Journal of the ACS, Vol. 8, May 2014

Arabic Sign Language Recognition Using Neural network

Abdelatif Hussein A.ALI,

High Institute of Computers and Information Technology, Computer Dept., El-Shorouk Academy, Cairo, Egypt,
E-mail: aabouali@hotmail.com

Ahmed Fouad Kamal
Maryam Saeed Mostafa
Dalia Atef Hassan
Modern Academy for Engineering, Cairo, Egypt

Abstract

Sign language maps letters, words, and expressions of a certain language to a set of hand gestures enabling an individual to communicate by using hands and gestures rather than by speaking. Systems capable of recognizing sign-language symbols can be used as a means of communication between Hearing-impaired and vocal people. This paper represents an attempt to recognize handed signs from the Unified Arabic Sign Language Dictionary using a webcam and artificial neural networks. Hu moments is used for feature extraction. 200 samples of each of 5 two-handed signs were collected from an adult signer. 150 samples of each sign were used for training an artificial neural networks to perform the recognition. The performance is obtained by testing the trained system on the remaining 50 samples of each sign. A recognition rate of 87.6% on the testing data was obtained. When more signs will be considered, the artificial neural networks must be retrained so that signs are recognized and categorized.

I. Introduction

Signing has always been part of human communications [1]. Newborns use gestures as a primary means of communication until their speech muscles are mature enough to articulate meaningful speech. For thousands of years, deaf people have created and used signs among themselves. These signs were the only form of communication available for many deaf people. Within the variety of cultures of deaf people all over the world, signing evolved to form complete and sophisticated languages. Sign language is a form of manual communication and is one of the most natural ways of communication for most people in deaf community. There has been a re-surging interest in recognizing human hand gestures. The aim of the sign

language recognition is to provide an accurate and convenient mechanism to transcribe sign gestures into meaningful text or speech so that communication between deaf and hearing society can easily be made. Hand gestures are spatio-temporally varying and hence the automatic gesture recognition turns out to be very challenging [2-3]. As in oral language, sign language is not universal; it varies according to the country, or even according to the regions. The significance of using hand gestures for communication becomes clearer when sign language is considered. Sign language is a collection of gestures, movements, postures, and facial expressions corresponding to letters and words in natural languages, so the sign language has more than one form because of its dependence on natural languages. The sign language is the fundamental communication method between people who suffer from hearing impairments. In order for an ordinary person to communicate with deaf people, an interpreter is usually needed to translate sign language into natural language and vice versa [4]. Human-Computer Interaction (HCI) is getting increasingly important as a result of the increasing significance of computer's influence on our lives [4]. Researchers are trying to make HCI faster, easier, and more natural. To achieve this, Human-to-Human Interaction techniques are being introduced into the field of Human-Computer Interaction. One of the richest Human-to-Human Interaction fields is the use of hand gestures in order to express ideas.

II. Related Work

Signing has always been part of human communications. The use of gestures or sign is not tied to ethnicity, age, or gender [5]. In recent years, several research projects in developing sign language systems have been presented [6]. In [5], an Arabic Sign Language Translation Systems (ArSL-TS) model has been introduced. That presented model runs on mobile devices to develop an avatar based sign language translation system that allows users to translate Arabic text into Arabic Sign Language for the deaf on mobile devices such as Personal Digital Assistants (PDAs). In [7], a virtual signer technology was described. The ITC (Independent Television Commission-UK) has specially made Televirtual to develop "Simon", the virtual signer in order to translate printed text - television captions - into sign language. The proposed model tried to solve some of the problems

resulted from adding sign language to television programs. Also, authors discussed the language processing techniques and models that have been investigated for information communication in a transaction application in Post Offices, and for presentation of more general textual material in texts such as subtitles accompanying television programs. The software proposed in [7] consists of two basic modules: linguistic translation from printed English into sign language, and virtual human animation. The animation software allows Simon to sign in real-time. A dictionary of signed words enables the system to look up the accompanying physical movement, facial expressions and body positions, which are stored as motion-capture date on a hard disk. The motion-capture data that includes hand, face and body information is applied to a highly detailed 3D graphic model of a virtual human. This model includes very realistic and accurate hand representations, developed within the project. Moreover, natural skin textures are applied to the hands and face of the model to create the maximum impression of subjective reality. In [8], Data acquisition, feature extraction and classification methods employed for the analysis of sign language gestures have been examined. These were discussed with respect to issues such as modeling transitions between signs in continuous signing, modeling inflectional processes, signer independence, and adaptation. Also, it has been stated that non-manual signals and grammatical processes, which result in systematic variations in sign appearance, are integral aspects of this communication but have received comparatively little attention in the literature. Works that attempt to analyze non-manual signals have been examined. Furthermore, issues related to integrating these signals with (hand) sign gestures and the overall progress toward a true test of sign recognition systems dealing with natural signing by native signers have been discussed. Moreover, a summary of selected sign gesture recognition systems using sign-level classification has been presented in [8]. According to that summary, the two main approaches in sign gesture classification either employ a single classification stage, or represent the gesture as consisting of simultaneous components that are individually classified and then integrated together for sign-level classification. Another summary indicated the variety of classification schemes and features used under the two broad approaches. In each approach, methods that use both, direct-measure devices and vision are included. In [9], an automatic Thai finger-spelling sign language translation system was developed using Fuzzy C-

Means (FCM) and Scale Invariant Feature Transform (SIFT) algorithms. Key frames were collected from several subjects at different times of day and for several days. Also, testing Thai fingerspelling words video was collected from 4 subjects. The system achieves 79.90% and 51.17% correct alphabet translation and the correct word translation, respectively, with the SIFT threshold of 0.7 and 1 nearest neighbor prototype. However, when the number of nearest neighbor prototypes was increased to 3, the system yields higher percentages, 82.19% and 55.08% correct alphabet and correct word translation, respectively, at the same SIFT threshold. Also, a system for automatic translation of static gestures of alphabets in American Sign Language (ASL) was developed in [10]. Three feature extraction methods and neural network were used to recognize signs. The developed system deals with images of bare hands, which allows the user to interact with the system in a natural way. An image is processed and converted to a feature vector that will be compared with the feature vectors of a training set of signs. The system is implemented and tested using data sets of number of samples of hand images for each signs. Three feature extraction methods are tested and best one is suggested with results obtained from Artificial Neural Network (ANN). The system is able to recognize selected ASL signs with the accuracy of 92.33%. In [11], Authors discussed the development of a data-driven approach for an automatic machine translation (MT) system in order to translate spoken language text into signed languages (SLs). They aimed at improving the accessibility to airport information announcements for deaf and hard of hearing people. [11] also demonstrates the involvement of deaf members of the deaf community in Ireland in three areas, which are: the choice of a domain for automatic translation that has a practical use for the deaf community; the human translation of English text into Irish Sign Language (ISL) as well as advice on ISL grammar and linguistics; and the importance of native ISL signers as manual evaluators of our translated output. The proposed system achieved a reasonable job of translating English into ISL with scores comparable to mainstream speech-to-speech systems. More than two thirds of the words produced are correct and almost 60% of the time the word order is also correct.

A comprehensive study for Arabic Sign Language, ASL, based on pulse coupled neural network in [24-27]. In this study snapshots from two video cameras at different angles from the signer are used to provide images of resolution 160 X 120X24. The features extracted using the pulse coupled

neural network, which are invariant to translation, rotation, and scaling, from the two sources are combined using weighting that depend on signature quality of each. The study also includes the use of natural language processing to aid the recognition process.

III. Pre-Processing

1. **Binary skin classifiers:**
   The methods considered in this paper separate skin and non-skin colors using a piecewise linear decision boundary. These explicit skin cluster methods propose a set of fixed skin thresholds in a given color space. Some color spaces permit searching skin color pixels in the 2D chromatic space, reducing dependence on lighting variation, others, such as the RGB space, address the lighting problem by introducing different rules depending on illumination conditions (uniform daylight, or flash). Working within different color spaces, we have implemented the two different algorithms analyzed in this paper. They are named for the color space adopted: YCbCr [12] and HSV [13]. The details of their implementation can be found in the referenced papers and are summarized in the subsections here below.

   a. **YCbCr:**
      Chai and Ngan [12] develop an algorithm that exploits the spatial distribution characteristics of human skin color. A skin color map is derived and used on the chrominance components of the input image to detect pixels that appear to be skin. The algorithm then employs a set of regularization processes to reinforce those regions of skin-color pixels that are more likely to belong to the facial regions. We use only their color segmentation step here. Working in the YCbCr space the authors find that the ranges of Cb and Cr most representative for the skin-color reference map were:

      $$77 \leq Cb \leq 127$$
      $$133 \leq Cr \leq 173.$$

   b. **HSV:**
      Starting from a training data set composed of skin color samples,

Garcia and Tiziritas [13] compute the color histogram in hue-saturation-value (HSV) color space, and estimate the shape of this skin color subspace. They find a set of planes by successive adjustments depending on segmentation results, recording the equations shown below which define the six bounding planes found in the HSV color space case, where H ∈[−180°180°]:

$V \geq 40$
$H \leq (-0.4V + 75)$
$10 \leq S \leq (-H - 0.1V + 110)$
If $H \geq 0$
$S \leq (0.08\,(100 - V)\,H + 0.5V)$
if $H < 0$ $S \leq (0.5\,H + 35)$.

2. Background Subtraction:

The CB algorithm adopts a quantization technique [14], to construct a background model. Samples at each pixel are clustered into the set of codewords. The background is encoded on a pixel by pixel basis

Construction of the initial codebook [15]:
The algorithm is described for color imagery, but it can also be used for gray-scale imagery with minor modifications. Let X be a training sequence for a single pixel consisting of N RGB-vectors: $X = \{x_1; x_2; \ldots; x_N\}$: Let $C = \{c_1; c_2; \ldots; c_L\}$ represent the codebook for the pixel consisting of L codewords. Each pixel has a different codebook size based on its sample variation. Each codeword c_i : i = 1 . . . L; consists of an RGB vector $v_i = \{\bar{R}_i, \bar{G}_i, \bar{B}_i\}$ and a 6-tuple $aux_i = (I_{i(min)}, I_{i(max)}, F_i, \lambda_i, P_i, Q_i)$: The tuple aux_i contains intensity (brightness) values and temporal variables described below:

$I_{(min)}$, $I_{(max)}$: the min and max brightness, respectively, of all pixels assigned to this codeword

F: the frequency with which the codeword occurred

λ: the maximum negative run-length (MNRL) defined as the longest interval during the training period that the codeword has NOT recurred

P, Q: the first and last access times, respectively, that the codeword has occurred

IV. Feature Extraction Of The Hand Gesture

Image classification is a very mature field today. There are many approaches to finding matches between images or image segments. Starting from the basic correlation approach to the scale-space technique, they offer a variety of feature extraction methods with varying success. However, it is very critical in hand gesture recognition that the feature extraction is fast and captures the essence of a gesture in unique small data set. Neither the Fourier descriptor, which results in a large set of values for a given image, nor scale space succeed in this context. The proposed approach of using moment invariants stems from our success in developing the "Wave Controller". Gesture variations caused by rotation, scaling and translation can be circumvented by using a set of features, such as moment invariants, that are invariant to these operations. The moment invariants algorithm has been recognized as one of the most effective methods to extract descriptive feature for object recognition applications and has been widely applied in classification of subjects such as aircrafts, ships, and ground targets, [19, 20]. Let $f(i,j)$ be a point of a digital image of size $M \times N\ (I = 1,2,\dots,M\ and\ j = 1,2,\dots,N)$. The two dimensional moments and central moments of order $O(p+q)\ of\ f(i,j)$, are defined as:

$$m_{pq} = \sum_{i=1}^{M} \sum_{j=1}^{N} i^p j^q f(i,j)$$

$$U_{pq} = \sum_{i=1}^{M} \sum_{j=1}^{N} (i-\bar{\imath})^p (j-\bar{\jmath})^q f(i,j)$$

Where

$$\bar{\imath} = \frac{m_{10}}{m_{00}} \quad and\ \bar{\jmath} = \frac{m_{01}}{m_{00}}$$

From the second order and third order moments, a set of seven (7) moment invariants are derived as follows [19-20]:

$$\phi_1 = \eta_{20} + \eta_{02}$$

$$\phi_2 = (\eta_{20} - \eta_{02})^2 + 4\eta_{11}^2$$

$$\phi_3 = (\eta_{30} - 3\eta_{12})^2 + (3\eta_{21} - \eta_{03})^2$$

$$\phi_4 = (\eta_{30} + \eta_{12})^2 + (\eta_{21} + \eta_{03})^2$$

$$\varphi_5 = (\eta_{30} - 3\eta_{12})(\eta_{30} + \eta_{12})\left[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2\right] + (3\eta_{21} - \eta_{03})(\eta_{21} + \eta_{03})\left[3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2\right]$$

$$\varphi_6 = (\eta_{20} - \eta_{02})\left[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2\right] + 4\eta_{11}(\eta_{30} + \eta_{12})(\eta_{21} + \eta_{03})$$

$$\varphi_7 = (3\eta_{21} - \eta_{03})(\eta_{30} + \eta_{12})\left[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2\right] - (\eta_{30} - 3\eta_{12})(\eta_{21} + \eta_{03})\left[3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2\right]$$

Where η_{pq} is the normalized central moments defined by:
$r = [(p + q)/2] + 1$
$p + q = 2, 3,...$
It is possible that using few features than seven values will still provide a very useful set of features.

V. Neural Network Classifcation

The Artificial Neural Network or ANN algorithms are the commonly used as base classifiers in classification problems [21]. An artificial neural network is a powerful data modeling and information-processing paradigm that is able to capture and represent complex input/output relationships [22]. The advantage of neural networks mainly lies in that they are data driven self-adaptive methods, which can adjust themselves to the data without any explicit specification of functional or distributional form for the underlying model. Also, they are universal functional approximations in that neural networks can approximate any function with arbitrary accuracy [22], [23]. The function of the neural network is transforming inputs into meaningful outputs. It inspired by the way of biological nervous systems, neural networks look like the human brain in two stages learning stage and testing

stage. Moreover, neural network is able to represent both linear and non-linear relationships and the way it can learn these relationships directly from the modeled data. The most common neural network model is the multilayer perceptron (MLP). This type called supervised network because it requires a desired output in order to learn. The goal of this type of network is to create a model that correctly maps the feature vector of a single sample (input) to the class of the input sample (output) using historical data so that the model can then be used to produce the output when the desired output is unknown.

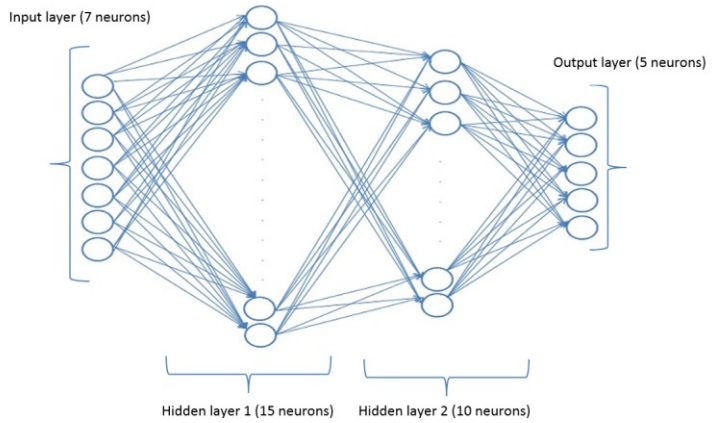

Figure 1 Neural Network Architecture

The neural network consists of three types of layers input, hidden and output layer. Each of them consists of number of perceptrons or neurons and these neurons connected together from layer according to specific network architecture. Each connection has a very important unit called "weight". The weight unit controls the degree of intelligence of the neural network. The input layer is the layer that represents the input data so the length of this layer is equal to the length of the input data (feature vector), and there is only one input layer in the neural network. It consists of a set of input values (X_i) and associated weights (W_i). The hidden layer is the kernel of the network because it controls the number of thinking equations and by which the result gets better. The neural network may contain several hidden layers. The last one is the output layer, which returns the result. The length of this layer equals to the number of classes. There is only one output layer in the neural network. The MLP neural network looks like any neural network so it goes through two stages. The first one is the learning stage, which trains the network to be able to think and return the best result. The learning process comes by updating the weights.

VI. Experimental Results:

The proposed system was implemented on a Core i5 (1.7GHz) laptop computer with Microsoft Windows 8 platform using MathWorks MATLAB R2012a, OpenCV-2.4.7 and Microsoft Visual Studio 2013. To evaluate the performance of the proposed system, we ran a Batch test which consisted of 50 samples for the testing stage paired with their targets, we made the network go through recognition on these samples and compared the network results with the actual targets the results as shown in the graph the network successfully recognized 219 samples out of 250 samples giving a recognition rate of 87.6%.

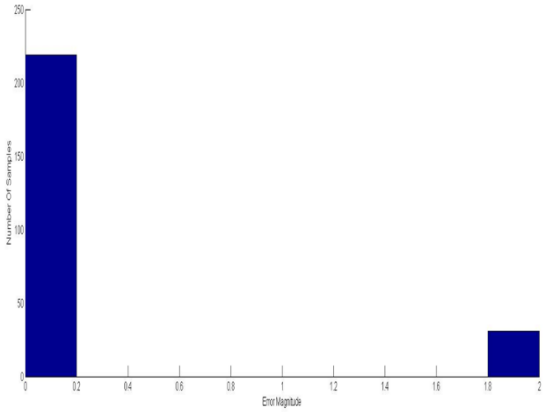

Figure 2 Batch Testing Results

VII. Conclusions and Future Work

In this paper, a simple system for the purpose of the recognition and translation of the alphabets in the Arabic sign language were designed. The proposed Arabic Sign Language Alphabets Translator system composed of five main phases; Pre-processing phase, Best-frame Detection phase, Category Detection phase, Feature Extraction phase, and finally Classification phase. The extracted features are translation, scale, and rotation invariant, which make the system more flexible. Experiments revealed that the proposed system was able to recognize representing subsets (5 letters) of the Arabic manual alphabets with an accuracy of 85.7% using multilayer perceptron (MLP). There still a lot of room for further

research in performance improvement considering different feature sets, considering natural language processing as remedy for recognition, cameras with different views as in [25-27] and classifiers. Moreover, additional improvements can be applied for this system to be used on mobile to provide easy communication way among deaf/hearing-impaired people. Also, this system could be developed to be provided as a web service used in the field of conferences and meetings attended by deaf people. Furthermore, this system could be used by deaf and normal people for controlling their computers and performing actions to them without the need for touching any device. Finally, it can be used in intelligent classrooms and intelligent environments for real-time translation for sign language.

References

[1] K. Assaleh, and M. Al-Rousan. "Recognition of Arabic Sign Language Alphabet Using Polynomial Classifiers", EURASIP Journal on Applied Signal Processing (JASP), 2005(13), pp. 2136-2145, 2005.

[2] M. AL-Rousan, K. Assaleh, and A. Tala'a. "Video-based Signer-independent Arabic Sign Language Recognition Using Hidden Markov Models", Applied Soft Computing, 9(3), pp. 990-999, 2009.

[3] Y. Wu and T. S. Huang. "Vision-based Gesture Recognition: A Review", In Proceedings of 3rd International Gesture Workshop (GW'99), pp. 103-115, France, March 1999.

[4] O. Al-Jarrah and A. Halawani. "Recognition of Gestures in Arabic Sign Language Using Neuro-Fuzzy Systems", Artificial Intelligence, 133(1-2), pp. 117-138, 2001.

[5] S. M. Halawani. "Arabic Sign Language Translation System on Mobile Devices", International Journal of Computer Science and Network Security (IJCSNS), 8(1), pp. 251-256, 2008

[6] M. Huenerfauth. "Generating American Sign Language Classifier Predicates For English-To ASL Machine Translation", Ph.D dissertation, University of Pennsylvania, Department of Computer and Information Science, Philadelphia, PA, USA, 2006.

[7] J.A. Bangham, S.J. Cox, M. Lincoln, ITutt, and M. Wells. "Signing for the Deaf Using Virtual Humans", IEE Seminar on Speech and Language Processing for Disabled and Elderly People 2000/025, London, UK, pp. 4/1-4/5, April, 2000.

[8] S. C.W. Ong and Surendra Ranganath. "AutomatiLanguage Analysis: A Survey and the Future beyond Lexical Meaning", IEEE Transactions on Pattern Analysis and Machine Intelligence, 27(6), June, 2005.

[9] S. Phitakwinai, S. Auephanwiriyakul, and N. Theera-Umpon. "Thai Sign Language Translation Using Fuzzy C-Means and Scale Invariant Feature Transform", In Proceedings of International Conference of Computational Science and Its Applications 2008, Lecture Notes in Computer Science, vol. 5073/2008, pp. 1107-1119, 2008.

[10] V. S. Kulkarni and S.D. Lokhande. "Appearance Based Recognition of American Sign Language Using Gesture Segmentation", International Journal on Computer Science and Engineering (IJCSE), 2(3), pp. 560 2010.

[11] S. Morrissey and A. Way. "Joining Hands: Developing a Sign Language Machine Translation System with and for the Deaf Community", In Proceedings of Conference and Workshop on Assistive Technologies for People with Vision and Hearing Impairments: Assistive Technology for All Ages (CVHI-2007), Granada, Spain, pp. 28 August, 2007.

[12] D. Chai and K. N. Ngan. Face segmentation using skin colour map in videophone applications, IEEE Transactions on Circuits and Systems for Video Technology pp. 551-564 April, 1999.

[13] C.Garcia and G. Tziritas. "Face detection using quantized skin colour regions merging and wavelet packet analysis", IEEE Transaction on Multimedia pp. 264-277 January, 1999.

[14] T. Kohonen. "Learning vector quantization", Neural Networks, Vol. 1, pp. 3-16, 1988.

[15] K. Kim , T.H. Chalidabhongse , D. Harwood and L. Davis, "Background Modeling and Subtraction by Codebook

Construction", Proc. Intâl Conf. Image Processing (ICIP), vol. 5, pp.3061 -3064, 2004.

[19] Q. Zhongliang, and W. Wenjun, "Automatic ship classification by superstructure moment invariants and two-stage classifier", ICCS/ISITA '92 Communications on the Move, pp. 544-547, 1992.

[20] P. Premaratne, "ISAR ship classification: An alternative approach", CSSIP-DSTO Internal Publication, Australia, March, 2003.

[21] F. Roli, G. Giacinto, and G. Vernazza. "Methods for Designing Multiple Classifier Systems", In of 2nd International Workshop on Multiple Classifier Systems, Lecture Notes in Computer Science, Cambridge, UK, Springer-Verlag, vol. 2096, pp. 78, 2001.

[22] G. P. Zhang, "Neural Networks for Classification: A Survey", IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications And Reviews pp. 451-462, November, 2000.

[23] K. Hornik. "Approximation Capabilities of Multilayer Feedforward Networks", Neural Networks pp.251–257, 1991.

[24] A. SAMIR Elons,"3D CONTINUOUS REAL-TIME ARABIC SIGNLANGUAGE RECOGNITION", A Thesis Submitted to the Department of Scientific Computing, Faculty of Computer &Information sciences Ain Shams University, Cairo 2012

[25] A. Samir Elons, and Magdy Aboul-Ela, and M. Fahmy Tolba; "A Proposed PCNN Features Quality Optimization Technique for Cameras Weighting in Pose-Invariant 3D Arabic Sign Language Recognition", Applied Soft Computing Journal, Springer, ELSEVIER, 2012.

[26] A. Samir Elons, and Magdy Aboul-Ela, and M. Fahmy Tolba; "3D Object Recognition Using Multiple 2D Views for Arabic Sign Language", Journal of Experimental & Theoretical Artificial Intelligence, Taylor & Francis, 2012.

[27] A. Samir Elons, and Magdy Aboul-Ela, and M. Fahmy Tolba; "Arabic sign language continuous sentences recognition using PCNN and graph matching", Neural Computing and Applications Journal, Springer, 2012.