3D Hand Pose Estimation with Neural Networks

Jose Antonio Serra[1], Jose Garcia-Rodriguez[1], Sergios Orts[1], Juan Manuel Garcia-Chamizo[1], Anastassia Angelopoulou[2], Alexandra Psarou[2], Markos Mentzelopoulos[2], Javier Montoyo-Bojo[1] and Enrique Dominguez[3]

[1]Dept. of Computing Technology, University of Alicante, PO.Box. 99. E03080. Alicante, Spain
{jserra,jgarcia,sorts,juanma,jmontoyo}@dtic.ua.es

[2]Dept. of Computer Science & Software Engineering (CSSE), University of Westminster, Cavendish, W1W 6UW, United Kingdom
{agelopa,psarroa, mentzem}@wmin.ac.uk

[3]Dept . of Computer Science, University of Malaga, Spain
{enriqued}@lcc.uma.es

Abstract. We propose the design of a real-time system to recognize and interpret hand gestures. The acquisition devices are low cost 3D sensors. 3D hand pose will be segmented, characterized and track using growing neural gas (GNG) structure. The capacity of the system to obtain information with a high degree of freedom allows the encoding of many gestures and a very accurate motion capture. The use of hand pose models combined with motion information provide with GNG permits to deal with the problem of the hand motion representation. A natural interface applied to a virtual mirror writing system and to a system to estimate hand pose will be designed to demonstrate the validity of the system.

Keywords: Growing Neural Gas, 3D Sensor, Hand Pose Estimation, Hand Motion, Trajectories Description.

1 Introduction

The estimation of the 3D hand pose has special interest because by understanding the configuration of the hands, we will be able to build systems that can interpret human activities and understand important aspects of the interaction of a human with their physical and social environment. There are several works that address only visual data without using markers [1,2]. Existing approaches can be classified as model based and appearance based. The model based systems provide a continuum set of solutions, but are computationally expensive and dependent on the availability of a large amount of visual information (usually provided by a multi-camera system). Appearance-based systems are associated with a less computational cost and a much smaller hardware complexity, but only recognize a discrete number of hand poses that correspond generally to the training set.

Despite the large amount of work in this field [3,4], the problem is still open and has several theoretical and practical challenges, due to a number of difficulties common to most systems, among which stand out: a) High dimensionality problem: the hand is an articulated with more object than 20 DOF. b) Self- occlusions: since the hand is an articulated object, its projections generate a variety of ways with various self-occlusions which makes harder the segmentation of the various parts of the hand and the extraction of high-level features. c) Speed processing: even for a single image, a real-time computer vision system needs to process a lot of data. d) Uncontrolled Environments: for general use, many HCI systems must operate with unrestricted backgrounds and a wide range of lighting conditions. e) Rapid hand movements: hand has a very fast movement capability reaching speeds up to 5 m / s for translation and 300 m / s for the rotation of the wrist.

In this work, we propose a new model-based approach to address hands 3D tracking. The observations come from a low-cost 3D sensor (Kinect). The optimization is performed with a variant of growing neural networks GNG. We hope to achieve accurate and robust tracking with an acquisition frequency of at least 10-15Hz. Among the contributions of the proposed method is novel because (a) provides accurate solutions to the problem of 3D tracking of the hand (b) requires no complex hardware configuration (c) is based solely on visual data without using physical markers (d) is not very sensitive to lighting conditions and (e) run in real time.

The rest of the paper is organized as follows: section 2 presents the growing neural gas and describes its ability to obtain reduced representations of objects preserving the topological information. Also extend such capabilities to the representation of point cloud sequences and motion analysis. In section 3, the hand pose estimation system is described. Finally, section 4 presents experiments on the application to a virtual mirror writing system and a hand gesture recognition system, followed by our conclusions and future work.

2 Topological Representation with Growing Neural Gas

The approach presented in this paper is based on self-organising networks trained using the Growing Neural Gas learning method [5], an incremental training algorithm. The links between the units in the network are established through competitive hebbian learning. As a result, the algorithm can be used in cases where the topological structure of the input pattern is not known a priori and yields topology preserving maps of feature manifold.

2.1 Growing Neural Gas

From the Neural Gas model [6] and Growing Cell Structures [7], Fritzke developed the Growing Neural Gas model, with no predefined topology of a union between neurons. A growth process takes place from minimal network size and new units are inserted successively using a particular type of vector quantisation [8,9]. To determine where to insert new units, local error measures are gathered during the adaptation process and each new unit is inserted near the unit which has the highest accumulated error. At each adaptation step a connection between the winner and the second-nearest unit is created as dictated by the competitive hebbian learning algorithm. This is continued until an ending condition is fulfilled, as for example evaluation of the optimal network topology based on some measure. Also the ending condition could it be the insertion of a predefined number of neurons or a temporal constrain. In addition, in GNG networks learning parameters are constant in time, in contrast to other methods whose learning is based on decaying parameters.

2.2 Point Cloud Data Representation with Growing Neural Gas

The ability of neural gases to preserve the input data topology will be employed in this work for the representation and tracking of objects. Identifying the points of the input data that belong to the objects allows the network to adapt its structure to this input subspace, obtaining an induced Delaunay triangulation of the object.

Learning takes place following the GNG algorithm described in previous section. So, doing this process, a representation based on the neural network structure is obtained which preserves the topology of the object O from a certain feature $\mathcal{T}$. That is, from the visual appearance $\mathcal{A}_{\mathcal{V}}$ of the object is obtained an approximation to its geometric appearance $\mathcal{A}_{\mathcal{G}}$. In our case the 3D hands representation.

GNG has been adapted to represent Point Cloud Sequences. The main difference with the GNG original algorithm is the omission of insertion/deletion actions after the first frame. Since no neurons are added or deleted the system keeps the correspondence during the whole sequence, solving intrinsically the problem of correspondence. This adaptive method is also able to face real-time constraints, because the number λ of times that the internal loop is performed can be chosen according to the time available between two successive frames that depend on the acquisition rate. The mean time to obtain a GNG on a frame is about 10ms., using the adaptive method.

GNG provides a reduction of the input data, while preserving its structure. This gives us two advantages. First, we have to process less data, so we speed up the next step of feature extraction. Second, outliers are reduced. Outliers are one of the main sources of error in this kind of applications.

2.3 Motion Representation and Analysis

The motion can be classified according to the way of perceiving it: common and relative, and it can be represented using the graph obtained from the neural network structure for each input data acquisition.

In the case of the common motion, it can be performed the analysis of the trajectory followed by an object by tracking the centroid of the same along the sequence. This centroid may be calculated from the positions of the nodes in the graph that represents the object in each capture (GNG structure).

To follow the relative motion, it should be calculated the changes in position of each node with respect to the centroid of the object for each capture. By following the path of each node, it can be analyzed and recognized changes in the morphology of the object.

One of the most important problems in tracking objects, the correspondence between features along the sequence, can be solved of intrinsic form [10] since the position of neurons is known at any time without the need for additional processing.

The analysis of the trajectory described by each object is used to interpret its movement. In some cases, to address the movement recognition, a trajectory parameterization is performed. In [11], it can be found some suggestions for parameterization. Also, it can be used direct measures of similarity between paths, such as the modified Hausdorff distance [12], allowing comparison of trajectories.

3 3D Hand Pose Estimation with Growing Neural Gas

The previous section described the ability of neural gases for the representation and tracking of data streams in several dimensions. To analyze the movement, certain characteristics will be monitored in each of the shots of each sequence, not the object itself, but its representation that is obtained with the neural network. That is, using the position of the neurons of the network (their structure) as a further feature.

The construction of a system of representation, tracking and recognition of hand gestures based on the GNG and 3D sensors capable of rendering common and relative motion is proposed (figure 1).

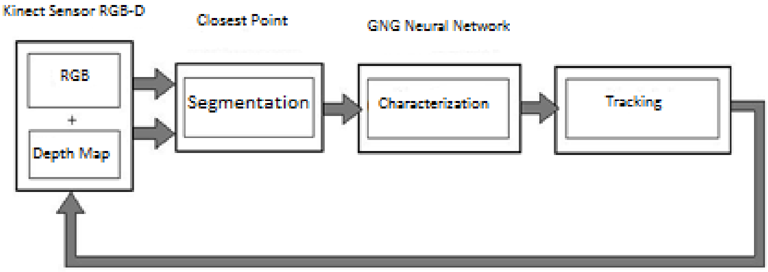


Fig. 1. Flowchart of the whole system.

3.1 Data Acquisition and Hands Segmentation

All the experimental phase is based on the use of real-world sequences obtained by a Kinect sensor. Such sensors belong to the so-called RGB-D cameras since they provide RGB format images with depth information per pixel. Specifically, Microsoft's Kinect sensor is able to get screenshots of 640x480 pixels and its corresponding depth information, based on an infrared projector combined with a CMOS sensor with a resolution of 320x240 pixels, and can reach rates of up to 30 frames per second. A first processing of sensor data enables obtaining the component in the z axis of coordinates of the points in the three dimensional space.

For the segmentation of the hands from the background, an hybrid technique based on depth information and the determination of appropriate thresholds on HSV model to filter skin color points have been developed. The system has been trained with multiple users.

3.2 Gesture Characterization and Recognition with Growing Neural Networks

Growing neural gas presented in section 2 is used to characterize hands allowing a reduced topological representation by a graph defining an induced Delaunay triangulation. In the experiments section, the minimum number of neurons necessary for an adequate representation of the hand is defined, which allows the subsequent tracking and recognition of gestures. Figure 2 shows an example of 3D hand characterization.

The path followed by neurons can be obtained and interpreted by processing the position information of them on the map along the sequence. This evolution can be studied at the level of the global movement, following the centroids of the map or locally, studying the deformation of the object. This is possible because the system does not restart the map for each shot, adapting the new input data to the previous map without inserting or deleting neurons. Thus, the neurons are used as stable visual markers that define the shape of objects.

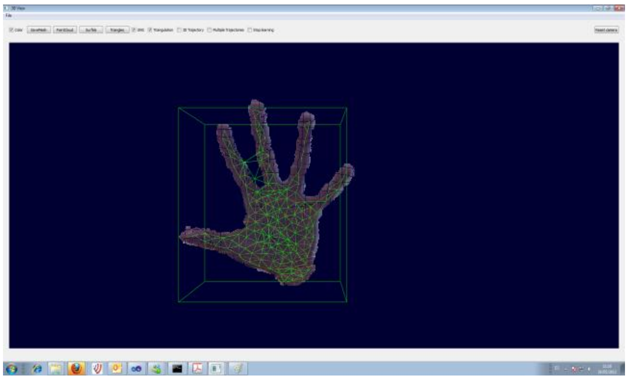


Fig. 2. Hand pose characterization with GNG.

4 Experimentation

Several experiments have been performed for the validation of our system. At first, it has been obtained a proper parametrization of the neural network and then a global motion experiments related to virtual mirror writing have been conducted. Finally, the system have learned and labeled various poses of the hand, performed by various users, and subsequently the same gestures, made by other users, have been presented to the system, with a high rate of correct recognition.

To carry out all the experiments we used the RGBDemo framework, a computer vision software written in C++, which allow to quickly and easily get started with Kinect-like camera and develop our own projects.

4.1 Optimal Parametrization of GNG

This experiment measured the mean square error of different representations of the hand obtained with a varied number of neurons. From the graph of figure 3, it can be noticed that with approximately 200 neurons, the error is low enough and the quality provided is adequate for 3D hands representation. We chose the minimum number of neurons with an acceptable quality as the computational cost is reduced and allows real-time management. The other GNG parameters used were: $\lambda = 2000$, $\varepsilon_1 = 0.1$, $\varepsilon_2 = 0.01$, $\alpha = 0.5$, $\beta = 0.0005$.

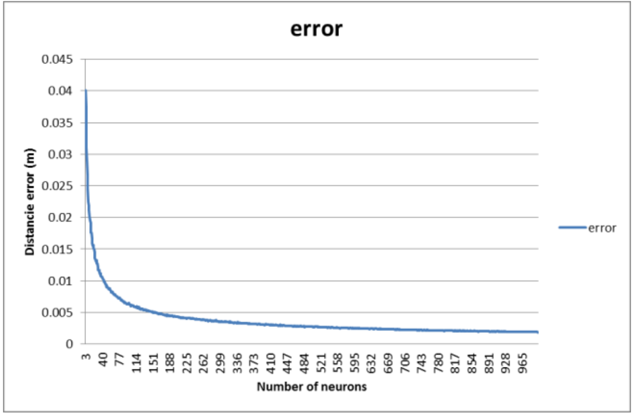
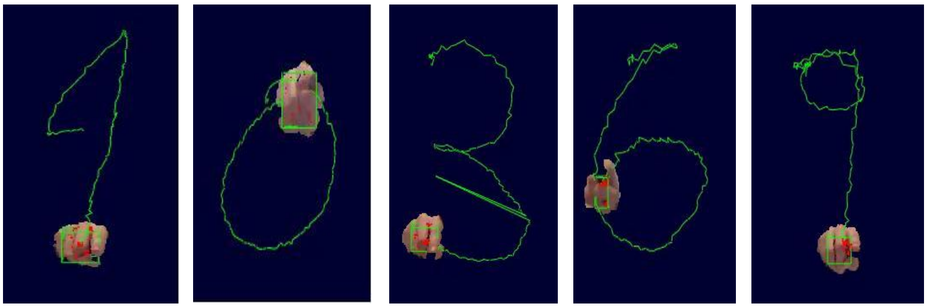


Fig. 3. Error representation based on the number of neurons in the network.

4.2 Virtual Mirror Writing Recognition

This section presents an application of a handwriting recognition system in a virtual mirror by tracking the trajectory of the centroid. Figure 4 shows a set of gestures.



(a) (b) (c) (d) (e)

Fig. 4. Mirror writing gestures corresponding to the numbers 0, 1, 3, 6 and 9.

Through a training phase, the path described by the centroid of the neural network representation of different characters virtually made by different users has been tagged. Subsequently, there has been conducted a recognition phase of the symbols created by new users by comparing the paths described and labeled using the Hausdorff distance [12] , with a success rate greater than 95%.

4.3 Recognition of Hand Poses

This section presents an application of the gesture recognition system. Figure 5 shows a set of gestures used as input to the system. Once the sequences have been learned by GNG, Hausdorff distances will be used for comparison of sets of points that define a path, in this case the ones followed by all neurons representing the gestures of the hand with respect to the centroid of the representation.

At figure 5, the pose of the left defines the start position and the center ones are different final positions for different gestures. The representation on the right shows the trajectories described by the neurons along the gesture realization.

As in the previous experiment stages of training / recognition and labeling were made. The results obtained for a reduced set of gestures are promising with a success rate of 95%.

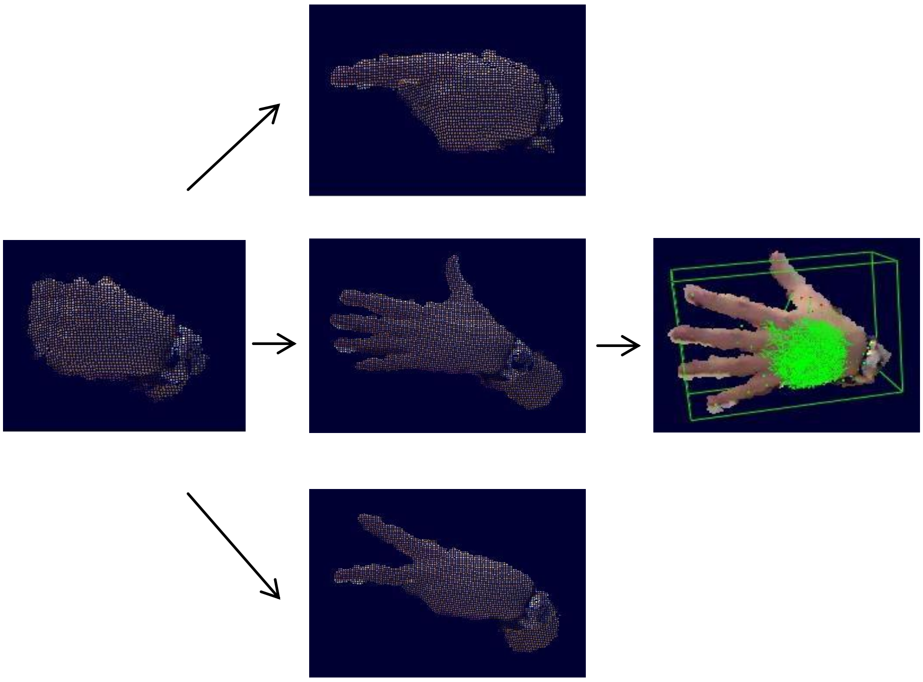
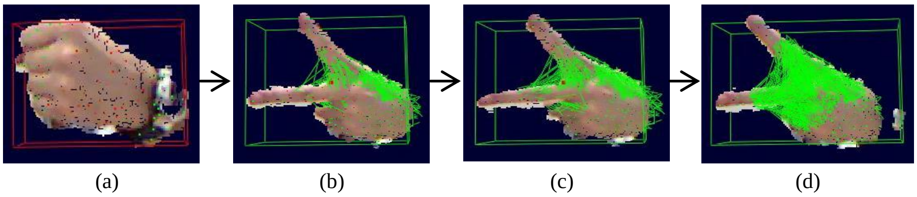


Fig. 5. Set of gestures used for trajectories study.

In this case, it is especially important to be able to represent 3D space, as can be seen in figure 6, since some gestures present trajectories in all axes that would be impossible to perceive with a system based only on the x and y axes.



(a) (b) (c) (d)

Fig. 6. 3D trajectories evolution during the gesture.

5 Conclusions and Future Work

In this paper we have presented a novel architecture to represent and recognize gestures based on neural networks and 3D sensors.

It has been shown that neural gases are suitable for reduced representation of objects in three dimensions by a graph which defines the interconnection of neurons. Moreover, the processing of the position information of these neurons over time, allows us to build the hand trajectories and interpret them.

For the recognition of gestures we used the Hausdorff distance to measure the similarity between the sets of points that define the different trajectories global and / or local of our markers (neurons).

Finally, to validate the system it have been developed a framework that have been used to test several global and local gestures made by different users, obtaining good results in the recognition task. However, the noisy images and occlusions with the environment are two major problems.

As future work, we will improve the system performance at all stages to achieve a natural interface that allows us to interact with any object manipulation system. Likewise, it is contemplated the acceleration of the whole system on GPUs.

References

1. I. Oikonomidis, N. Kyriazis, and A.A. Argyros. Full DOF Tracking of a Hand Interacting with an Object by Modeling Occlusions and Physical Constraints. In ICCV, (2011)
2. I. Oikonomidis, N. Kyriazi, A.A. Argyros. Efficient Model-based 3D Tracking of Hand Articulations using Kinect. BMVC (2011)
3. T.B. Moeslund, A. Hilton, V. Kruger: A survey of advances in vision-based human motion capture and analysis. CVIU 104 (2006) 90-126
4. A. Erol, G. Bebis, M. Nicolescu, R.D. Boyle, X. Twombly: Vision-based hand pose estimation: A review. CVIU 108 (2007) 52-73
5. Fritzke, B (1995). A Growing Neural Gas Network Learns Topologies, *In Advances in Neural Information Processing Systems 7, G. Tesauro, D.S. Touretzky y T.K. Leen (eds.), MIT Press, Cambridge, Mass*
6. Martinetz, T., Berkovich, S.G., & Schulten, K.J. (1993). “Neural-Gas” Network for Vector Quantization and its Application to Time-Series Prediction, *IEEE Transactions on Neural Networks*, 4(4):558-569
7. Fritzke, B. (1993). Growing Cell Structures – A Self-organising Network for Unsupervised and Supervised Learning, *Technical Report TR-93-026, International Computer Science Institute, Berkeley, California*
8. Bauer, H.-U., Hermann, M., & Villmann, T. (1999). Neural Maps and Topographic Vector Quantization. *Neural Networks*, 12(4–5), 659–676
9. Martinetz, T., & Schulten, K. (1994). Topology Representing Networks. *Neural Networks*, 7(3) 507-522
10. Zhang, Z (1993). Le problème de la mise en correspondance: L’état de l’art. Rapport de recherche nº 2146, Institut National de Recherche en Informatique et en Automatique.
11. Cédras, C. & Shah, M. (1995). Motion-based recognition: a survey, Image and Vision Computing, 13(2): 129-155
12. Dubbuison, M.P. & Jain, A.K. (1994). A Modified Hausdorff Distance for Object Matching, In Proceedings of the International Conference on Pattern Recognition, Jerusalem, Israel. 566-568.