Hand Gesture Recognition Via a New Self-organized Neural Network

E. Stergiopoulou, N. Papamarkos[1], and A. Atsalakis

[1] Image Processing and Multimedia Laboratory,
Department of Electrical & Computer Engineering,
Democritus University of Thrace,
67100 Xanthi, Greece
papamark@ee.duth.gr

Abstract. A new method for hand gesture recognition is proposed which is based on an innovative Self-Growing and Self-Organized Neural Gas (SGONG) network. Initially, the region of the hand is detected by using a color segmentation technique that depends on a skin-color distribution map. Then, the SGONG network is applied on the segmented hand so as to approach its topology. Based on the output grid of neurons, palm geometric characteristics are obtained which in accordance with powerful finger features allow the identification of the raised fingers. Finally, the hand gesture recognition is accomplished through a probability-based classification method.

1 Introduction

Hand gesture recognition is a promising research field in computer vision. Its most appealing application is the development of more effective and friendly interfaces for human–machine interaction, since gestures are a natural and powerful way of communication. Moreover, it can be used to teleconferencing and telemedicine, because it doesn't require any special hardware. Last but not least, it can be applied to the interpretation and the learning of the sign language.

Hand gesture recognition is a complex problem that has been dealt with many different ways. Huang et al. [1] created a system consisting of three modules: i) model based hand tracking that uses the Hausdorff distance measure to track shape–variant hand motion, ii) feature extraction by applying the scale and rotation invariant Fourier descriptor and iii) recognition by using a 3D modified Hopfield neural network (HNN). Huang et al. [2] developed also another model based recognition system that consists of three stages as well: i) feature extraction based on spatial (edge) and temporal (motion) information, ii) training that uses the principal component analysis (PCA), the hidden Markov model (HMM) and a modified Hausdorff distance and iii) recognition by applying the Viterbi algorithm. Yin et al. [3] used a RCE neural network based color segmentation algorithm for hand segmentation, extracted edge points of fingers as points of interest and matched them based on the topological features of the hand, such as the centre of the palm. Herpers et al. [4] used a hand

M. Lazo and A. Sanfeliu (Eds.): CIARP 2005, LNCS 3773, pp. 891 – 904, 2005.

segmentation algorithm that detects connected skin–tone blobs in the region of interest. A medial axis transform is applied, and finally, an analysis of the resulting image skeleton allows the gesture recognition.

In the proposed method, hand gesture recognition is divided into four main phases: the detection of the hand's region, the approximation of its topology, the extraction of its features and its identification. The detection of the hand's region is achieved by using a color segmentation technique based on a skin color distribution map in the YCbCr space [6-7]. The technique is reliable, since it is relatively immune to changing lightning conditions and provides good coverage of the human skin color. It is very fast and doesn't require post–processing of the hand image. Once the hand is detected, a new Self-Growing and Self-Organized Neural Gas (SGONG) [8] network is used in order to approximate its topology. The SGONG is an innovative neural network that grows according to the hand's morphology in a very robust way. The positions of the output neurons of the SGONG network approximate the shape and the structure of the segmented hand. That is, as it can be viewed in Fig. 1(c), the grid of the output neurons takes the shape of the hand. Also, an effective algorithm is developed in order to locate a gesture's raised fingers, which is a necessary step of the recognition process. In the final stage, suitable features are extracted that identify, regardless to the hand's slope, the raised fingers, and therefore, the corresponding gesture. Finally, the completion of the recognition process is achieved by using a probability–based classification method.

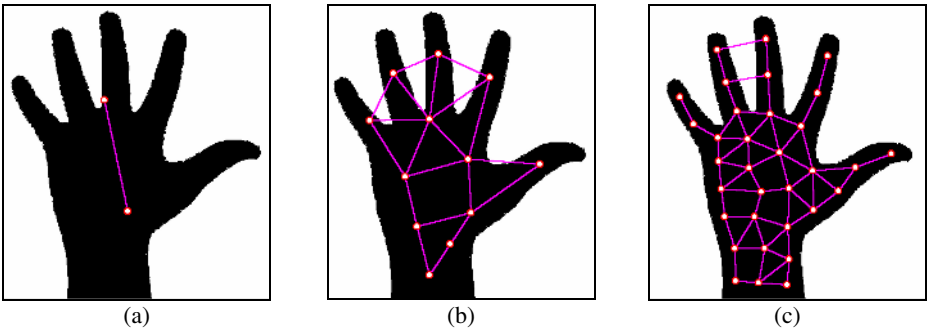


Fig. 1. Growth of the SGONG network: (a) starting point, (b) a growing stage, (c) the final output grid of neurons

The proposed gesture recognition system has been trained to identify 26 hand gestures. It has been tested by using a large number of gestures and the achieved recognition rate is satisfactory.

2 Description of the Method

The purpose of the proposed gesture recognition method is to recognize a set of 26 hand gestures. The principal assumption is that the images include exactly one hand.

Furthermore, the gestures are made with the right hand, the arm is roughly vertical, the palm is facing the camera and the fingers are either raised or not. Finally, the image background is plain, uniform and its color differs from the skin color.

The entire method consists of the following four main stages:

- Color Segmentation
- Application of the Self-Growing and Self-Organized Neural Gas Network
- Finger Identification
- Recognition Process

Analysis of these stages follows.

2.1 Color Segmentation

The detection of the hand region can be achieved through color segmentation. The aim is to classify the pixels of the input image into skin color and non-skin color clusters. This can be accomplished by using a thresholding technique that exploits the information of a skin color distribution map in an appropriate color space.

It is a fact that skin color varies quite dramatically. First of all, it is vulnerable to changing lightning conditions that obviously affect its luminance. Moreover, it differs among people and especially among people from different ethnic groups. The perceived variance, however, is really a variance in luminance due to the fairness or the darkness of the skin. Researchers, also, claim that the skin chromaticity is the same for all races [5]. So regarding to the skin color, luminance introduces many problems, whereas chromaticity includes the useful information. Thus, proper color spaces for skin color detection are those that separate luminance from chromaticity components.

The proposed color space is the YCbCr space, where Y is the luminance and Cb, Cr the chrominance components. RGB values can be transformed to YCbCr color space using the following equation [6-7]:

$$\begin{bmatrix} Y \\ Cb \\ Cr \end{bmatrix} = \begin{bmatrix} 16 \\ 128 \\ 128 \end{bmatrix} + \begin{bmatrix} 65.481 & 128.553 & 24.966 \\ -37.797 & -74.203 & 112 \\ 112 & -93.786 & -18.214 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix} \quad (1)$$

Given that the input RGB values are within range [0,1] the output values of the transformation will be [16, 235] for Y and [16, 240] for Cb and Cr. In this color space, a distribution map of the chrominance components of skin color was created, by using a test set of 50 images. It is found that Cb and Cr values are narrowly and consistently distributed. Particularly, the ranges of Cb and Cr values are, as shown in Fig. 2, R_{Cb} = [80, 105] and R_{Cr} = [130, 165], respectively. These ranges were selected very strictly, in order to minimize the noise effect and maximize the possibility that the colors correspond to skin.

Let C_{bi} and C_{ri} be the chrominance components of the i-th pixel. If $C_{bi} \in R_{Cb}$ and $C_{ri} \in R_{Cr}$, then the pixel belongs to the hand region.

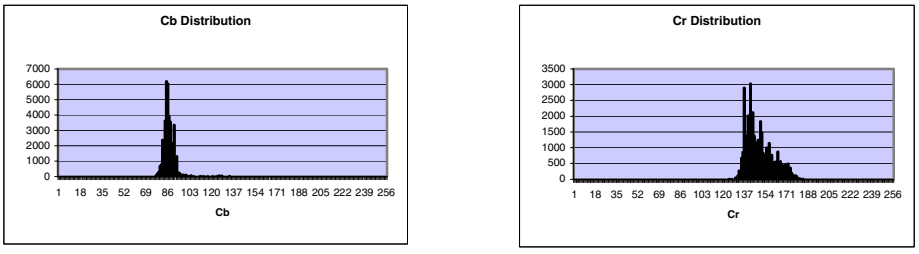
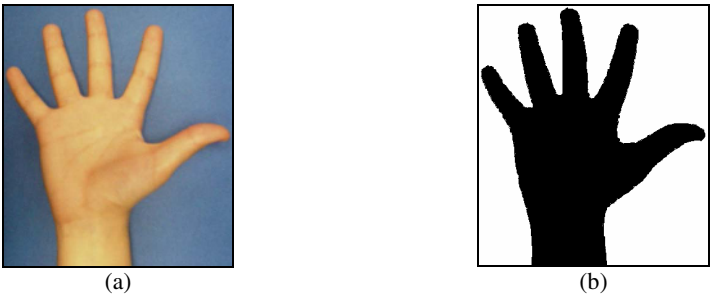


Fig. 2. Distribution of: Cb component and Cr component

Finally, a thresholding technique completes the color segmentation of the input image. The technique consists of the following steps.

- Calculation of the Euclidean distance between the C_{bi}, C_{ri} values and the edges of R_{Cb} and R_{Cr}, for every pixel.
- Comparison of the Euclidean differences with a proper threshold. If at least one difference is less than the threshold, then the pixel belongs to the hand region. The proper threshold's value is taken equal to 18.

The output image of the color segmentation process is considered as binary. As illustrated in Fig. 3 the hand region, that is the region of interest, became black and the background white. The hand region is normalized to certain dimensions so as the system to be invariant of the hand's size. It is worth to underline also, that the segmentation results are very good (almost noiseless) without further processing (e.g. filtering) of the image.



(a) (b)

Fig. 3. (a) Original image, (b) Segmented image

2.2 Application of the Self-growing and Self-organized Neural Gas Network

The next stage of the recognition process is the application of the Self Growing and Organized Neural Gas (SGONG) [8] on the segmented (binary) image.

The SGONG is an unsupervised neural classifier. It achieves clustering of the input data, so as the distance of the data items within the same class (intra-cluster variance) is small and the distance of the data items stemming from different classes (inter-

cluster variance) is large. Moreover, the final number of classes is determined by the SGONG during the learning process. It is an innovative neural network that combines the advantages both of the Kohonen Self-Organized Feature Map (SOFM) and the Growing Neural Gas (GNG) neural classifiers.

The SGONG consists of two layers, i.e. the input and the output layer. It has the following main characteristics:

a. Is faster than the Kohonen SOFM,
b. The dimensions of the input space and the output lattice of neurons are always identical. Thus, the structure of neurons in the output layer approaches the structure of the input data,
c. Criteria are used to ensure fast converge of the neural network. Also, these criteria permit the detection of isolated classes.

The coordinates of the output neurons are the coordinates of the classes' centers. Each neuron is described by two local parameters, related to the training ratio and to the influence by the neighbourhood neurons. Both of them decrease from a high to a lower value during a predefined local time in order to gradually minimize the neurons' ability to adapt to the input data. As it is shown in Fig. 1, the network begins with only two neurons and it inserts new neurons in order to achieve better data clustering. Its growth is based on the following criteria:

- A neuron is inserted near the one with the greatest contribution to the total classification error, only if the average length of its connections with the neighbor neurons is relatively large.
- A neuron is removed if no input vector is classified to its cluster for a predefined number of epochs.
- All neurons are classified according to their importance. The less valuable neuron is removed, only if the subsequent increase in the mean classification error is less than a predefined value.
- A neuron is removed, if it belongs to an empty class.
- The connections of the neurons are created dynamically by using the "Competitive Hebbian Learning" method.

The main characteristic of the SGONG is that both neurons and their connections approximate effectively the input data's topology. This is the exact reason for using the specific neural network in this application. Particularly, the proposed method uses the coordinates of random samples of the binary image as the input data. The network grows gradually on the black segment, i.e. the hand region and a structure of neurons and their connections is finally, created that describes effectively the hand's morphology. The output data of the network, in other words, is an array of the neurons' coordinates and an array of the neurons' connections. Based on this information important finger features are extracted.

2.3 The Training Steps of the SGONG Network

The training procedure for the SGONG neural classifier starts by considering first two output neurons ($c = 2$). The local counters N_i , $i = 1,2$ of created neurons are set to

zero. The initial positions of the created output neurons, that is, the initial values for the weight vectors W_i, $i=1,2$ are initialized by randomly selecting two different vectors from the input space. All the vectors of the training data set X' are circularly used for the training of the SGONG network.

The training steps of the SGONG are as follows:

Step 1. At the beginning of each epoch the accumulated errors $AE_i^{(1)}$, $AE_i^{(2)}$, $\forall i \in [1,c]$ are set to zero. The variable $AE_i^{(1)}$ expresses, at the end of each epoch, the quantity of the total quantization error that corresponds to $Neuron_i$, while the variable $AE_i^{(2)}$, represents the increment of the total quantization error that we would have if the $Neuron_i$ was removed.

Step 2. For a given input vector X_k, the first and the second winner neurons $Neuron_{w1}$, $Neuron_{w2}$ are obtained:

$$\text{for } Neuron_{w1}: \left\|X_k - W_{w1}\right\| \le \left\|X_k - W_i\right\|, \ \forall i \in [1,c] \tag{2}$$

$$\text{for } Neuron_{w2}: \left\|X_k - W_{w2}\right\| \le \left\|X_k - W_i\right\|, \ \forall i \in [1,c] \text{ and } i \ne w1 \tag{3}$$

Step 3. The local variables $AE_{w1}^{(1)}$ and $AE_{w1}^{(2)}$ change their values according to the relations:

$$AE_{w1}^{(1)} = AE_{w1}^{(1)} + \left\|X_k - W_{w1}\right\| \tag{4}$$

$$AE_{w1}^{(2)} = AE_{w1}^{(2)} + \left\|X_k - W_{w2}\right\| \tag{5}$$

$$N_{w1} = N_{w1} + 1 \tag{6}$$

Step 4. If $N_{w1} \le N_{idle}$ then the local learning rates $\varepsilon 1_{w1}$ and $\varepsilon 2_{w1}$ change their values according to equations (7), (8) and (9). Otherwise, the local learning rates have the constant values $\varepsilon 1_{w1} = \varepsilon 1_{\min}$ and $\varepsilon 2_{w1} = 0$.

$$\varepsilon 2_{w1} = \varepsilon 1_{w1} / r_{w1} \tag{7}$$

$$\varepsilon 1_{w1} = \varepsilon 1_{\max} + \varepsilon 1_{\min} - \varepsilon 1_{\min} \cdot \left(\frac{\varepsilon 1_{\max}}{\varepsilon 1_{\min}}\right)^{\frac{N_{w1}}{N_{idle}}} \tag{8}$$

$$r_{w1} = r_{\max} + 1 - r_{\max} \cdot \left(\frac{1}{r_{\max}}\right)^{\frac{N_{w1}}{N_{idle}}} \tag{9}$$

The learning rate $\varepsilon 1_i$ is applied to the weights of $Neuron_i$ if this is the winner neuron ($w1 = i$), while $\varepsilon 2_i$ is applied to the weights of $Neuron_i$ if this belongs to the neighborhood domain of the winner neuron ($i \in nei(w1)$). The learning rate $\varepsilon 2_i$ is used in order to have soft competitive effects between the output neurons. That is, for each output neuron, it is necessary that the influence from its neighboring neurons to be gradually reduced from a maximum to a minimum value. The values of the learning rates $\varepsilon 1_i$ and $\varepsilon 2_i$ are not constant but they are reduced according to the local counter N_i. Doing this, the potential ability of moving of neuron i toward an input vector (plasticity) is reduced with time. Both learning rates change their values from maximum to minimum in a period, which is defined by the N_{idle} parameter. The variable r_{wi} initially takes its minimum value $r_{\min} = 1$ and in a period, defined by the N_{idle} parameter, reaches its maximum value r_{max}.

Step 5. In accordance with the Kohonen SOFM, the weight vector of the winner neuron $Neuron_{w1}$ and the weight vectors of its neighboring neurons $Neuron_m$, $m \in nei(w1)$, are adapted according to the following relations:

$$W_{w1} = W_{w1} + \varepsilon 1_{w1} \cdot \left(X_k - W_{w1} \right) \tag{10}$$

$$W_m = W_m + \varepsilon 2_m \cdot \left(X_k - W_m \right), \ \forall m \in nei(w1) \tag{11}$$

Step 6. With regard to generation of lateral connections, SGONG employs the following strategy. The CHR is applied in order to create or remove connections between neurons. As soon as the neurons $Neuron_{w1}$ and $Neuron_{w2}$ are detected, the connection between them is created or is refreshed. That is

$$s_{w1,w2} = 0 \tag{12}$$

With the purpose of removing of superfluous lateral connections, the age of all connections emanating from $Neuron_{w1}$, except the connection with $Neuron_{w2}$, is increased by one:

$$s_{w1,m} = s_{w1,m} + 1, \ \forall m \in nei(w1), \text{ with } m \neq w2 \tag{13}$$

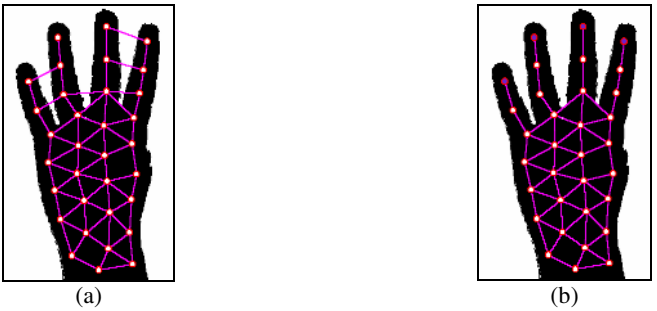
Step 7. At the end of each epoch it is examined if all neurons are in *idle state*, or equivalently, if all the local counters N_i, $\forall i \in [1,c]$ are greater than the predefined value N_{idle} and the neurons are well trained. In this case, the training procedure stops, and the convergence of SGONG network is assumed. The number of input vectors needed for a neuron to reach the "*idle state*" influences the convergence speed of the proposed technique. If the training procedure continues, the lateral connections between neurons with age greater than the maximum value α are removed. Due to

dynamic generation or removal of lateral connections, the neighborhood domain of each neuron changes with time in order to include neurons that are topologically adjacent.

2.4 Finger Identification

2.4.1 Determination of the Raised Fingers' Number

An essential step for the recognition is to determine the number of fingers that a gesture consists of. This is accomplished by locating the neurons that correspond to the fingertips.



(a) (b)

Fig. 4. (a) Hand image after the application of the SGONG network, (b) hand image after the location of the raised fingers

Observations of the structure of the output neurons' grid leads to the conclusion that fingertip neurons are connected to neighbourhood neurons by only two types of connections: i) connections that go through the background, and ii) connections that belong exclusively only to the hand region. The crucial point is that fingertip neurons use only one connection of the second type. Based on this conclusion, the determination of the number of fingers is:

- Remove all the connections that go through the background.
- Find the neurons that have only one connection. These neurons are the fingertips, as indicated in Fig. 4.
- Find successively the neighbor neurons. Stop when a neuron with more than two connections is found. This is the finger's last neuron (root-neuron).

Find the fingers' mean length (i.e. the mean fingertip and root neuron distance). If a finger's length differs significantly from the mean value then it is not considered to be a finger.

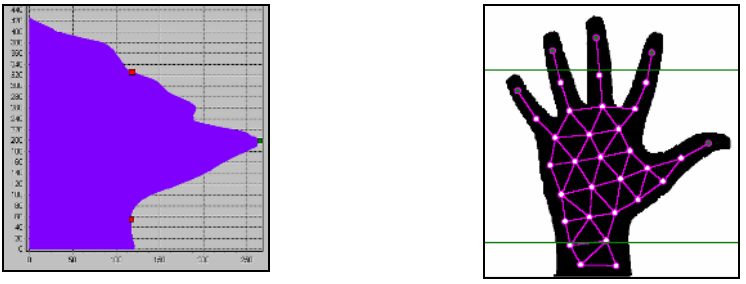
2.4.2 Extraction of Hand Shape Characteristics

Palm Region

Many images include redundant information that could reduce the accuracy of the extraction techniques and lead to false conclusions. Such an example is the presence

of a part of the arm. Therefore, it is important to find the most useful hand region, which is the palm.

The algorithm of finding the palm region is based on the observation that the arm is thinner than the palm. Thus, a local minimum should appear at the horizontal projection of the binary image. The minimum defines the limits of the palm region as it is shown in Fig. 5.



(a) (b)

Fig. 5. (a) Horizontal projection, (b) Palm region

This procedure is as follows:

- Create the horizontal projection of the image *H[j]*:
- Find the global maximum $H\left[j^{\max} \right]$ and the local minima $H\left[j_i^{\min} \right]$ of $H[j]$.
- Calculate the slope of the lines segments connecting the global maximum and the local minima, which satisfy the condition $j_i^{\min} < j^{\max}$. The minimum j_{lower} that corresponds to the greatest of these slopes defines the lower limit of the palm region, only if its distance from the maximum is greater than a threshold value equal to ImageHeight/6.
- The point that defines the upper limit of the palm region is denoted as j_{upper} and is obtained by the following relation:

$$H\left[j_{upper} \right] \le H\left[j_{lower} \right] \quad \text{and} \quad j_{upper} > j^{\max} > j_{lower} \tag{14}$$

Palm Centre

The coordinates of the centre of the palm are taken equal to the mean values of the coordinates of the neurons that belong to the palm region.

Hand Slope

Despite of the roughly vertical direction of the arm, the slope of the hand varies. This fact should be taken under consideration because it affects the accuracy of the finger features, and consequently, the efficiency of the identification process. The recognition results depend greatly on the correct calculation of the hand slope.

The hand slope can be estimated by the angle of the left side of the palm, as it can be viewed in Fig. 6(a). The technique consists of the following steps:

- Find the neuron N_{Left}, which belongs to the palm region and has the smallest horizontal coordinate.
- Obtain the set of palm neurons N_{set} that belong to the left boundary of the neurons grid. To do this, and for each neuron, starting from the N_{Left}, we obtain the neighborhood neuron which has, simultaneously, the smallest vertical and horizontal coordinates.
- The first and the final neurons of the set N_{set} define the hand slope line (HSL) which angle with the horizontal axis is taken equal to the hand's slope.

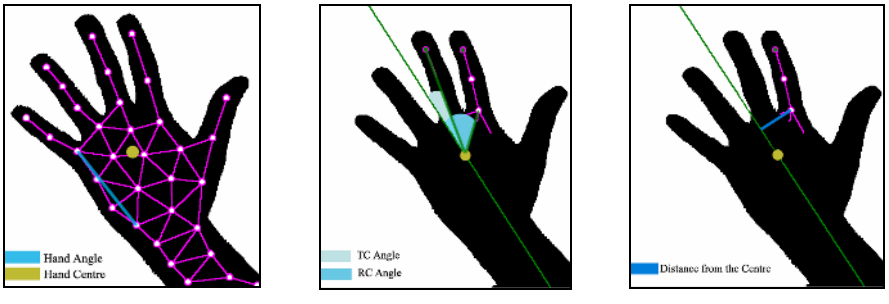
The hand slope is considered as a reference angle and is used in order to improve the feature extraction techniques.

2.4.3 Extraction of Finger Features

Finger Angles

A geometric feature that individualizes the fingers is their, relative to the hand slope, angles. As it is illustrated in Fig. 6(b), we extract two finger angles.

- RC Angle. It is an angle formed by the HSL and the line that joints the root neuron and the hand center. It is used directly for the finger identification process.
- TC Angle. It is an angle formed by the HSL and the line that joints the fingertip neuron and the hand center. This angle provides the most discrete values for each finger and thus is valuable for the recognition.



(a) (b) (c)

Fig. 6. (a) Hand slope and centre, (b) Fingers' angles, (c) Distance from the centre

Distance from the Palm Centre

A powerful feature for the identification process is the vertical distance of the finger's root neuron from the line passing through the palm centre and having the same slope as the HSL. An example is illustrated in Fig. 6(c).

3 Recognition Process

The recognition process is actually a choice of the most possible gesture. It is based on a classification process of the raised fingers into five classes (thumb, index, middle, ring, little) according to their features. The classification depends on the probabilities of a finger to belong to the above classes. The probabilities derive from the features distributions. Therefore, the recognition process consists of two stages: the off–line creation of the features distributions and the probability based classification.

3.1 Features Distributions

The finger features are naturally occurring features, thus a Gaussian distribution can model them. Their distributions are created by using a test set of 100 images from different people.

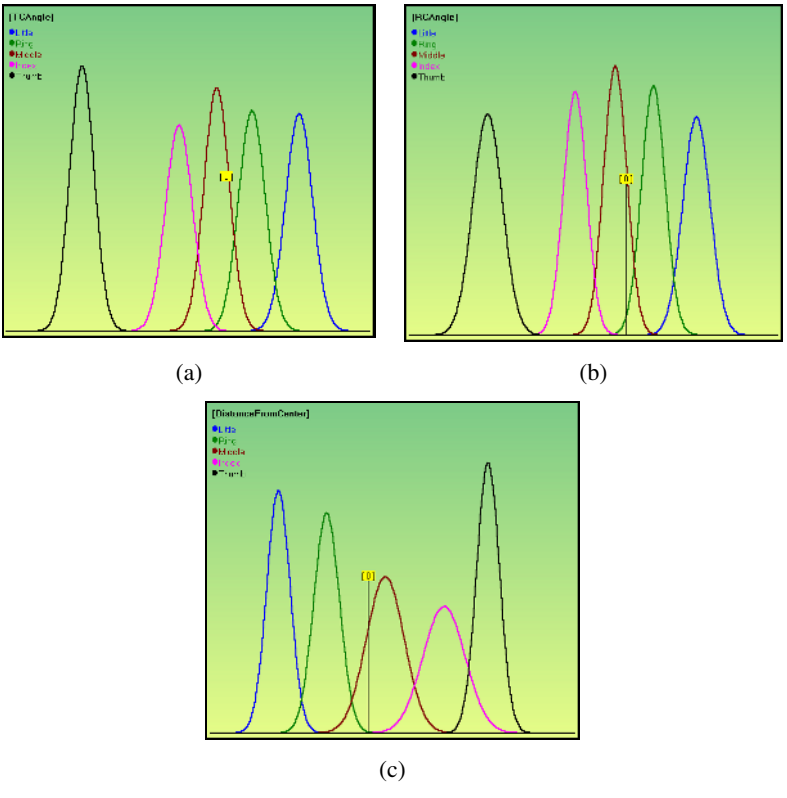


Fig. 7. Features distributions (a) TC Angle, (b) RC Angle, (c) Distance from the centre

If f_i is the i-th feature ($i \in [1, 3]$), then its Gaussian distributions for every class c_j ($j \in [1, 5]$) are given by the relation:

$$p_{fi}^{c_j}(x) = \frac{e^{-\frac{\left(x - m_{pi}^{c_j}\right)^2}{2\sigma^2}}}{\sigma\sqrt{2\pi}} \tag{15}$$

where, $j = 1,...,5$, m_{fi}^{cj} is the mean value and σ_{fi}^{cj} the standard deviation of the f_i feature of the c_j class. . The Gaussian distributions of the above features are shown in Fig. 7. As it can be observed from the distributions, the five classes are well defined and are well discriminated.

3.2 Classification

The first step of the classification process is the calculation of the probabilities RPc_j of a raised finger to belong to each one of the five classes. Let x_0 be the value of the i-th feature f_i. Calculate the probability $p_{fi}^{cj}(x_0)$ for $i \in [1, 3]$ and $j \in [1, 5]$. The requested probability is the sum of the probabilities of all the features for each class

$$RPc_j = \sum_{i=1}^{3} p_{f_i}^{c_j} \tag{16}$$

where, $j = 1,...,5$, m_{fi}^{cj} is the mean value and σ_{fi}^{cj} the standard deviation of the f_i feature of the c_j class. . The Gaussian distributions of the above features are shown in Fig. 7. As it can be observed from the distributions, the five classes are well defined and are well discriminated.

This process is repeated for every raised finger.

Knowing the number of the raised fingers, one can define the possible gestures that can be created. For each one of these possible gestures the probability score is calculated, i.e. the sum of the gesture' s each raised finger to belong to each one of the classes. Finally, the gesture is recognized as the one with the higher probability score.

4 Experimental Results

The proposed hand gesture recognition system, which was implemented in DELPHI, was tested with 158 test hand images 1580 times. It is trained to recognize up to 26 gestures. The recognition rate, under the conditions described above, is 90.45%. Fig. 8 illustrates recognition examples.

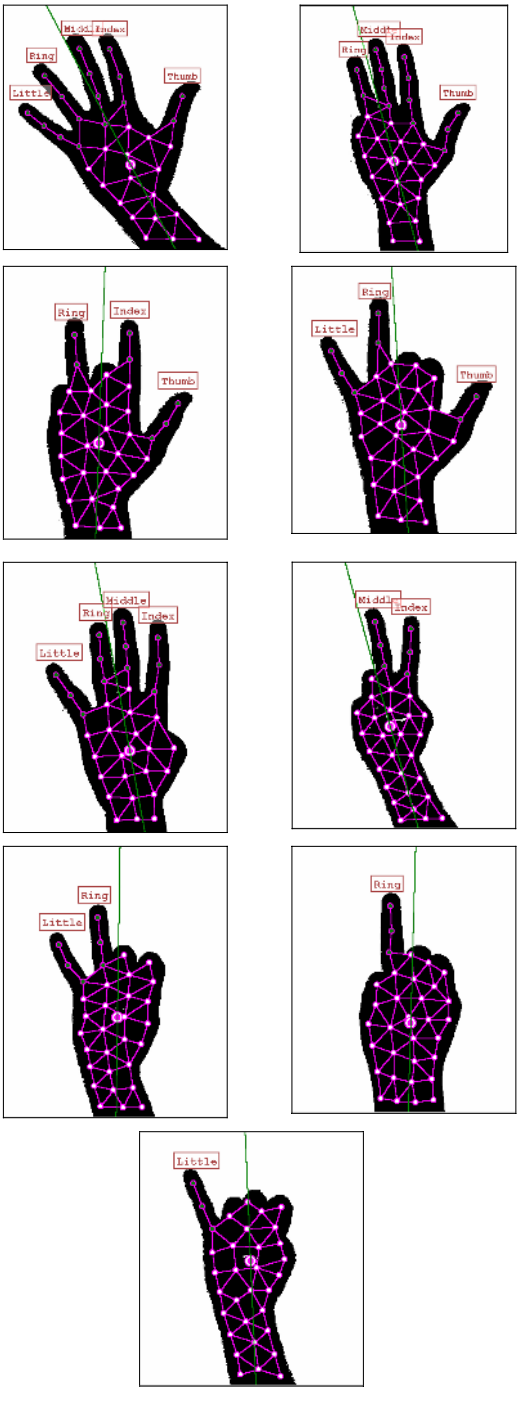


Fig. 8. Gesture recognition examples

5 Conclusions

This paper introduces a new technique for hand gesture recognition. It is based on a color segmentation technique for the detection of the hand region and on the use of the Self-Growing and Self-Organized Neural Gas network (SGONG) for the approximation of the hand's topology. The identification of the raised fingers, which depends on hand shape characteristics and fingers' features, is invariant of the hand's slope. Finally, the recognition process is completed by a probability-based classification with very high rates of success.

References

[1] Huang Chung-Lin, Huang Wen-Yi (1998). Sign language recognition using model-based tracking and a 3D Hopfield neural network. Machine Vision and Applications, 10:292-307. Springer-Verlag.

[2] Huang Chung-Lin, Jeng Sheng-Hung (2001). A model-based hand gesture recognition system. *Machine Vision and Applications*, 12:243-258. Springer-Verlag.

[3] Yin Xiaoming, Xie Ming (2003). Estimation of the fundamental matrix from uncalibrated stereo hand images for 3D hand gesture recognition. *Pattern Recognition*, 36:567-584. Pergamon.

[4] Herpers R., Derpanis K., MacLean W.J., Verghese G., Jenkin M., Milios E., Jepson A., Tsotsos J.K. (2001). SAVI: an actively controlled teleconferencing system. *Image snd Vision Computing*, 19:793-804. Elsevier.

[5] O' Mara David T. J. (2002). Automated Facial Metrology. Ph.D. Thesis, University of Western Australia, Department of Computer Science and Software Engineering.

[6] Chai Douglas, Ngan N. King (1999). Face segmentation using skin color map in videophone applications. *IEEE Transactions on Circuits and Systems for Video Technology,* 551-564.

[7] Chai Douglas, Ngan N. King (Apr. 1998). Locating facial region of a head–and–shoulders color image. Third IEEE International Conference on Automatic Face and Gesture Recognition, Nara, Japan, 124-129 .

[8] Atsalakis Antonis (2004). Colour Reduction in Digital Images. Ph.D. Thesis, Democritus University of Thrace, Department of Electrical and Computer Engineering.