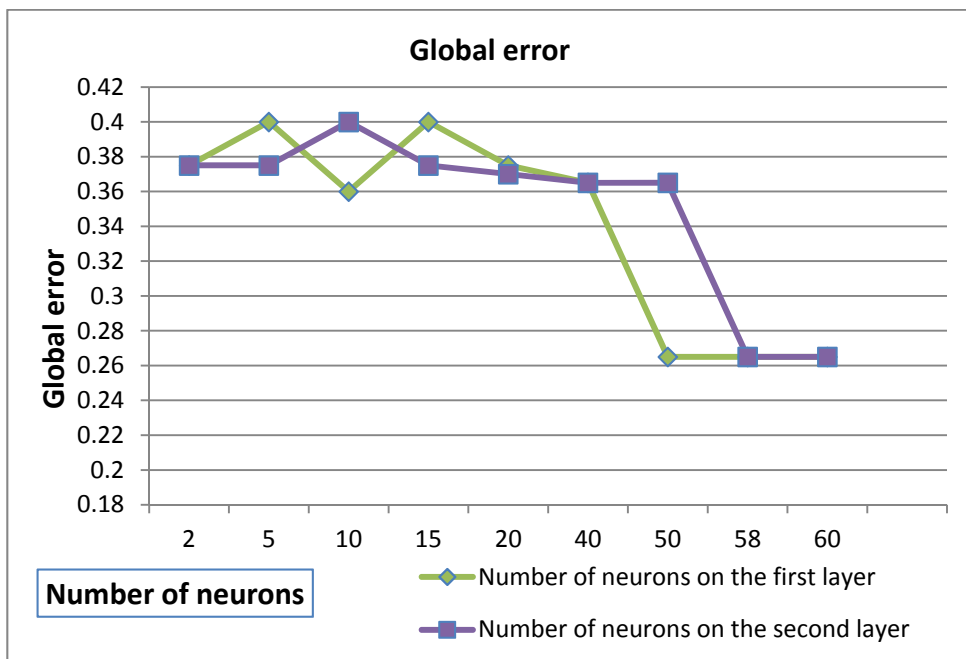


Figure 2 (b) Global error depending of the number of neurons

5. CONCLUSIONS AND FUTURE WORK

The coordinates of the skeletal provided by the Kinect are processed in order to compute a set of angles for gesture recognition. Angles are calculated based on their relevance to detect gestures: the arm - forearm, neck - shoulders, arms - shoulders. Next step is made to identify the actual gesture based on the information processed previously. Gesture recognition is achieved using a Time Delay Neural Network, implemented in two ways: with delay layer, and second with delay synapse.

The resulting system was trained on a set of gestures, such as: hands up and elbows bent, lifting arms, lowering arms, circling, greeting gesture and pointing assertion. Each gesture was repeated at least 5 times with different speed. The accuracy of the proposed method is approximately 80%. The best results were obtained for the following configurations: (i) network with delay layers with time windows: 28, 6, 3, and the number of neurons in the hidden levels 60 and 60 and (ii) network type time-delay type synapses implemented using delay with time windows: 4, 6, 9, and the number of neurons in the hidden levels 3 and 4.

As future work we will intend to recognize more gestures and we also will integrate the recognition system in an ambient intelligent system for supervising people. The ambient intelligent system will detect the gestures of a supervised person in order to interact with the intelligent devices from the house and also to recognize daily activities performed by the person.

6. REFERENCES

- [1] Cook D, Das S. Smart Environments Technology, Protocols and Applications. 2004
- [2] Modler, P. and Myatt, T. Recognition of Separate Hand Gestures by Time-delay Neural Networks Based on Multi-state Spectral Image Patterns from Cyclic Hand Movements. *IEEE International Conference on Systems, Man, and Cybernetics*, 2008, p. 1539-1544
- [3] Moeslund TB, Granum, E. A Survey of Computer Vision-Based Human Motion Capture. *In: Computer Vision and Image*; 2001; 81(3), p. 231-268
- [4] Mocanu, I, Florea, A. M. A Multi-Agent system for Human Activity Recognition in Smart Environments. *Intelligent Distributed Computing V, Studies in Computational Intelligence*, 382, 2012, p. 291-301
- [5] Yang MH, Ahuja N. Recognizing Hand Gesture Using Motion Trajectories. *In: Proc. IEEE Conf. Computer Vision and Pattern Recognition*; 1999, p. 466-472
- [6] Ivanov YA, Bobick A.F. Recognition of Visual Activities and Interactions by Stochastic Parsing. *In: IEEE Trans. Pattern Analysis and Machine Intelligence*; 2000; 22(8), p. 852-871
- [7] Aarob B, Davis J. W. Action Recognition using Temporal Templates. <http://www.cse.ohio-state.edu/~jwdavis/CVL/Publications/motrec.pdf>, accessed November 2013
- [8] Galata A, Johnson N, Hogg D. Learning Variable-Length Markov Models of Behavior. *In: Computer Vision and Image Understanding*; 2001; 81(3), p. 398-413
- [9] Chu CW, Cohen I. Posture and Gesture Recognition using 3D Body Shapes Decomposition. <http://iris.usc.edu/outlines/papers/2005/cwchu-v4hci05.pdf>, accessed November 2013
- [10] Waibel A, Hanazawa T, Hinton G, Lang KJ, Shikano K. Phoneme Recognition Using Time – Delay Neural Networks. <http://www.cs.toronto.edu/~hinton/absps/waibelTDNN.pdf>, accessed November 2013

- [11] Microsoft Kinect for Windows SDK, Programming Guide.
http://research.microsoft.com/redmond/kinectsdk/docs/programmingguide_kinectsdk.pdf,
accessed November 2013
- Memsic - <http://www.memsic.com/wireless-sensor-networks>
- [12] Marrone P. Java Object Oriented Neural Engine. The Complete Guide. All you need to know
about Joone, www.joone.org, accessed September 2012