Proceedings of the 17th World Congress
The International Federation of Automatic Control
Seoul, Korea, July 6-11, 2008

Combined Gesture-Speech Recognition and Synthesis Using Neural Networks

C. Roncancio Valencia*. J. Gomez Garcia-Bermejo.**
E. Zalama Casanova**

**Robotics, Machine Vision and Real Time Systems Division. CARTIF*
Parque Tecnologico de Boecillo, Valladolid, Spain; Tel: (+34) 983-54-6504;
(e-mail: catron@cartif.es).
***ETSII, Valladolid University*
Valladolid,Spain, (e-mail:jaigom@eis.uva.es, ezalama@eis.uva.es)

Abstract: *Sign languages such as Spanish Sign Language (LSE) are the primary communication way among members of the Deaf community. However this language is not widely known outside of this community. The techniques for automatic recognizing hand signs proposed in this paper allow creating systems which can help deaf people to communicate with others, by providing them with computer tools for assisted communication or potentially with a fully automatic portable sign-language-to-speech translation system. The aim of this research it to implement a self-organizing neural network based technique for hand sign recognition, using self organizing map (SOM) and learning vector quantization (LVQ) based algorithms. The two classifiers are then combined to make the final decision. SOM and LVQ classifiers are training for the 26 signs of the manual alphabet and the speech synthesizer concatenate the different phones, diphones and triphones stored in the database to generate the right words in artificial speech. Finally, the system is to contribute to the implementation of meaningful human machine interactions in a workstation, mainly for welfare applications.*

Keywords: Computer Vision; Image Prosessing; SOM and LVQ based Classifiers; Sign Language, Speech synthesis.

1. INTRODUCTION

Speech and gestures are the most commonly used means of communication among humans. Yet, when it comes to communicating with computers, the typical home or business user is still bound to devices such as the keyboard and the mouse. While speech recognition systems are finding their way into low-cost computers, there is a real need for gesture recognition systems that provide robust, real-time operation at low cost, so as to be readily available to the typical deaf communities.

In recent years, hand gesture recognition has become a very active research theme because of its potential use in human computer interaction (García I. 2006). Furthermore, a successful hand gesture recognition and translation into sound will provide to the deaf or hard of hearing people a tool to establish an easy communication with other people that unknown these language. This could be a common tool that provides access of the deaf people into the society. Our experimental setup is described as follows: We first propose to apply a combination of histogram analysis-based image segmentation with contrast stretching-based edge detection (Ortiz M. 2005) to efficiently detect hands. A novel approach to hand gesture recognition based on Som/Lvq network (Knerr S. 1992) is then adopted to classify a subset of static hand postures of the Spain Sign Language (LSE), each posture representing a given phoneme, and also to discriminated between hand postures and the image scene background.

In this paper, we propose a system for the robust simultaneous detection of human hand and classification of hand postures of the Spain Sign Language (LSE) in grey scaled images. The system can adapt to slowly varying illumination conditions, and it does not imply any a priori assumption about the presence of a hand (posture) in an image.

1.1 Machine Vision

Recently, so-called "intelligent" environments, in which a range of human activities can be automatically sensed, analysed and "understood" by use of various machine vision technologies (Anil K. 2000) that are the least conspicuously embedded in the environment but are ubiquitous, have been developed (Weiser M. 1991). Meaningful human machine interactions require as a first step the automatic detection of human hands, for higher-level gesture recognition tasks. In particular, such human-machine interfaces allow a human user to control a variety of devices without any physical contact with remote controls, keyboards, etc.

The majority of the analyses realized on manual gestures have been generated by the existence of the sign language. The most common method for description of manual gestures

by means of the artificial vision bears four fundamental components in mind: size of the hand, orientation, place of the joint and movement (Yuntao C. 1995), the last one is the most complex, since it involves the three remaining ones. For this motive this research is based in the 26 static forms of the alphabet, and not in the words.

1.2 Sing Language

The sign language is a natural language of expression, configuration spatial gesture and visual perception, thanks to which the deaf people can establish a channel of basic information for the relationship with the social environment. Inside the sign language we find the manual or dactylology alphabet. For countries of Hispanic speech, this one is known as Latin dactylology alphabet (Fig. 1), it is conformed by 26 signs, which correspond to each of the letters of the alphabet.

Fig. 1. LSE examples

1.3 Speech Synthesis

The synthesis of voice is the artificial production of human speech. The system used for this purpose can be implemented in software or hardware.

While text analysis algorithms are relatively standard, there are three widely different paradigms for waveform synthesis: Concatenation synthesis, formant synthesis, and articulatory synthesis. The architecture of most modern commercial systems is based on concatenation synthesis, in which samples of speech (phones and diphones) are chopped up, stored in a database, and combined and reconfigured to create new sentences. The complexity is high, but the obtained quality is very good (Ahuactzin L. 1999).

2. DESCRIPTION OF THE PROBLEM

At present the language of signs is not only a tool for the persons who present an auditory fault, but also for the interaction with the computers, especially in virtual environments and applications with a high degree of complexity.

The workstation has a camera subject to a board that acts as interface among the human being and the computer (Fig.2). The Hand Sing system, detects the hand on the board, captures the signs and reproduces in artificial speech the words.

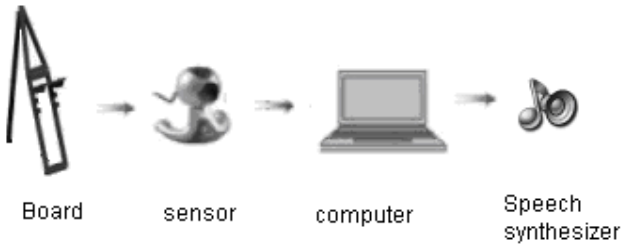

Fig. 2. General configuration of the Hand Sing system

3. DETECTION OF THE HAND BY MEANS OF THERSHOLDING

One of the main objectives of this work is to recognize manual signs, but before this, it is necessary to be located inside the scene, the hand. The implementation of segmentation techniques avoids us to separate the background of the main object in the image, in this case the hand.

Based in the contrast stretching (Ortiz M. 2005), a level of grey u Î[0, L] (L=255 as the maximum value in grey levels) is transformed in another level of grey v (Hany F. 2000), usually of the same range (Fig. 3), by means of :

$$v = T(u) \tag{1}$$

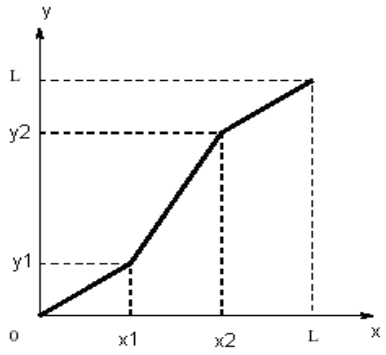

Fig.3. Contrast stretching function

The values of v are situated doing a contrast stretching used to enhance low contrast images by boosting the lighter pixels to a higher intensity level, and doing exactly the opposite to the lower intensity pixels. This transformation gives better quality to the image, with clearer characteristics to discriminate one sign from another one.

The contrast stretching function is defined as:

$$V = \begin{cases} m_0 * u & 0 \le u < x_1 & \text{where } m_0 = y_1 / x_1 \\ m_1 * u & x_1 < u < x_2 & \text{where } m_1 = (y_2 - y_1)/(x_2 - x_1) \\ (u - x_2) + y_2 & x_2 < u \le L & \text{where } m_2 = (L - y_2)/(L - x_2) \end{cases} \tag{2}$$

After pre-processing the image, a threshold T value over the image histogram is defined. Every pixel of the image is compared with this value, identifying if the threshold is exceeded (Ortiz M 2005). The pixels that exceed the threshold are grouped into a *D* vector, which allow to treat this vector as a region. Then the region is explored and its contour is extracted (Extraction of borders).

$$B(i,j) = \begin{cases} \mathrm{D} & if \quad f(i,j) \geq T \\ 0 & if \quad f(i,j) < T \end{cases} \tag{3}$$

The extraction of the contour consist in going pixel by pixel in a determined direction, which depends on if the pixel is inside or outside the region.

In this work, three basic cases are defined for starting the scanning of the contour:

1. Image in which the hand enters by the right side of the board (Fig. 4a). These images are classified as *Scanning by Columns.*
2. Image in which the hand enters by the top of the board (Fig. 4b). These images are classified as *Scanning by the Top*.
3. Image in which the hand enters by the bottom of the board (Fig5 4.c). These images are classified as *Scanning by Rows*.

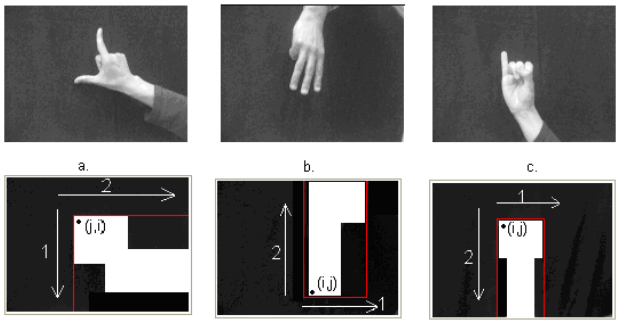

Fig. 4. Hand ways positions. (a) Hand entering by the right side. (b) Hand entering by the top. (c) Hand entering by the bottom.

Finally, when the hand is detected, it is framed inside a red edged region. This sub-image becomes the image to be scanned to determine the hand gesture (Fig. 6a and Fig. 5b).

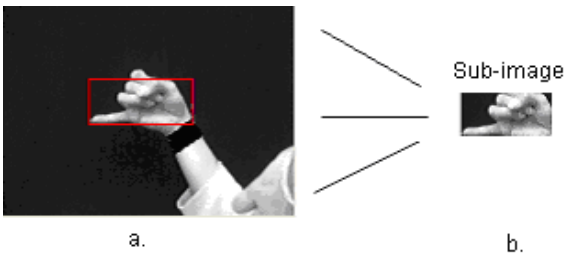

Fig. 5. a.) Image captured. b.) Result image in the detection process (sub-image).

4. HAND GESTURE RECOGNITION

In order to hand gesture recognition be performed a neuronal network is implemented. The network uses a set of 546 patterns, 26 images for any one of the 21 user. Each patter image is in grey levels to reduce the processing time.

The sub-image (Fig. 5b) is scaled in a image of 40X40 pixels. Each of these images are represented in a vector of 1X1600 data, which is obtained by concatenating the rows of the image matrix, this vector is the input to the network. Each element of the vector represents the intensity of each pixel of the image in grey levels [0 255] (Fig. 6).

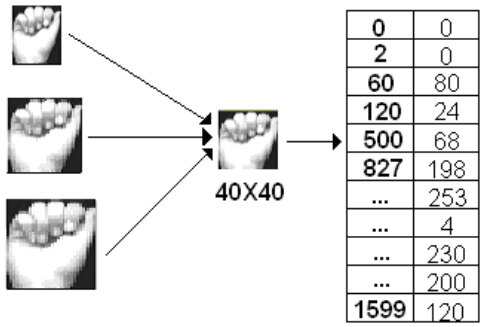

Fig. 6. Image scaling and representation as input vector to SOM+LVQ

It is important to notice that the images of the hand include an area from the wrist up to the fingers in its major extension without the intervention of external characteristics like clothes or accessories (Fig. 5b).

4.1 Som/Lvq network

Different types of neuronal networks exist to generate structures of processing adapted to a particular application. The Kohonen model turns out to be useful for the study of manual signs and belongs to the group of algorithms of codification vector. This model generates an eigenspace to locate in an ideal way a fixed number of vectors in the space of entry, of major dimension than the space of exit to facilitate the compression of information (Haykin S. 1994). Two variants of this model are self-organizing map (SOM) and learning vector quantization (LVQ). Both variants are based on the principle of eigenspaces formation to establish common characteristics between the input vectors of the neuronal network.

4.2 Som/Lvq Algorithms

After the hand has been detected, this sub-image of the scene is extracted as entry to the algorithm of sign recognition. The SOM neuronal network, consists of one input layer (Fig.7) (vector X) of 1600 data, is due to the fact that the input pattern is of 40X40 pixels and one exit layer of 26 neurons, equivalent to the letters of the dactylology alphabet and the *W* vector of 26 x 1600 elements, size defined by the number of

inputs and exits of the neuronal network. This *W* vector gives a value to each exit depending on the present pattern.

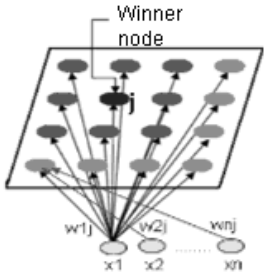

Fig. 7. One layer SOM classifier for hand sign recognition
The Network at time *t* is presented as:

$$x(t) = [x_1(t), x_2(t), ..., x_n(t)]^T \quad (4)$$

The supervised learning procedure consists of the following steps:

1. Initialize weights w_{ij} ($1 \leq i \leq N$) to small random values. Set the initial radius of the neighbourhood around node j to $N_j(t)$.
2. Present input x(t).
3. Calculate distance dj between the input and each node j, given by:

$$d_j = \sum_{i=0}^{N} [x_i(t) - w_{ij}(t)]^2 \quad (5)$$

4. Determine d_j^* which is the minimum value of d_j for all j values.
5. Update weights for node j* and its neighbours within $N_j^*(t)$. The new weights are:

$$wij(t+1) = w_{ij}(t) + \alpha(t)[x_i(t) - w_{ij}(t)] \quad (6)$$

6. Go to step 2 if the stopping condition is not satisfied. Learning rate α(t) is specifies as follows:

$$\alpha(t) = \alpha(1 - t/\beta) \quad (7)$$

In the ordering phase α= 0.99 and β=1000. After the ordering phase, α(t) is given a small value of 0.09 for the following steps. The neighbourhood around node j, Nj(t), is set to half of the map size at the beginning of the training and Nj(t) decreases as time increases. The maximum iteration number is used as the stopping condition. It is 3 to 5 times the number of training patterns. The learning rate α(t) and the stopping condition used in the experiments were tested on several groups of training data and small changes of the values did not significantly affect the results.

The W weight obtained in this neuronal network passes to the LVQ, which allows us to do a final touch of the weights in order to define more clarity the typical features of every sign (Fig.8).

Fig. 8. Output Eigenspaces for each one of the hand sing.

The architecture of the LVQ neuronal network have 1600 data, a *W* vector of 26 x 1600, and 26 output neurons, the same configuration as SOM. With this network what was obtained is a better response of the system since thanks to its competitive learning, only it updated the weight of the winning neuron and depending on it, adjust the weight of the connections (Fig. 9) (Hilera. J. 1995).

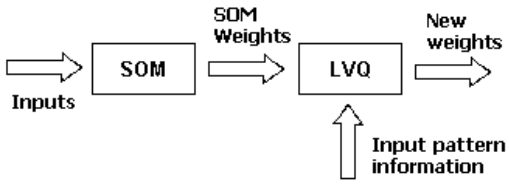

Fig.9. SOM/LVQ model

After the training and once we obtain the better output weights of the network (Fig. 10), a text file is used to save the output weights *W* of each network. In the classification process, each input pattern is led to the layer maps by looking for the several closest nodes and then identifying the class of the nearest prototype in the corresponding maps. To interpret this result, the index of the letter to which the above mentioned argument belongs is saved, considering that for the index 0 corresponds the *a*, for *b* corresponds the index 1 and this way successively. These indexes are saved in a vector that is used in the process of speech synthesis.

Fig. 10. Output images of SOM and LVQ for 1 user.

Once the indexes corresponding to the input information of the neuronal network are obtained, they are sent to the speech synthesizers in order to reproduce in sound the hand signs.

5. SPEECH SYNTHESIZER

The speech synthesizer concatenates pieces of recorded speech, which are stored in a database (Ahuactzin L 2003). This database stored speech units as phones, diphones and triphones to reproduce the input words clearly and similarly to the human voice.

5.1 Speech Synthesizer Algorithm

When the network identifies every sign, each of these signs are converted to a chain of numbers in the same order that they have been introduced by the user. Every letter is identified by a number, for example, the letter '*a*' is identified by the number '*1*', the letter '*b*' is identified by the number '*2*', and this way successively, Then the system accedes to the information of the chain.

The last step of the process has to concatenate the two first data of the chain and compared if these are a valid diphone, but also is needed to compare if the tow first data have the possibility to make a triphone, For example the diphone *'gu'* could be part of a triphone like *'gue'* or *'gui'* and the diphone *'br'* could make the triphone *'bra'*. Like that, the diphonems are validated if could be a part of some of these special cases, when it happens the third character is concatenate. After fulfilling all the verifications, each of the diphones and triphones are reproduced using the Builder C++ media player.

6. EXPERIMENTAL RESULT

The described procedure has been implemented and tested in a 3 GHz, Pentium IV personal computer. A Logitech colour web cam has been used. Signing person may be reasonably placed from the web cam.

The hand detection algorithm was probed with images that only have the hands area. When used alone, the two classifiers give recognition rates of 82,3% and 76% respectively. When the two classifiers are combined a recognition rate of 83% is achieved.

The system validation was made taken samples of 3 different users (Fig. 11). Every sample includes 3 images for every sign. The obtained results are show in the table 1.

Fig. 11. D letter made by the same user at different times

TABLE 1.
Obtained recognition ratios probe the suitability of the proposed approach.

Hand Sign		Response of the System		
Letter	Class	User 1	User 2	User 3
A	Class 1	2	3	2
B	Class 2	3	3	3
C	Class 3	3	2	3
D	Class 4	2	3	3
E	Class 5	2	2	2
F	Class 6	2	3	3
G	Class 7	2	1	1
H	Class 8	1	2	2
I	Class 9	3	3	3
J	Class 10	3	2	3
K	Class 11	2	3	2
L	Class 12	3	2	2
M	Class 13	2	3	2
N	Class 14	2	2	2
O	Class 15	2	2	3
P	Class 16	3	3	3
Q	Class 17	3	3	3
R	Class 18	2	3	2
S	Class 19	1	2	1
T	Class 20	2	2	2
U	Class 21	2	3	3
V	Class 22	3	3	3
W	Class 23	3	2	3
X	Class 24	2	2	3
Y	Class 25	2	2	3
Z	Class 26	2	3	3
Recognition ratios		76%	83%	83%

7. CONCLUSIONS

Based on results of the system used by three different users SOM and LVQ provide an attractive performance in detection and recognition of the 26 hand gestures. This solution combined with the speech synthesizer allows deaf or hard of hearing people to establish an easy communication way with other user that unknown the Sign Language.

Results suggest that the system could help deaf people improve communication, adapt to their environment, and function in society by a common and necessary tool as the computer. This device, which could be found in places like homes, offices, restaurants, etc., makes the system accessible to everybody.

ACKNOWLEDGEMENTE

This work has been partly supported by the Militar University of Colombia with the participation of Laura Pardo and Juan Jose Padilla, the Spanish Profit program (Project Nb. 3-690-2005) ant the Spanish Ministry of Work

and Social Affairs (Imserso R&D program, Project Nb.125/06).

REFERENCES

Ahuactzin L. (2003). *Diccionario español/ingles para el aprendizaje de vocabulario utilizando una interfaz de voz.* Cap 1. http://catarina.udlap.mx/u_dl_a/tales /documentos/lis/ahuactzin_1_a/capitulo1.pdf

Anil K. J. Robert P. W. and Jianchanng M.(2000). *Statical Patter Recongnition: A Review*. IEEE Transaction on pattern Analysis and Machine Intelligence, Vol. 22, No. 1, pp. 4-37.

Extraction of borders. http://wgpi.tsc.uvigo.es/libro/procesim/node8.htm.

García I. Gómez J. Zalama E. (2006). *Hand Gesture Recognition for Deaf People Interfacing*. Proc. Of the 18th International Conference on Pattern Recognition, IEEE Computer Society, Vol. 2, pp.100-104, 08.

Hany F. (2000). *Fundamentals of digital image processing Tutorials.* Pp 43. http://www.cs.darmouth.edu/~fari d/tutorials.

Haykin S. (1994). *Neural networks: a comprehesive foundation, Englewood Cliffs*. New Jersey: IEEE Press Macmillan.

Hilera J.R and V.J Martinez (1995). *Redes Neuronales Artificiales: Fundamentos, modelos y aplicaciones*. Madrid, España. Ed RAMA.

Knerr S. Personnaz L. and Dreyfus G.(1992) . *Handwritten digit recognition by neural networks with single-layer training.* IEEE Trans. Neural Networks, 3(6):962-968, 1992.

Ortiz M. M. (2005). *Operaciones orientadas al punto* Procesamiento digital de imágenes. Notas del curso. Capitulo 2. pp. 2. http://www.cs.buap.mx/~mmartin/pdi/PDI-Cap2.pdf

Ortiz M. M. (2005). *Introducción.* Procesamiento digital de imágenes. Notas del curso. Capitulo 1. pp. 6-8 http://www.cs.buap.mx/~mmartin/pdi/PDI-Cap1.pdf

Speech synthesis: *Definition* http://www.portaldigitro.com.br/_esp/laboratorio_ tecnologias/descritivos/index.php, sf

Weiser M.(1991). *The computer for the 21st century.* Scientific American, 256(3):66-76.

Yuntao C. Daniel. L. Swets and John J. Weng (1995). *Learning Based hand sing recognition using Shoslif-M.* Int'1 Conf. On Computer Vision. pp.631- 636.