Neural Networks Based Recognition System for Isolated Arabic Sign Language

Mohammad AL-Rousan
Jordan University of Science and technology, Computer Engineering Department,
P.O. Box 3030, Irbid, Jordan
alrousan @just.edu.jo

Omar Al-Jarrah
Jordan University of Science and technology, Computer Engineering Department,
P.O. Box 3030, Irbid, Jordan
aljarrah @just.edu.jo

N. Nayef
Jordan University of Science and technology, Computer Engineering Department,
P.O. Box 3030, Irbid, Jordan
n_nayef81@hotmail.com

ABSTRACT

In this paper, we present a vision based intelligent system that provides a feasible solution to Arabic Sign Language recognition at word level. The proposed system is based on applying image processing and intelligent techniques on digital images of the video-captured signs. The outcome of the proposed method is the interpretation of these sequences of images (signs) into the corresponding spoken words. The system uses neural networks as its main classification engine. A large data set has been collected to train and test the system. Experimental results showed that the proposed recognition system provides high recognition accuracy for tested gestures ranging from 95% to 100%. The system does not require the use of any special gloves for inputting gestures.

Keywords: Arabic Sign Language, DCT, Neural Networks, Recognition.

1. INTRODUCTION

Humans express their emotions not only verbally but expressing the emotions also involves non-verbal means and physically sensible actions. We use our faces, hands and body as an integral part of our communication when talking to another human; faces change expressions continuously and both deliberate and unintentional gestures accompany our speech [1], we even make spontaneous gestures while speaking on the telephone. So, the ability to interpret the context of communication is one of the necessary skills for the computers to understand human emotions and hence, to interact intelligently with their human users. In order to attain this ability; we need to develop technologies that can effectively track human movements, body behavior, facial expressions and speech, and interpret these to a computer readable form. Recent advances in image analysis and machine learning open up the possibility for these developments in Human Computer Interaction (HCI) research.

Arabic sign language (ArSL) is the first and primary language of communication for Arabian Deaf and speech-impaired people, and to date; a small amount of work has been done on ArSL recognition with respect to research in other sign languages, whilst there is a growing need for such systems. In this work we aim at designing and implementing an automated robust signer system for Arabic sign language (ArSL) recognition from videos of signs at word level that is capable of recognizing signer manual movements. Samples of ASL gestures are shown in Figure 1. The system does not force the user to wear any cumbersome device or any type of gloves.

B. Bauer and H. Hienz [2] in 2000 developed a GSL (German Sign Language) recognition system that uses colored cloth gloves in both hands. The system is based on Hidden Markov Models (HMM) with one model for each sign. A lexicon of 52 signs was collected form one signer both for training and

classification. A 94% recognition percentage was achieved, and when enlarging the lexicon the recognition rate drops. The system is not real time.

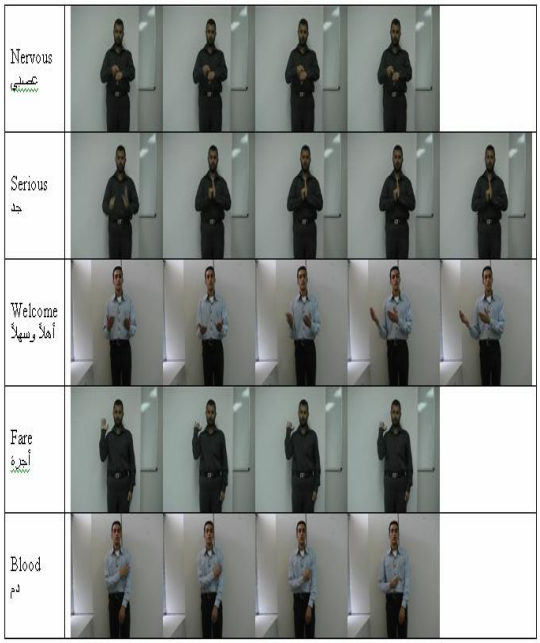


Figure1 Samples of ArSL gestures

M. Al-Rousan et al. [3] and O. Aljarrah et al. [4] in 2000 and 2001 respectively, developed two systems for recognizing 30 static gestures of Arabic sign language using a collection of Adaptive Neuro-Fuzzy Inference System (ANFIS) networks for training and classification depending on spatial domain features, the first system used a colored cloth glove to aid segmentation and feature extraction, and achieved an average of 93.5% recognition rate, but for the testing set the rate was only 84.5%, while the second system did not use any type of gloves or visual markings, and it was able to recognize the 30 Arabic manual alphabets with an accuracy of 93.55% with 87% success rate for the testing set.

N. Tanibata et al. [5] -in 2001- proposed a method of extraction of hand features and recognition of JSL (Japanese Sign Language) words. For tracking the face and hands, they initialized the face and hand regions by matching the initial pose template, and decided the range of skin color at the first frame, under the assumption that an image for the background is given by a fixed camera, and that each word starts from the same initial pose. They tracked the overlapping face and hands by matching the texture template of the previous face and hands. 6 spatial features of the hands were extracted for recognition using the Hidden Markov Models (HMM), the samples were taken in a complex background whose color is not similar to the skin color, and the clothes of the signer are long sleeved and not similar to the skin color. Depending only on the probability of the HMM, they could recognize 64 out of 65 words successfully, that means a percentage of 98.4%, but the vocabulary that they used includes no minimal pairs nor signs that are similar to a part of another sign, they only word that matches the latter condition was not recognized. The system is person dependent, but it was trained and tested on different persons. No details were mentioned on the speed of recognition.

In 2003 k. Assaleh et al. [6] used colored gloved for collecting a varying size data samples for 30 manual alphabet of Arabic sign language, polynomial classifiers were used as a new approach for classification. Polynomial classifiers outperform the previous approaches in [7] on the same data

Chen et al. [8] introduced in 2003 a system for recognizing dynamic gestures (word signs) for TSL (Taiwanese Sign Language), they used frequency domain features (Fourier Transform) plus some information from motion analysis for recognizing 20 words, the data set was collected from 20 signers but the system is person dependent, they tracked the hand depending on that it is the only moving object in the background then applying skin color segmentation, their tracking method is real time but the system is not. HMMs were used as the classifier. An average of 92.5% recognition rate was achieved.

Zahedi et al. [9] developed in 2003 an appearance based recognition system for ASL, they extracted features from the image directly without applying any segmentation or tracking of hands, and they found that high recognition rates still can be achieved with the advantage of lowering the complexity of the system, but they used a small vocabulary of 10 words and a varying number of samples per sign, the system is person dependent. The performance varies from 74% to 93% using HMM classifiers. No details about the speed of the system.

In 2004 and 2005, J. Zieren et al. [10, 11] presented two systems for isolated recognition, the first is for recognizing GSL, on a vocabulary of 152 signs achieving a rate of 97% using HMM, but the rate decreases for the group of signs that contain overlaps in either hands or face and hands, they used sophisticated tracking and segmentation methods and yet put restrictions on the image of the signer, the system is person dependent. The second system they made after, was to recognize BSL vocabulary of 232 signs, but the high recognition rates (average of 87%) were achieved on a very small vocabulary of 6 to 18 signs!, this is for signer dependent, they collected data from 6 signers and experimented person independent recognition and achieved 44.1% rate, this is of course not counted as signer independence. As in the first system a lot of work is done on tracking and segmenting hands, but in controlled environments. Both systems are too slow to be real time.

In a recent (2005) work in Arabic Sign Language, Mohandes et al. [12] developed a system that recognized 50 signs of words performed by one person having 10 samples per sign, they used a training: testing ratio of 70:30, their system has the limitation of using colored cloth gloves in both hands, they detect the signer face by a hybrid technique of Gaussian skin color model and region growing, so the detection of the face is computationally expensive, the hands were tracked by their color, and the system operates under controlled environmental conditions. HMMs were used for classification and achieved a recognition accuracy of 98% based on spatial domain features. The system is not real time.

The paper is organized as follows, In Section 2 we discuss the system architecture, followed by a discussion about the feature extraction method we used in our system. Section 4 presents the system modeling using neural networks, then the results are discussed in Section 5. Finally, we summarize in Section 6.

2. System Architecture

Figure 2 illustrates the proposed system architecture; it manifests the system constituting components and the way they are connected to each other. The system is mainly composed of several modules including the input, preprocessing, feature extraction, and recognition modules.

The first module (input) acquires signs performed by a deaf person communicating with the system using Arabic sign language; the video camera captures this. A group of signs that represent words and phrases are collected as the data set for this system. The collected data is in form of variable length video segments.

In the second module, the preprocessing phase, the captured video segments are converted into image frames. Now each sign is represented by a sequence of frames. Since the camera we used is of good quality; no filtering or any other noise reduction operations are needed. The region of interest in the images is the face and hands of the signer, so we first subtract the background then we use a simple algorithm that depends on the skin color to extract the required parts of the signer body.

In the third module, a set of features will be extracted from the word/ phrase signs, and among the characterizing features of the signs images, the best are chosen to be extracted and put in the feature vector.

We have two main approaches in doing this; the first is taking the feature vector from each image in one repetition and then constitute the final feature vector from putting the images' feature vectors one after another, and of course the sequence of the vectors is a feature by itself.

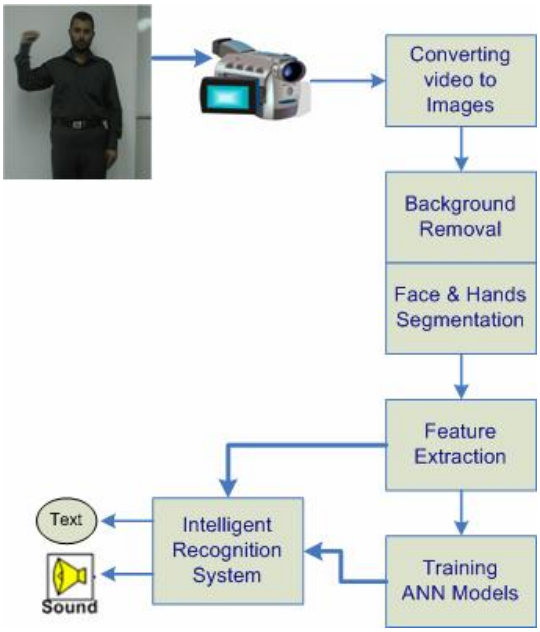


Figure 2: Arabic SLR System Layout

The second approach is projecting all the frames of one repetition into a single frame by image arithmetic addition, so that each sign repetition is represented by one image and consequently one feature vector. We add the second frame to the first, and then the third to the result of first addition and so on.

The Discrete Cosine Transform (DCT) is used to obtain the feature vectors in our work. Section 3 describes the DCT theory and fundamentals.

Recognition process in the fourth module is carried out as follows; first the classification tool –Artificial Neural Networks- is trained on recognizing signs from their representative features; the training is carried out on 60% of the data set which is referred to as the training set. The training process is performed before the system can be used (i.e. off line), after training the classifier; it is used for recognizing signs form the test set. The trained classifier has the extracted feature vector of the sign to be recognized - which comes from the previous phase- as an input, and outputs the recognized sign in text form for simplicity.

3. Feature Extraction

Generally, feature extraction falls into two domains: spatial and frequency domains. Spatial domain features include area, centroid, eccentricity, and orientation. Frequency domain features can be obtained via transformations such as the fast Fourier transform (FFT), discrete cosine transform (DCT), and discrete wavelet transform (DWT). Among spectral transformations, DCT is known for its energy compaction property which implies preserving the

image information via a reduced number of coefficients.
The two dimensional DCT coefficients C(u,v) of the image x(m,n) can be obtained from

$$C(u,v) = a_u a_v \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} x(m,n) \cos\left[\frac{\pi u(2m+1)}{2M}\right] \cos\left[\frac{\pi v(2n+1)}{2N}\right] \quad (1)$$

$$\text{Where } a_v = \begin{cases} \frac{1}{\sqrt{N}}, b = 0 \\ \sqrt{\frac{2}{N}}, 1 \le b \le N-1 \end{cases} \quad (2)$$

$$a_u = \begin{cases} \frac{1}{\sqrt{M}}, a = 0 \\ \sqrt{\frac{2}{M}}, 1 \le a \le M-1 \end{cases} \quad (3)$$

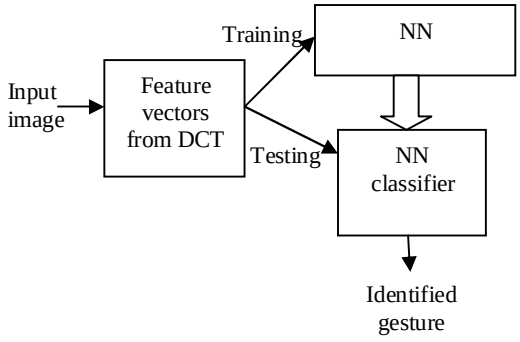
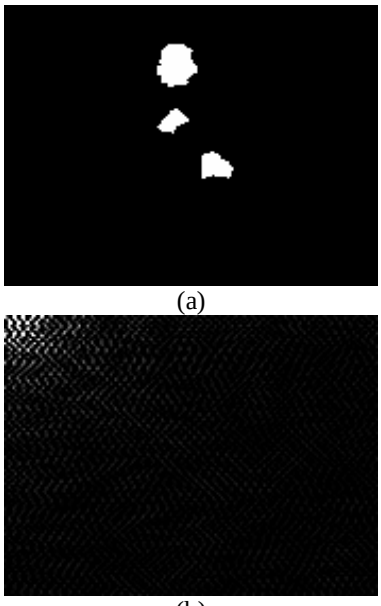


Figure 3: The training and testing

In Discrete Cosine Transform (DCT) almost all information in the image transform is encompassed in less than 1/4 of the coefficients (in the top left corner). So we apply a coding theme; the zigzag coding; to have the number of features we want to extract. An illustration of the DCT of an image frame from our data is shown in Figure 4(b). The extracted coefficients construct the feature vector used to represent the sign.
We have experimented different numbers of features from the DCT coefficients, for example, for the resolution 234x192, we have tested 6, 10, 15, 21, 28, 36, 45, 55, 66, 78, 91, 105 and 120 coefficients. The size of feature vector for one sign is the number of coefficients multiplied by the number of images in one repetition, if the first approach in feature extraction is adopted, while if the second approach (the frames projection approach) is used then the feature vector size for one sign equals the number of extracted coefficients.



(a)

(b)

Figure 4-6: (a) part of the sign "Bus" (b) DCT of (a)

Once the features are extracted, they are used for training and testing the system (see Figure 3).

4. Neural Network Classifier

Many types of ANN have been tested; two models have given the best results; the probabilistic neural networks and the multilayer feed forward neural network with back propagation (see Figure 5).
A two layer feed forward neural network with back propagation is built to be used as the classifier in our ArSL system. Many trials have been made to find the best design for the network in order to achieve the best performance, among these we tried to increase the number of hidden layers, change the number of neurons in each layer and test many different learning algorithms. In the 2 layer FFBP network; we assign different sizes for the only hidden layer that determines the amount of complexity embedded in the network. The size that gives the best performance varies according to the size of feature vector, i.e., number of features. We have used sizes from 120 to 380 neurons. The output layer size is 40 neurons, which is the number of the classes. The network is illustrated in Figure 5.
We create the network by giving to it some design parameters, like specifying the learning rate, the performance function type and error goal among others. Then we invoke the train function with the input matrix that has all feature vectors of all repetitions and all signs; and the right targets, so that the network knows to which sign class the input

feature belongs. The training is done by updating the weights of the neurons till reaching the error goal defined by us, if goal is not reached; the training process stops by one of the stopping criteria, for example reaching the maximum number of epochs which is defined also by us.

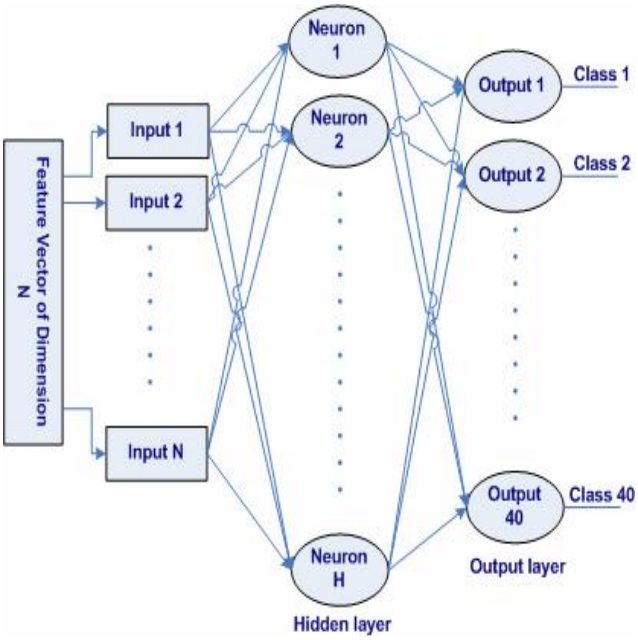


Figure 5: The 2-layer FFBP Neural Network used in our system

Having the trained network; we can use it now in recognizing signs from the testing set by invoking the simulation function, which takes the sign's feature vector and the trained network as inputs and computes the outputs according to the weights of the neurons then it finds the output of the maximum weight and recognizes the sign to be of its class.

5. EXPERIMENTAL RESULTS

All video capturing is done in in-door environments under normal lighting conditions; a digital video camera of high resolution is used. The data set of the system is collected from 14 signers; the vocabulary is 40 different signs taken from 6 different domains. The used vocabulary is shown in Table 1. Each signer is to repeat each sign 5 times, this makes the data size 14*40*5 (2800) samples.

In all tests we present here, the training set size is 60% of the data set size. That means each signer contributes to the training set by 3 random repetitions out of the 5 repetitions performed by him/ her. Therefore, a total of 1680 gestures (3 repetitions*40 gestures* 14 signers) are used for training. The other 2 repetitions are part of the testing set, which has the size of 40% of the data set, with a total of 1120 gestures.

We start by a testing a factor that is important to improving the system speed, this factor is the image resolution. We would like to make each image frame as small as possible so as to make all the computations faster. For example, if the image is small; the speed of the data preparation module improves, as the subtraction and segmentation operations operate faster. The speed of the feature extraction module improves too, as the transforms are computed in less time, and moreover, a smaller number of features encompass most of image information. The results in Table 2 show the recognition accuracy using the DCT as the feature extractor. 0.06% of the transform coefficients are used to represent the feature vector for one image.

Table 2: Performance against image resolution

Image Resolution	Recognition Rate
140x115	90%
176x144	92%
234x192	97.3%
351x288	94%
702x576	89.2%

Form these results we can see that the bigger the resolution is, the better the results. But also it is noticed that after certain resolution, the difference between the rates gets smaller. The reason behind these results is that when the image is of high resolution, its details become clearer, and its transform carry more discriminating features.
The next test in isolated recognition is related to vocabulary size and characteristics. The effect of reducing vocabulary size on the recognition rate is tested, first we test removing random signs, then removing the error pone signs, which are the minimal pairs. Also we test the effect of putting each group of the minimal pairs into one class. Again these tests are carried out using the parameters that gave the best results in the previous tests.

Table 3: Performance against vocabulary size with removing random signs

.

Vocabulary Size	Recognition Rate
10	100%
20	98.5%
30	98%
35	97.6%

It shown in Table 3 that with decreased number of classes- the NN classifier produces better recognition results. We have three groups of minimal pairs, the first group consists of two signs; sign 1 and sign 15, the second consists of four signs; sign 2, 31, 36 and

37, and the third consists of two signs; sign 17 and 34. Table 4 shows the results when the signs: 15,31,36,37 and 34 are removed. However, we want to maximize the recognition rate for all gestures. One way to achieve this is by grouping the signs that have close similarity with each other. We have tested the system after rearranging the gestures into 35 groups (classes). The result is shown in Table 5.

Table 4: Performance against vocabulary size with removing minimal pairs

Vocabulary Size (removed Sign)	Recognition Rate
39 (15)	97.3%
36(15,31,36,37)	97.6%
35(15,31,36,37,34)	98%
34(15,31,36,37,34,40)	98.5%

Table 5: Performance against vocabulary size with grouping minimal pairs

Vocabulary Size (40)	Recognition Rate
Classes (35)	97.6%

By comparing Tables 4 and 5, we can see a slightly better performance when removing the minimal pairs than when grouping them. For grouping a number of signs, the training is done for each of the signs in the group, so the recognition rate improves because we count recognizing the sign as any sign in its group, a correct recognition. But the existence of these models will affect to some degree the recognition of other signs outside the group. Therefore, removing the error-prone signs completely gives better results. This also is consistent with the result of decreasing the vocabulary size.

6. CONCLUSION AND DISCUSSION

We have introduced a vision based recognition system for Arabic Sigh Language (ArSL) based on DCT for feature extraction and neural networks for classification. We have validated our system design on a database that we collected in Jordan. The recognition results were 100% on the training data and exceeding 95% on the test data. When the confusable sets were grouped, the overall recognition rate jumped to 97%. The obtained recognition rates are viewed to be satisfactory considering that the database collection was done in a totally unsupervised manner without any special instructions to the signers.

It is worth mentioning that the system we built in this work provides high freedom to signers who have used it; it requires no gloves of any type and no special lightning.

Currently, we are considering building the system using Hidden Markov Model for classification, and compare its performance with NN-based system. We do so, because it is expected that NN would not perform well for large number of signers. We also are aiming at extending our system so it can recognize continuous sentences in ArSL. Since previous work on other sign languages have not proposed satisfactory approaches for such problem, we are testing several choices to tackle this challenging issue.

Table 1: 40-Signs vocabulary used in our system

Sign No.	Arabic Sign (word/phrase)	English Meaning	Sign No.	Arabic Sign (word/phrase)	English Meaning
Domain 1: Adjectives and Feelings			Domain 4: Money and Commerce		
1	مبسوط	Happy	22	نقود	Money
2	مهم	Important	23	ربح	Profit
3	عصبي	Nervous	24	مصرف	Bank
4	جميل	Beautiful	25	دينار	1 JD
5	حب	Love	26	مجاناً	Free of Charge
6	قلق	Worry	Domain 5: Hospital		
7	جد	Serious	27	مستشفى	Hospital
8	جديد	New	28	طبيب	Doctor
Domain 2: Home and Visits			29	مريض	Patient
9	جار	Neighbor	30	صداع	Headache
10	صديق	Friend	31	دواء	Medication
11	هدية	Gift	32	حقنة	Injection
12	أهلاً وسهلاً	Welcome	33	فحص	Laboratory Examination
13	بيت	Home	34	دم	Blood
14	كيف حالك	How are you	35	عملية جراحية	Surgical Operation
15	السلام عليكم	Hello	Domain 6: Miscellaneous		
16	مشاكل	Problem(s)	36	طعام- يأكل	Food/ eat
Domain 3: Roads and Transportation			37	ماء- يشرب	Water/ drink
17	حافلة	Bus	38	كم	How many/much
18	تكسي	Taxi	39	أين	Where
19	مجمع	Station	40	السبب - لماذا	Why /because
20	أجرة	Fare			
21	شارع	Street			

REFERENCES

[1] Mehrabian, "Communication Without Words", Psychol. Today, Vol 2, No. 4, , 1968, pp. 53–56

[2] B. Bauer and H. Hienz, "Relevant Features for Video-Based Continuous Sign Language Recognition", Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition 2000, pp. 64–75

[3] M. Al-Rousan, O. Aljarrah and M. Hussain, "Automatic Recognition of Arabic Sign Language Finger Spelling", International Journal of Computers and Their Applications, Issue 1076-5204, Vol. 8, No.2. 2001, pp. 80-88.

[4] O. Al-Jarrah and A. Halawani, "Recognition of Gestures in Arabic Sign Language Using Neuro-Fuzzy Systems", ACM journal of Artificial Intelligence, Volume 133, No. 1-2, 2001, pp. 117-138.

[5] N. Tanibata, N. Shimada and Y. Shirai, "Extraction of Hand Features for Recognition of Sign Language Words", Proceedings of the 15th International Conference on Vision Interface, Calgary, Canada, 2001.

[6] K. Assaleh and M. Al-Rousan, "Recognition of Arabic Sign Language Alphabet Using Polynomial Classifiers", EURASIP Journal on Applied Signal Processing society No. 13, 2005, pp. 2136-2145

[7] W. M. Campbell, K. T. Assaleh, and C. C. Broun, "Speaker recognition with polynomial classifiers," IEEE Trans. Speech Audio Processing, vol. 10, no. 4, pp. 205–212, 2002.

[8] F-S. Chen, C-M. Fu and C-L. Huang, "Hand gesture recognition using a real-time tracking method and hidden Markov models", Image and Vision Computing, 2003; No. 21, pp. 745–758.

[9] M. Zahedi, D. Keysers, and H. Ney, "Appearance-Based Recognition of Words in American Sign Language", 2nd Iberian Conference on Pattern Recognition and Image Analysis, 2005, pp. 511–519.

[10] J. Zieren and K-F. Kraiss, "Non-Intrusive Sign Language Recognition For Human-Computer Interaction", in 9th IFAC/IFIP/IFORS/IEA Symposium Analysis, Design, and Evaluation of Human-Machine Systems, Atlanta, GA, 2004, pp. 221-228.

[11] J. Zieren and K-F. Kraiss, "Robust Person-Independent Visual Sign Language Recognition", Proceedings of Pattern Recognition and Image Analysis, Second Iberian Conference, Estoril, Portugal, 2005.

[12] M. Mohandes and M. Deriche, "Image based Arabic sign language recognition", Signal Processing and Its Applications, 2005. Proceedings of the Eighth International Symposium, Vol. 1, 2005, pp. 86- 89.