STUDIA UNIV. BABEŞ–BOLYAI, INFORMATICA, Volume **LVI**, Number 4, 2011

REAL TIME SIGN LANGUAGE RECOGNITION USING ARTIFICIAL NEURAL NETWORKS

CORNELIU LUNGOCIU

Abstract. This paper focuses on the problem of recognizing in real time the sign language used by the community of deaf people. The problem is addressed using digital image processing methods in combination with a surpervised learning approach to recognize the signs made by a deaf person. For the recognition step, an artificial neural network will be used. The main goal of the paper is to show that a good performance can be achieved without using any special hardware equipment, so that such a system can be implemented and easily used in real life. An experiment is provided and directions to further improve our work are also emphasized.

1. INTRODUCTION

The goal of our work is to make possible the communication between deaf people and the rest of the world in daily life. According to some unofficial statistics [1], the sign language is known and used in daily communication by deaf people, and persons leaving close to them. This is why people with hearing disabilities are unable to communicate with other hearing people without a translator. For this reason, the implementation of a system that recognize the sign language on daily live (public spaces, banks, post offices, schools, etc) or video chat applications, would have a significant benefic impact on deaf people social live.

In this paper we propose a system that is supervised for recognizing one component of the sign language communication: finger spelling in English. For the supervised learning scenario, an *artificial neural network* will be used.

The rest of the paper is structured as folows. The problem statement, its relevance as well as the difficulties that arise in the study of this field are

Received by the editors: 09-11-2011.

2010 *Mathematics Subject Classification.* 68T05, 65K05.

1998 *CR Categories and Descriptors.* I.2.10 [**Vision and Scene Understanding**]: Shape; I.4.8 [**Image Processing and computer vision**]: Shape; I.5.1 [**Pattern recognition**]: Neural nets .

Key words and phrases. Sign Language Recognition, Artificial Neural Networks, Image Processing.

presented in Section 2. Section 3 briefly reviews the fundamentals of artificial neural networks, also presenting existing approaches to the sign language recognition problem. Our supervised learning based approach, as well as details of our implementation are introduced in Section 4. An analysis of the obtained experimental results is also provided. Section 5 presents the advantages and drawbacks of our system, emphasizing a set of possible further improvements.

2. PROBLEM DESCRIPTION AND RELEVANCE

To better understand the problem, we start by defining the signs in the context of gesture communication and sign language.

A sign is a form of non verbal communication done with body parts and used instead of oral communication, or in combination with it. Most people use both words and signs during communication. A sign language is a language that uses signs to communicate instead of sounds, where the signs are the hand shapes, positions and movements of the hands, arms or body, facial expressions or movements of the lips.

According to the above definition, a sign language has three major components [11]:

(1) *Finger-spelling*: for each letter of the alphabet there is a corresponding sign. This type of communication is used mainly for spelling names.
(2) *Word level sign vocabulary*: for each word of the vocabulary there is a corresponding sign in the sign language. This is the most commonly used type of communication between people with hearing disabilities in combination with the third type.
(3) *Non-manual-features*: facial expressions and tongue, mouth and body position.

As defined above, it is obvious the fact that a sign language is not international. Each language has its particularities regarding the alphabet and vocabulary. In different languages there are different words with different meanings, or words for which there is no equivalent word in other language, etc.

Like the spoken languages and dialects currently used, the sign language has developed differently depending on the region and culture. That is why it would be very useful to have an automatic sign language recognition system.

3. BACKGROUND

In this section we briefly review the fundamentals of *artificial neural networks*, and present several existing approaches to the problem considered in this paper.

3.1 Artificial Neural Networks

Artificial neural networks are emerging as the technology of choice for many applications, such as pattern recognition, speech recognition[9], prediction [8], system identification and control. An *artificial neural network* [13] is a system based on the operation of biological neural networks, in other words, is an emulation of biological neural system. An Artificial Neural Network is an adaptive system that learns to perform a function (an input/output map) from data. Adaptive means that the system parameters are changed during operation, normally called the *training phase*. After the training phase the Artificial Neural Network parameters are fixed and the system is deployed to solve the problem at hand (the *testing phase*).

In a supervised learning scenario, an input is presented to the neural network and a corresponding desired or target response set at the output. These input-output pairs are often provided by an external teacher (supervisor). An error is composed from the difference between the desired response and the system output. This error information is fed back to the system and adjusts the system parameters in a systematic fashion (the learning rule). The process is repeated until the performance is acceptable.

3.2 Sign Language Recognition Systems

The problem of recognizing the sign language in real time was intensively studied in the past in prestigious universities like MIT, University of Milan and the Royal Melbourne Institute of Technology, as well as in private companies such as Fujitsu.

To be able to recognize the signs, a set of measurable features of the body that make difference between signs is needed. The body characteristics that make the difference between the signs are the shape of the hand, the angle from each joint of the fingers and wrist, or arm position and trajectory.

To implement such a system, researches have been conducted in two main directions [10]:

(1) systems that use specialized hardware devices for data acquisition: robotic glove to measure finger and hand joint angles, and various mechanical, optical, magnetic and acoustic devices to detect hand position and trajectory;

(2) systems that use image processing and computer vision techniques to detect the characteristics of the hand in images taken with a video or web camera.

If using imaging systems for detecting body characteristics, data preprocessing is required to obtain numerical features describing the characteristics of the body. This preprocessing step is quite difficult, so the best results were obtained using specialized hardware for data acquisition.

Several existing sign language recognition systems are:

(1) *SLARTI* [10]: for data acquisition it uses a robotic glove and a system based on magnetic fields. For classifying data, a set of neural networks is used, each one trained to classify the sign according to a set of features.
(2) *Glove-Talk* [14]: for data acquisition it uses a robotic glove and for classification uses multiple artificial neural networks.
(3) *Talking Glove* [6]: for data acquisition uses a robotic glove and for classification uses a neural network. This system achieved only finger spelling recognition.
(4) *University of Central Florida gesture recognition system* [3]: uses a webcam and computer vision techniques to collect the data and a neural network to classify shapes. For recognition, this system requires wearing a specially colored glove to facilitate the imaging process.
(5) *ASLR* [12]: uses a webcam and computer vision techniques to collect field data and for classification of signs uses Bayesian learning [15].

The results obtained by each of the above described systems are presented in Table 1.

System	Vocabulary size	Accuracy
SLARTI	52	94%
Glove-talk	203	94%
Talking Glove	28	-
PlaceTypeCentral Florida	8	94%
ASLR	201	17%

Table 1. Comparative results.

Still a comparison between all these systems can not be made, because not all of them have the same purpose, and there is no benchmark database to test all the applications. Each tries to recognize a particular component of sign language. Some intend only to recognize the finger-spelling; others recognize a vocabulary of signs (SLARTI, Glove-talk) or take into account other features of the body such as head position (ASLR).

4. OUR APPROACH

The purpose of this paper is to implement a system that recognizes a component of a sign language, namely: finger-spelling in English. We also aim to develop a simply to use project, that is why, in addition to other implementations, we are trying to achieve this without using any specialized hardware, and without requiring the user to use gloves or other clothing of certain colors.

To make this possible, the project will use a web camera which takes images of the signs made by the user. The images are processed to extract the characteristics necessary for recognition, which are then used as inputs for an artificial neural network that will recognize the sign. The project steps are described in Figure 1.

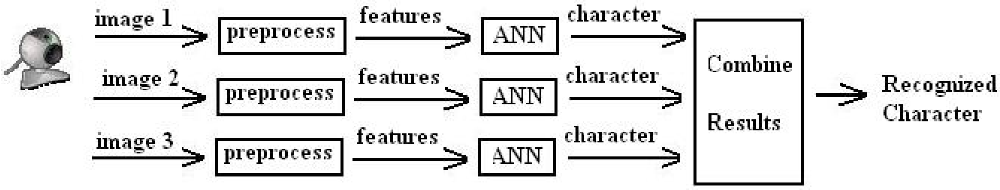


Figure 1. The recognition steps.

To overcome the errors that may occur because of noise from the images, we are using three consecutive images taken from the webcam to recognize a single character. Each image is individually processed and recognized, and then the results are combined using a heuristic method to obtain the final result.

As shown in Figure 1, after taking the image from web camera, three steps are required for recognition of one sign:

(1) *Preprocessing*: uses digital image processing techniques to extract important features from the image, that will be used for recognition.
(2) *Sign recognition*: the data obtained at the first step are used as inputs for a neural network that recognizes the sign.
(3) *Results combination*: the results obtained from these three images are then combined using a heuristic to produce the final result.

The first two steps are done for each of the three images separately. Then the last step is done only one time for all three images.

In the following subsections we will detail each phase of the proposed recognition system.

4.1 Preprocessing

The purpose of this step is to obtain, after the initial image processing, a set of numerical features that uniquely describe a sign. Depending on the goal and available devices, these values may relate to each finger joint angle, hand position and orientation, body position, facial expressions, etc.

Considering the purpose of this work, namely to recognize the fingerspelling in English, detecting the hand shape from the image is sufficient for the recognition of a sign. For this reason, the problem becomes a matter of recognizing objects in an image based on their shape.

The preprocessing is done according to the following diagram:

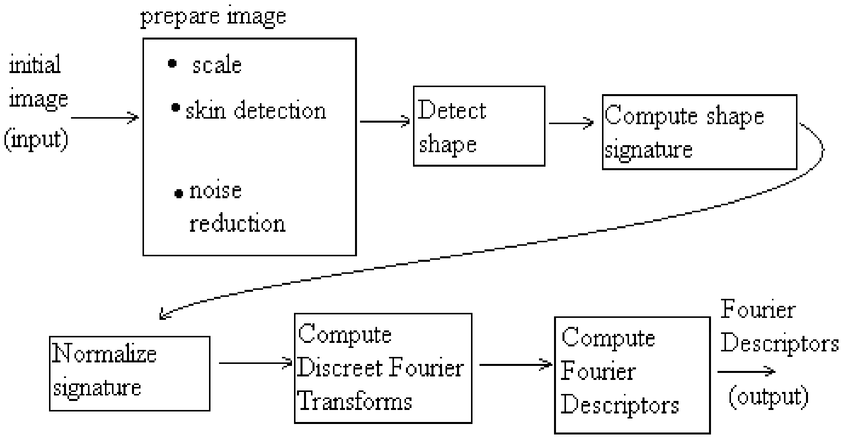


Figure 2. The preprocessing steps.

Image preparation. For this step we propose the following substeps:

(1) *Image scaling* is done to reduce the computational effort needed for image processing.

(2) *Skin detection.* The result of this processing is a binary image in which those pixels that define the hand are colored with white and all the others are black. This processing involves classification of each pixel of the image as part of a human skin or not. There are several techniques developed for skin detection in images, as described in [16]. In this paper, pixels are categorized based on an explicit relationship between the color components red, green and blue. Thus, a pixel is categorized as belonging to the skin if the color components (R, G, and B) meet the following relation:

$R > 95$ and $G > 40$ and $B > 20$ and $\max\{R, G, B\} - \min\{R, G, B\} > 15$ and $|R - G| > 15$ and $R > G$ and $R > B$

(3) *Noise reduction.* After the skin detection, not all pixels are correctly categorized. Noise reduction is meant to overcome these errors, coloring the pixels according to the colors of neighboring pixels [7].

Shape detection. After image processing, the outline of a hand is detected, as a vector of (x, y) coordinates, being the coordinates of the pixels that define the contour of the hand. For this we have used an adaptation of the algorithm described in [17].

Shape signature. Shape signature is a one dimensional vector characterizing the outline of a two-dimensional shape. Four types of shape signatures are described in [4]:

(1) *Complex coordinates*: the vector components are complex numbers. For each pixel of coordinates (x, y) that defines the outline of the shape, the complex number c = x + i*y is saved in the vector.
(2) *Centroid distance*: first, the centroid of the shape is computed, which is defined as the mean of all coordinates of the pixels from the outline. For each pixel from the outline, the distance from its position to the shape centroid is saved in the vector.
(3) *Curvature signature*: for each pixel that defines the outline of the shape the second derivative of the outline is saved.
(4) *Cumulative angular function*: the elements of the vector are the values of the angels between consecutive pixels from the outline.

Each of these variants were analyzed according to the rotation, translation and scaling invariance, and the best results were obtained using the centroid distance, which provides translation and scaling invariance. For this reason, this form signature was used in this paper.

Shape signature is normalized to reduce the size of the signature to a fixed number of values. This is done to reduce the size of the data, thus reducing the computational effort, to allow the uniform treatment of various forms of signatures, and to remove insignificant details form the outline of the shape. There are several ways of normalizing the signature [4]. Here we have used the "*equal point sampling*" method. The equal points sampling method selects candidate points spaced at equal number of points along the shape boundary.

Fourier Descriptors. For a normalized shape signature obtained as described in the previous paragraph, *s(t)*, $t = 0, 1, \dots$ *N*, the Discreet Fourier Transformations are calculated as indicated in Formula (1).

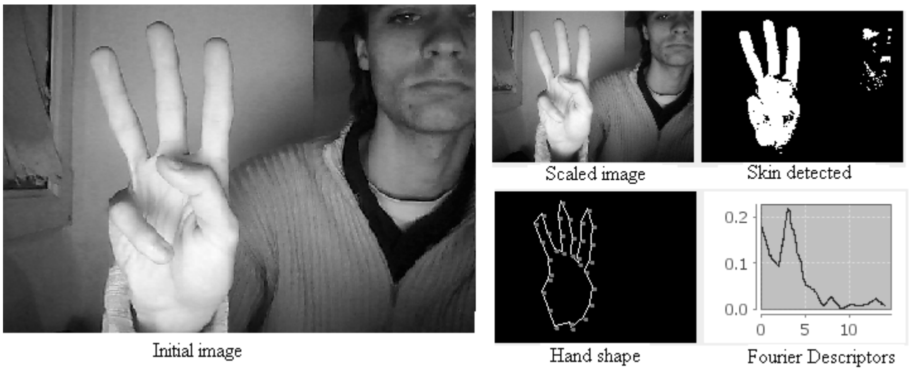
$$FD_n = \frac{1}{N} \sum_{t=0}^{N-1} s(t) \cdot e^{\frac{-j \cdot 2 \cdot \pi \cdot n \cdot t}{N}}, n = 0,1,\dots,N-1 \qquad (1)$$

The FD coefficients are called Fourier descriptors of the form. The calculation of these descriptors given by the above formula requires a large computational effort. For this implementation we have used the FFD algorithm (Fast Fourier Descriptors), which can reduce the complexity of the calculations if the input vector dimension is power of 2. This is why we chose N = 32.

Fourier transformation concentrates the information from the shape signature in the first terms of the result vector. To normalize the values of descriptors, each term is divided by the first term (called the DC component, which depends only on the position form the image). Since the signature form contains real values, only N / 2 distinct values will be obtained. For this reason, the Fourier descriptors are finally obtained as follows:

$$f = \left[\frac{|FD_1|}{|FD_0|}, \frac{|FD_2|}{|FD_0|}, \dots, \frac{|FD_{(N-1)/2}|}{|FD_0|} \right] \qquad (2)$$

Figure 3 illustrates the preprocessing steps for a sample image.



Initial image

Scaled image

Skin detected

Hand shape

Fourier Descriptors

Figure 3. The preprocessing steps

4.2 The Neural Network Model

The Fourier descriptors obtained after the preprocessing step (Subsection 4.1) will be used as input for a neural network that will recognize the shape based on these features. The response of the network is a vector V, with x$v_i \in [0,1]$, $i = \overline{1,24}$, representing the probability for the sign made by the user to be the ith letter from the English alphabet. For one sign made by the user there are three vectors of this form obtained (see Section 4): V^1,V^2,and V3.Combining these vectors we obtain the final result, in the form of a vector V^f, with $v_i^f = v_i^1 + v_i^2 + v_i^3$.

The neural network used is a feed-forward network [10] with layered architecture. It has one input layer, one output layer and one hidden layer, with each layer fully connected with the following layer. The number of the neurons from the hidden layer is the mean between the number of neurons from the input layer and the number of neurons from the output layer. The aggregation function for each neuron is the weighted sum of its inputs and the transfer function is the sigmoid function [15][2].

The network is trained in a supervised learning scenario. The most commonly used learning algorithm for neural networks with layered architectures is the *backpropagation algorithm* [15]. In the current implementation we have used an adaptation of this algorithm, known as *stochastic backpropagation*, which updates the weights after each backward propagation of the error for one training element, not only once for all training set. We have also used the *momentum technique* [15] to avoid local minima in the solution space, and we decreased the *learning rate* during the learning process. The following values for the parameters were chosen in our implementation: 0.3 for learning rate that linearly decreases to 0.01, and 0.1 for the momentum.

For training a set of 84 elements was used, representing the following letters of the English alphabet: A, B, C, D, E, F, I, K, L, R, U, V, W, X. For each letter there are 6 training instances. To select the training elements, we use the *cross validation* method, thus the data set is randomly partitioned, obtaining a *training* set, and a *validation* set. To avoid the *overfitting* of the network on the training set, every 1000 iterations the initial set is repartitioned. The learning process stops when the network error computed on the validation set, becomes less than a certain value.

4.3 Experimental results

The system was trained to recognize a set of 14 letters from the English alphabet: A, B, C, D, E, F, I, K, L, R, U, V, W, X, and the current results are promising, showing the fact that an approach like this, with a few improvements, could produce good results. A statistic regarding the performance of this system can be found below:

Recognized characters	Recognition accuracy
A, B, C, D, E, F, I, K, L, R, U, V, W, X	80 %

Our approach could be significantly improved if a skin detection algorithm to classify pixels according to the image histogram would be used. This would increase the robustness of the algorithm at the light intensity and color. Also, taking into consideration both the inner and outer contour of shapes would make a better differentiation between the signs, and therefore make the learning process easier.

5. CONCLUSIONS AND FUTURE WORK

We have proposed in this paper a neural network based approach for the sign language recognition. As shown by the experimental results, the solution we have proposed in this paper can be efficiently used in real life situations.

The main advantage of our approach over the other attempts is the fact that it requires no additional hardware equipment or special clothes to recognize the signs. As it happens in most applications that use computer vision techniques to collect data, the main drawback of our solution is the image processing part. The image processing phase of sign recognition process requires a large amount of calculations, which introduces latency in the video stream. The accuracy of the system could be also increased if a more robust skin detection algorithm will be used.

The approach proposed in this paper can be easily extended to recognize the hand trajectory in the image, by this being able to recognize the word level sign vocabulary. Future work may also be done in order to use another classification model in the supervised learning scenario, such as *support vector machines* [5] or *Bayesian Learning* [15].

References

[1] A.N.S.R, *Manual de Limbaj Mimico-Gestual Românesc*, Editura Mega, 2008.

[2] Dave Andersoan, George McNeill, *Artificial Neural Networks Technology*, Rome Laboratory, 1992

[3] Davis, J and Shas M., *Gesture Recognition,* Technical Report CS-TR-93-11,University of Central Florida, 1993

[4] Dengsheng Zhang, Guojun Lu - *Content-Based Shape Retrieval Using Different Shape Descriptors: A Comparative Study*, ICME, IEEE International Conference on Multimedia and Expo (ICME'01), 2001, pp. 1139–1142.

[5] I. Steinwart, A. Christmann, *Support Vector Machines*, Springer Publishing Company, Incorporated, 2008.

[6] J Kramer and L Leifer. *The Talking Glove: A Speaking Aid for Nonvocal Deaf and Deaf-Blind Individuals*, Proc. of the RESNA 12th annual Conf. (1993) pp. 471-472.

[7] Jerry Huxtable, Java Image Processing Pages: http://www.jhlabs.com/ip/filters/index.html

[8] K. R. Linstrom and A.J. Boye. *A neural network prediction model for a psychiatric application*, International Conference on Computational Intelligence and Multimedia Applications, pp. 36-40, 2005.

[9] M. D. Skowronski and J.G. Harris, *Automatic speech recognition using a predictive echo state network classifier*, Neural Networks, Volume 20, Issue 3, pp:414-423, April 2007.

[10] Peter Wray Vamplew , PhD Thesis: *Recognition of Sign Language Using Neural Networks*, Flinders University of South Australia, 1990

[11] Philippe Drew, Research on Sign Language Recognition: http://www-i6.informatik.rwth-aachen.de/~dreuw/database.php

[12] Philippe Drew, David Rybach, Thomas Deselaers, Morteza Zahedi, Herman Ney, *Speech recognition techniques for sign language recognition system,* In Interspeech, pages 2513-2516, Antwerp, Belgium, August 2007.

[13] R. Rojas, *Neural Networks: A Systematic Introduction*, Springer, 1996.

[14] S. Sidney Fels and Geo_rey E. Hinton, *Glove-TalkII: A neural network interface which maps gestures to parallel formant speech synthesizer controls*, Transactions on Neural Networks, 9 (1), 205–212. 1998.

[15] Tom M. Mitchell, *Machine Learning*, McGraw-Hill, 1997

[16] Vladimir Vezhnevets, Vassili Sazonov, Alla Andreeva, *A Survey on Pixel-Based Skin Color Detection Techniques*, Graphics and Media Laboratory, Faculty of Computational Mathematics and Cybernetics, Moscow State University, 2002.

[17] Wilhelm Burger, Mark J. Burge, *Digital Image Processing - An Algorithmic Introduction using Java*, Springer-Verlag Berlin, Heidelberg, New York, 2008.

Babe-Bolyai University, Faculty of Mathematics and Computer Science, 1, M. Kogalniceanu Street, 400084, Cluj-Napoca, Romania

E-mail address: `lcsi0417@scs.ubbcluj.ro`