User-independent recognition of Arabic sign language for facilitating communication with the deaf community

T. Shanableh[1] and K. Assaleh[2]
Department of Computer Science and Engineering[1]
Department of Electrical Engineering[2]
American University of Sharjah
tshanableh@aus.edu

Abstract

This paper presents a solution for user-independent recognition of isolated Arabic Sign language gestures. The video based gestures are preprocessed to segment out the hands of the signer based on color segmentation of the colored gloves. The prediction errors of consecutive segmented images are then accumulated into two images according to the directionality of the motion. Different accumulation weights are employed to further help preserve the directionality of the projected motion. Normally, a gesture is represented by hand movements; however, additional user-dependent head and body movements might be present. In the user-independent mode we seek to filter out such user-dependent information. This is realized by encapsulating the movements of the segmented hands in a bounding box. The encapsulated images of the projected motion are then transformed into the frequency domain using Discrete Cosine Transformation (DCT). Feature vectors are formed by applying Zonal coding to the DCT coefficients with varying cutoff values. Classification techniques such as KNN and polynomial classifiers are used to assess the validity of the proposed user-independent feature extraction schemes. An average classification rate of 87% is reported.

Keywords: Digital video/image processing; Sign language recognition; motion analysis; feature extraction; pattern classification.

1. Introduction

Recently, there has been a serious need for the deaf community in the Arab World to be able to communicate and integrate with the rest of the society. The deaf community has been accustomed to conducting most of its daily affairs in isolation and only with people capable of understanding sign language. This isolation deprives this sizable segment of the society from proper socialization, education, and aspiration to career growth. This lack of communications hinders the deaf community from deploying their talents and skills in benefiting the society at large. This paper addresses this problem by proposing a solution for the user-independent Arabic Sign Language recognition technique that facilitates communication between the deaf community and the rest of the society.

Interest in automatic Sign Language Recognition (SLR) research has started about twenty years ago particularly for American [1], Australian [2], and Korean [3] sign languages. Since then many techniques and algorithms have been proposed using a variety of methods based on sensor fusion, signal processing, image processing, and pattern recognition methods. The application was extended to several international sign languages including Japanese [4], Chinese [5], and to a lesser extent Arabic.

Although used in over 21 countries covering a large geographical and demographical portion of the world, Arabic sign language (ArSL) has received little attention in SLR research. To date, only small number of research papers, mainly on finger spelling or isolated gesture recognition, has been published on ArSL.

Work on the recognition of Arabic sign language started with classification of isolated alphabets as reported in [6]. The dataset is compiled from static images of hand postures that represent different letters. Signers are asked to wear gloves with colored finger tips where each fingertip has a distinguished color. In the feature extraction processes, vectors from the center of the hand to each fingertip are computed. The feature vector is then comprised of the magnitude and phase of each vector. Model estimation and validation were carried out using Polynomial classifiers.

More recently, recognition of video-based isolated Arabic sign language gestures are reported by the authors in [7] and [8]. The dataset is based on 23 gestures performed by 3 signers. The data collection phase did not impose any restrictions on clothing or background. A variety of novel feature extraction techniques were proposed. For instance, in [7] the forward or bi-directional prediction error of the input video sign was accumulated and thresholded into a single image. The still image is then transformed into the frequency domain and a feature vector for each gesture is derived accordingly. Simple

classification techniques such as KNN, linear and Bayesian were used to classify the feature vectors. The work in [7] work was extended in [8] where block-based motion estimation techniques are used to find motion vectors between successive images. Such vectors are then rearranged into intensity images and transformed into the frequency domain. Since in this case the temporal dimension of the input video gesture is retained, Hidden Markov Models (HMM) are used for model estimation and validation.

In any case, the reported work on isolated Arabic sign language gestures thus far is limited to user-dependent mode of recognition. Therefore, the main contribution of this paper is to extend the reported work into user-independent classification of isolated gestures. For this task, we propose a number of user-invariant pre-processing and feature extraction schemes.

Relevant work on other sign languages is reported in the literature. For instance, user dependent recognition of British isolated gestures was proposed in [9]. Their work demonstrated that detection can be achieved with reasonable accuracy compared to segmentation using colored gloves. A face detection module is used and the image background is modeled using a normalized histogram. Skin segmentation is realized by applying a likelihood ratio of face to background for each and every pixel in the image. Feature extraction is performed on the segmented images by calculating a set of invariant moments. Nevertheless, no results are available for user independent classification; thus the applicability of such solutions to user independent classification is unclear. Moreover, the use of face detection to segment out the hands leads to restrictions on clothing where skin-like colors should be avoided.

A more sophisticated HMM-based system reported in [10] introduced user independent recognition of isolated British gestures. A classification rate of 87.8% is achieved for six signers in uncontrolled environments for 18 isolated gestures. The extracted features are normalized for person-independence and robustness. Again, hand segmentation is realized through face detection. For the latter, the skin probability threshold is computed automatically by defining a metric to quantify the deviation of a boundary from that of an average face. Background modeling and removal are realized by calculating the median of a given pixel location over time. Motion tracking is based on the color of the hands. This results in many motion transition hypotheses. The most probable path based on high level semantics is found. Feature extraction is then based on segmented images where geometric features of the hands are calculated.

Unlike our proposed solution, image temporal dependencies are retained; hence the need for HMMs. Additionally, finding the most probable motion transition hypothesis is rather time consuming. Again, the use of face detection to segment out the hands leads to restriction in clothing where long sleeves are required and the color of the clothing should not resemble that of the skin.

The work reported in [10] is extended in [11]. The extended work uses a similar feature extraction technique. Isolated signs are collected from 4 native signers. The signers are asked to wear black cloths with long sleeves. The system is enhanced through the adaptation of methods from speech recognition such as Maximum Likelihood Linear Regression (MLLR) and Maximum A Posteriori (MAP) estimation. Very encouraging classification results are reported.

Lastly, [12] presented the recognition of user independent continuous Chinese sentences. The feature extraction is based on the use of datagloves and position sensors. The focus of the work is to segment continuous sentences using modified Simple Recurrent Networks (SRN) to alleviate the effect of motion epenthesis. The experiments are based on 3 signers acting continuous sentences. Note that the use of datagloves and position sensors are unnatural and might hinder the deployment of such a solution.

From the reviewed work it is clear that hand segmentation is a vital step in user independent recognition of sign language. As mentioned above, hand segmentation through face detection leads to imposing restrictions on clothing. On the other hand, the use of datagloves and position sensors is too restrictive. Thus, in this vision-based work we opt to use a simple hand segmentation method based on colored gloves. This is then followed by novel feature extraction techniques that constitute the main contribution of this paper.

This paper is organized as follows. Section 2 describes the dataset of video-based isolated gestures. Section 3 introduces the required preprocessing of such video-based gestures. Section 4 expands upon existing work to extract distinguishing features suitable for user-independent classification. Section 5 describes the various classification techniques used in this work in addition to a detailed description of the polynomial classifier. The experimental setup and results are given in Section 6. Finally the paper is concluded in Section 7.

2. Dataset description

To establish comparisons with exiting work we replicate the dataset of isolated Arabic sign language gestures as reported by the authors in [7]. However, the replicated dataset

was collected from signers with colored gloves. Apart from that, no other restrictions were imposed on clothing or image background. The data was originally collected in collaboration with Sharjah City for Humanitarian Services (SCHS) [13], UAE. We have collected a database of 23 Arabic gestured words/phrases from 3 different signers. The list of words is chosen from the greeting section which is the most commonly used in the communication between the deaf and hearing society. The list of words is shown in Table 1.

#	Arabic word	English Meaning	#	Arabic word	English Meaning
1	صديق	Friend	13	يأكل	To Eat
2	جار	Neighbor	14	ينام	To sleep
3	ضيف	Guest	15	يشرب	To Drink
4	هدية	Gift	16	يستيقظ	To wake up
5	عدو	Enemy	17	يسمع	To listen
6	السلام عليكم	Peace upon you	18	يسكت	To stop talking
7	اهلا وسهلا	Welcome	19	يشم	To smell
8	شكرا	Thank you	20	يساعد	To help
9	تفضل	Come in	21	امس	Yesterday
10	عيب	Shame	22	يدهب	To go
11	بيت	House	23	يأتي	To come
12	انا	I/me			

Table 1: A set of Arabic sign language gestures and their English meanings

Each signer was asked to repeat each gesture 50 times over 3 different sessions resulting in 150 repetitions of all the gestures resulting in a total of 3450 video segments. The videos were captured using a digital camcorder without imposing any restriction on clothing or image background. The signers are diverse in gender and body size which makes the problem more challenging.

3. Data preprocessing

Prior to feature extraction, the video sequences of a given gesture are segmented in the RGB color space. Clearly, this step takes advantage of the colored gloves worn by the signers. Samples of the RGB pixel vectors representatives of the color of the gloves are used to estimate the mean value and covariance matrix of the color to be segmented. Hence, the segmentation process is achieved automatically with no user intervention. The measure of pixel similarities used is the Mahalanobis distance. A pixel vector that falls within the locus of points that describe the 3D ellipsoid is classified as a glove pixel. The threshold used to define the locus of points is set to the maximum standard deviation of the three color components as suggested in [14]. Once an image is segmented, a 5X5 median filter is used to eliminate artifacts caused by the segmentation process.

4. Feature extraction

The purpose of the feature extraction is to extract the motion information from the temporal domain of the input image sequence through successive image differencing. Let $I_{g,i}^{(j)}$ denote image index j of the i^{th} repetition of gesture index g. The accumulated prediction errors or image differences (AD for short) can be expressed as [7]:

$$AD_{g,j} = \sum_{j=1}^{n-1} \partial_j \left(\left| I_{g,j}^{(j)} - I_{g,i}^{(j-1)} \right| \right) \tag{1}$$

Where n is the total number of images in the i^{th} repetition of a gesture at index g. ∂_j is a binary threshold function of the j^{th} frame.

As a result of this process, the temporal dimension of the input video is eliminated and the whole sequence onto one absolute AD image representative of the motion in the video sequence. However, it is worth mentioning that different sign gestures can have very similar AD images. One example, the sign "To Go" when performed in reverse order would look like the sign "To Come". As such, this section proposes the use of weighted directional accumulated image differences. The weighting here refers to the manner by which the predication errors are accumulated into one AD image. More specifically, the first half of the temporal sequence of predication errors are emphasized by assigning to them a larger accumulation weight than that assigned to the second half of the sequence. It is also realized through using two different thresholds for image differencing. Such differences can be categorized into positive and negative accumulated prediction errors (AD_+ and AD_- respectively). Formally, for a given gesture g of a given repetition i, the directional accumulated differences are expressed as:

$$AD_+(x,y) = \begin{cases} AD_+ + w_k & if \quad (f(x,y,t_k) - f(x,y,t_{k-1})) \geq Th_{(k,k-1)} \\ AD_+ & otherwise \end{cases} \tag{2}$$

$$AD_-(x,y) = \begin{cases} AD_- + w_k & if \quad (f(x,y,t_k) - f(x,y,t_{k-1})) \leq -Th_{(k,k-1)} \\ AD_- & otherwise \end{cases} \tag{3}$$

Where (x,y) are the pixel coordinates of the AD image and w_k is the accumulation weight at the k^{th} image difference. Lastly, the AD images are normalized by dividing their pixel values by the number of images in the underlying images sequence.

Once the temporal dimension is eliminated, spatial domain feature extraction schemes are applied on the resultant positive and negative AD images. Since the gloves are already segmented out then the AD images contain the motion of the hands only. To normalize

the differences in spatial coordinates of the hand movements a bounding box that encapsulates the movement is determined, the remaining blank parts of the AD images are discarded. The AD images are then transformed into the frequency domain using Discrete Cosine Transformation (DCT) given by:

/

$$F(u,v) = \frac{2}{\sqrt{MN}} C(u)C(v) \sum_{i=0}^{M-1} \sum_{j=0}^{N-1} f(i,j) \cos\left(\frac{\pi u}{2M} \cdot (2i+1)\right) \cos\left(\frac{\pi v}{2N} \cdot (2j+1)\right) \quad (4)$$

Where *NxM* are the dimensions of the input image *f*, and *F(u,v)* is the DCT coefficient at row *u* and column *v* of the DCT matrix. *C(u)* is a normalization factor equal to $1/\sqrt{2}$ for u=0, and 1 otherwise. More information about DCT transforms are found in [15].

An attractive property of this transformation is its effective energy compaction. Low frequencies are concentrated in the top left corner of the transformed image. Thus the AD images can be represented by traversing the DCT coefficients via zigzag scanning from the top left corner into an *n* dimensional vector. The dimensionality of this vector is empirically determined as illustrated in the experimental results section. This process is known as Zonal coding. The resultant DCT coefficients of both the AD images are interleaved to form a concise yet precise feature vector. Figure 1 illustrates the proposed feature extraction approach.

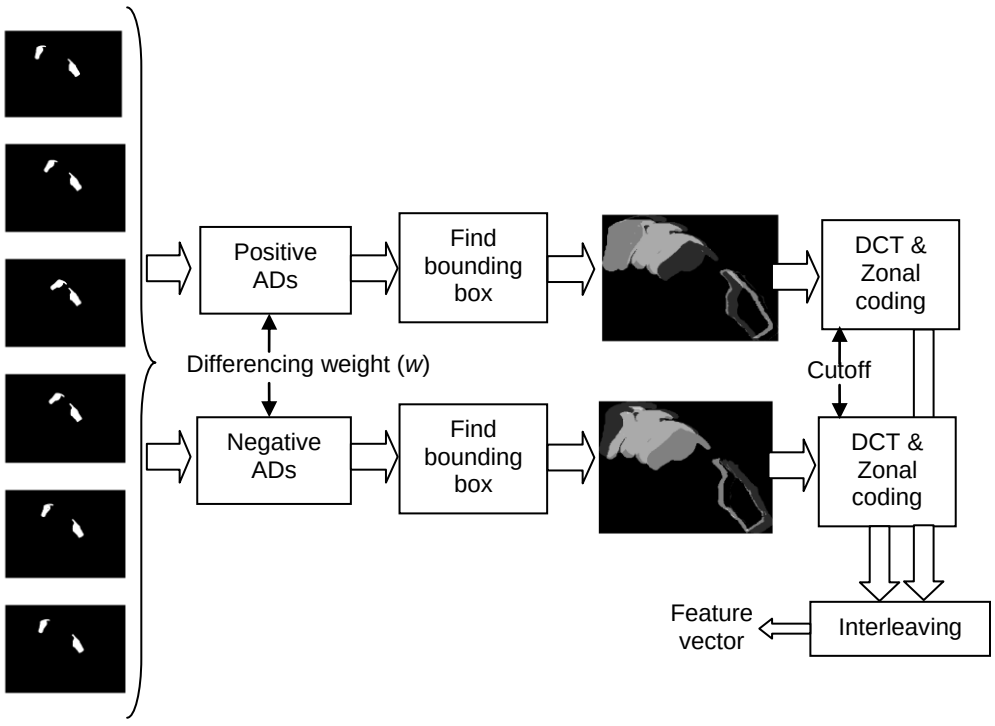

Figure 1. Block diagram of proposed preprocessing and feature extraction schemes

5. Classification techniques:

This work experiments with a number of simple classifiers such as KNN, linear discriminant functions and polynomial classifiers.

In the KNN classifier we experiment with both the Euclidean distance measure and the 'correlation factor' similarity measure. In the latter, the classification decision is based on finding the maximum correlation factor between two vectors which is defined as:

$$\arg\max_{\mathbf{x}_{\text{train j}}} \left(\frac{zscore(\mathbf{x}_{\text{test}})^T \cdot zscore(\mathbf{x}_{\text{train}_j})}{n-1} \right)$$

Where $\mathbf{x}_{\text{test}}$ is a feature vector from the testing set and $\mathbf{x}_{\text{train}_j}$ is the j^{th} feature vector from the training set and n is the dimensionality of the feature vector determined by the zonal cutoff.

We also experiment with linear discriminant functions of the following form: $d(\mathbf{x}) = \mathbf{w}^t\mathbf{x} + w_0$ where $\mathbf{w}$ is a weight vector that determines the orientation of the linear decision hyperplane, $\mathbf{w_0}$ is the bias and $\mathbf{x}$ is the feature vector. In this case the training is done in an offline mode and model parameters are uploaded to the recognition stage. Offline training mode is normally carried out in the user-independent mode or when the training data is composed of a large number of classes or exhibits an excessive variability within each class. Lastly, in this work we have also used polynomial classifiers.

5.1 Polynomial Classifiers

A polynomial classifier is a parameterized nonlinear map which expands a sequence of input vectors to a higher dimensionality for the purpose of making them linearly separable. Training of a P^{th} order polynomial classifier consists of two main parts. Part one is expanding the training feature vectors via polynomial expansion. The purpose of this expansion is to improve the separation of the different gestures in the expanded vector space. Ideally, it is aimed to have this expansion make all gestures linearly separable. Part two is linearly mapping the polynomial-expanded vectors to an ideal output sequence by minimizing an objective criterion. The mapping parameters represent the weights of the polynomial network.

5.1.1 Polynomial Expansion

Polynomial expansion of an M-dimensional feature vector $\mathbf{x} = [x_1 \ x_2 \ ... \ x_M]$ is achieved by combining the vector elements with multipliers to form a set of basis functions, **p(x).** The elements of **p(x)** are the monomials of the form

$\prod_{j=1}^{M} x_j^{k_j}$, where k_j is a positive integer, and $0 \le \sum_{j=1}^{M} k_j \le P$.

Therefore, the P[th] order polynomial expansion of an M-dimensional vector **x** generates an $O_{M,P}$-dimensional vector **p(x).** $O_{M,P}$ is a function of both M and P and can be expressed as

$$O_{M,P} = 1 + PM + \sum_{l=2}^{P} C(M,l) \tag{5}$$

where $C(M,l) = \binom{M}{l}$ is the number of distinct subsets of l elements that can be made out of a set of M elements.

Therefore, for class i the sequence of feature vectors $\boldsymbol{X}_i = [\boldsymbol{x}_{i,1} \ \boldsymbol{x}_{i,2} \cdots, \boldsymbol{x}_{i,N_i}]^T$ is expanded into

$$\boldsymbol{V}_i = [\boldsymbol{p}(\boldsymbol{x}_{i,1}) \quad \boldsymbol{p}(\boldsymbol{x}_{i,2}) \quad \cdots \quad \boldsymbol{p}(\boldsymbol{x}_{i,N_i})]^T \tag{6}$$

Notice that while $\boldsymbol{X}_i$ is a $N_i \times M$ matrix, $\boldsymbol{V}_i$ is a $N_i \times O_{M,\mathrm{p}}$ matrix.

Expanding all the training feature vectors results in a global matrix for all K classes obtained by concatenating all the individual $\mathbf{V}_i$ matrices such that $\boldsymbol{V} = [\boldsymbol{V}_1 \ \boldsymbol{V}_2 \ \cdots \ \boldsymbol{V}_K]^T$.

5.1.2 Solving for the classifier weights

For gesture i, the training problem reduces to finding an optimum weight vector. This weight vector is obtained by minimizing the distance between the ideal output vector y_i and a linear combination of the polynomial expansion of the training feature vectors $\mathbf{V}\mathbf{w}_i$ such that

$$\boldsymbol{w}_i^{opt} = \arg\min_{\boldsymbol{w}_i} \left\| \boldsymbol{V} \, \boldsymbol{w}_i - \boldsymbol{y}_i \right\|_p \tag{7}$$

The ideal output for the i[th] gesture y_i is a column vector comprised of ones and zeros such as $\mathbf{y_i} = [\mathbf{0}_{N_1}, \mathbf{0}_{N2}, ..., \mathbf{0}_{N_{i-1}}, \mathbf{1}_{N_i}, \mathbf{0}_{Ni+1}, ..., \mathbf{0}_{N_k}]^{\mathrm{T}}$. Equation 7 indicates that the weight vector is obtained by minimizing the L^P-norm of the error vector $\mathrm{e}_i = \mathrm{V}\mathrm{w}_i - \mathrm{y}_i$. In this work we use the special case of p=2 thus arriving at the L^2-regression problem. That is,

finding $\mathbf{w}_i^{opt}$ that attains the minimum of the L^2-norm of the error sequence $\mathbf{e}_i$. Or equivalently, minimizing the square of the L^2-norm. Fortunately, for this particular formulation with the L^2-norm there is an explicit formula for the solution $\mathbf{w}_i^{opt}$. This solution can be obtained by applying the normal equations method [16] such as

$$\boldsymbol{V}^T\boldsymbol{V}\boldsymbol{w}_i^{opt} = \boldsymbol{V}^T\boldsymbol{y}_i \tag{8}$$

By incorporating Equation 7, Equation 8 can be rearranged as

$$\sum_{j=1}^{K}\boldsymbol{V}_j^T\boldsymbol{V}_j\boldsymbol{w}_i^{opt} = \boldsymbol{V}_i^T\boldsymbol{1}_i \tag{9}$$

If we define $\boldsymbol{R}_j = \boldsymbol{V}_j^T\boldsymbol{V}_j$, $\boldsymbol{R} = \sum_{j=1}^{K}\boldsymbol{R}_j$, and $\boldsymbol{v}_i = \boldsymbol{V}_i^T\boldsymbol{1}_i$ then Equation 9 yields an explicit solution for $\boldsymbol{w}_i^{opt}$ expressed as

$$\boldsymbol{w}_i^{opt} = \boldsymbol{R}^{-1}\boldsymbol{v}_i \tag{10}$$

The set $\{\boldsymbol{w}_i^{opt}\}$ represents the weights of the K polynomial classifiers which we refer to as the gestures models.

In [17, 18, 19], Campbell and Assaleh discuss the computational aspects of solving for $\boldsymbol{w}_i^{opt}$ and they present a fast method for training polynomial classifiers by exploiting the redundancy of the $\boldsymbol{R}_j$ matrices. They also discuss in details the computational and storage advantages of their training method.

6. Experimental results:

As discussed in the dataset collection section, the data is collected from 3 different signers. Each gesture is repeated 50 times by each signer. The resultant videos are split into a sequence of still images followed by color segmentation. Considering the complexity of the data collection phase, only 3 users participated in the data collection. Therefore, to test the proposed feature extraction schemes in user-independent classification mode, the gestures of 2 users are used for training whilst the gestures of the remaining user are used for testing. The experimental results are based on three different combinations of training and testing as shown in Table 2.

	Training set (Gestures of)	Testing set (Gestures of)
Combination 1	Signers 2,3	Signer 1
Combination 2	Signers 1,3	Signer 2
Combination 3	Signers 1,2	Signer 3

Table 2. Three different combinations for user-independent classification.

Additionally, recall that the feature extraction schemes propose to eliminate the temporal dimension of the video-based gestures through the use of weighted and directional accumulated prediction errors. As such, time-sensitive modal estimators and classifiers such as Hidden Markov Models are not needed.

Common to all of the following experiments, the classification results shown are the average of the three signers using the three combinations given in Table 2. For completeness, the standard deviation values of the classification rates are also represented as error bars.

Figure 2 compares the recognition rate of the proposed work against that of [7]. The figure shows the recognition rate as a function of the DCT Zonal coding cutoff. Note that the objective of the work in [7] was user-dependent rather than independent classification. However, this experiment replicates that work and tests it with user-independent classification. This is referred to as "Existing work" in the figure. Additionally, this experiment re-implemented the existing work with the addition of the proposed weighted directional accumulated prediction errors. This is referred to as "Weighted directional ADs" in the figure.

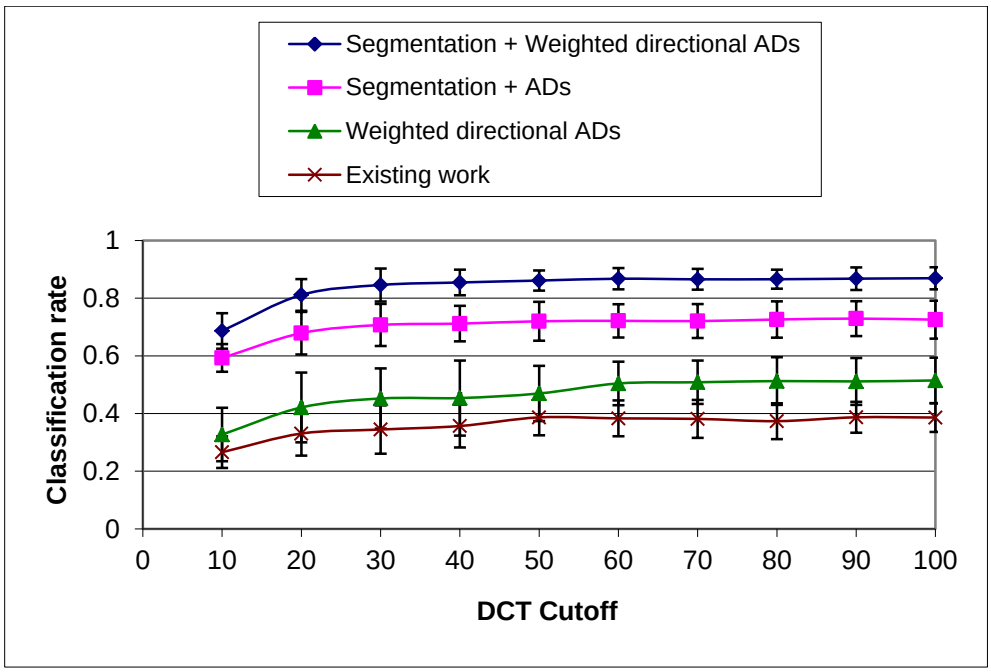

Figure 2. Classification rates of the proposed solution compared to existing work using KNN classifier with K=1.

Figure 2 shows that the existing work clearly fails with user-independent classification. The use of the proposed weighted directional accumulated differences did not result in a significant enhancement. Clearly, the glove-based segmentation and the encapsulation of

the ADs inside a bounding box was vital for the user-independent classification. Additionally, the most accurate classification results are attributed to the use of segmentation and weighted directional accumulated difference. In fact, the use of the latter resulted in an average increase in classification rate of around 14%.

Lastly, it is shown in Figure 2 that the standard deviation of the recognition rates across different users becomes smaller with the use of segmentation and weighted directional accumulated prediction errors.

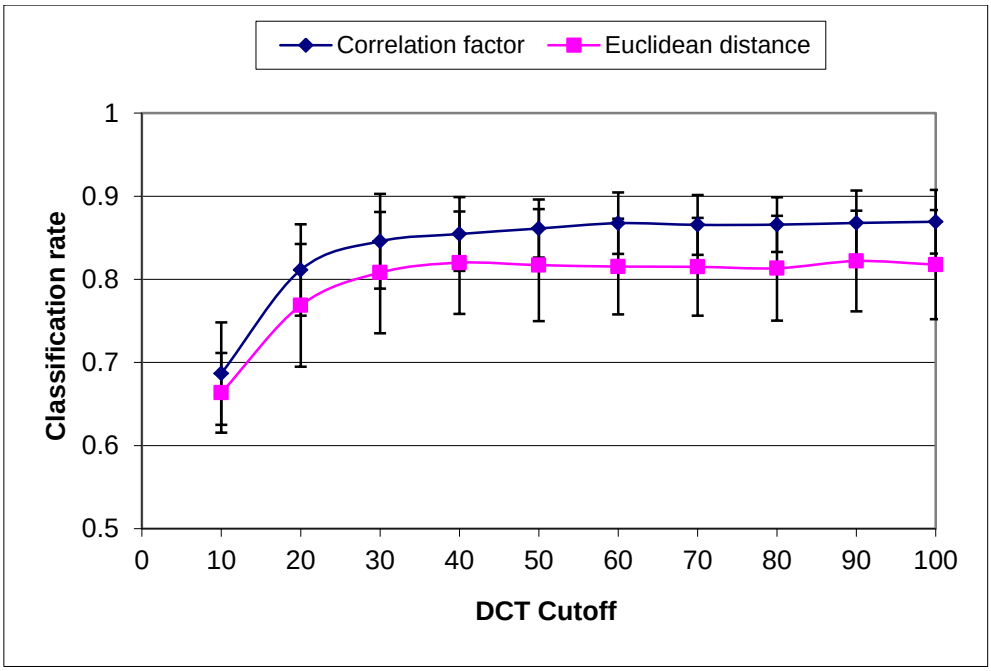

Figure 3. Classification rates of the proposed solution based on KNN classifier using the correlation similarity measure compared to the Euclidian distance measure.

Figure 3 shows the classification rates of the proposed solution based on the KNN classifier using the Euclidean distance measure and the correlation factor measure.

The normalization using z-scores in the correlation factor has improved the classification rates; on average, the rate has increased by 4% across different DCT cutoffs.

To study the effect of reducing the number of training samples per gesture, the KNN classification using the correlation measure is repeated with various numbers of training repetitions per gestures. The number of training samples per gesture used in this experiment are 1, 2, 5, 10, 25, and 50. Figure 4 shows the obtained classification rates for these different sizes of the training data.

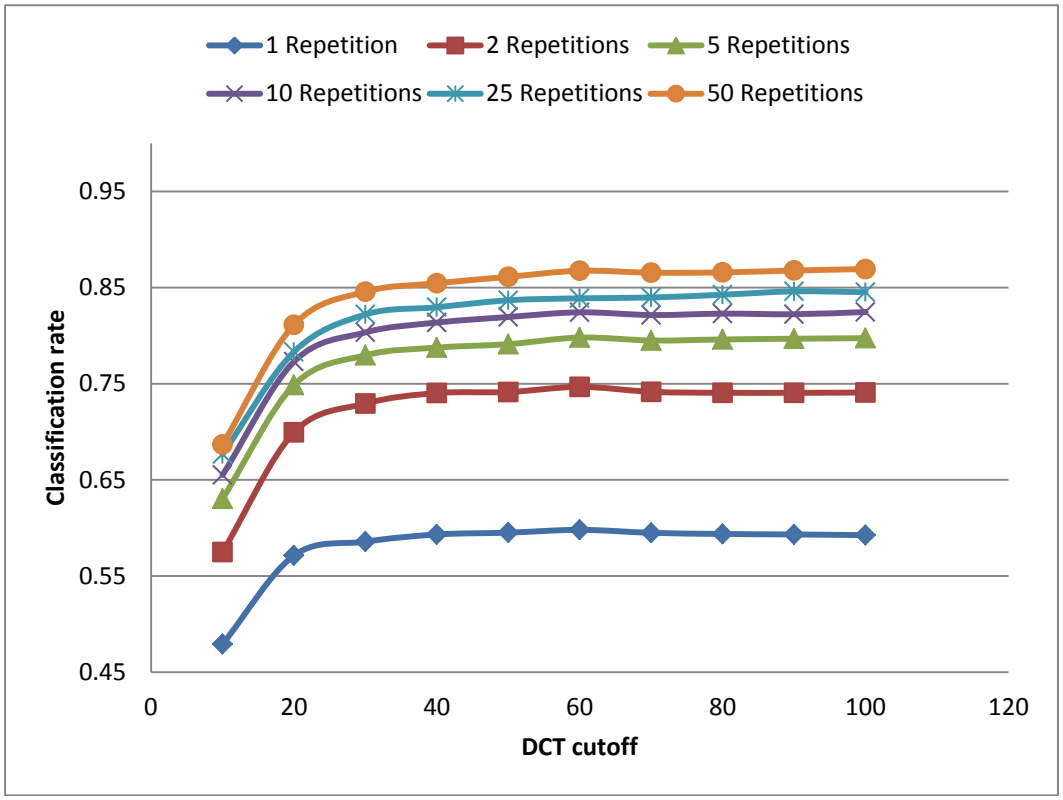

Figure 4.The effect of varying the number of training repetitions (samples per gesture) on the classification rates for different DCT cutoff values.

The repetitions used in training set are selected in a round robin fashion. The resultant classification results are averaged over all the combinations and shown in Figure 4.

The figure shows that training the system with 5 repetitions per gesture results in an encouraging classification rate of around 80% at a DCT cutoff of 60. Clearly, by increasing the number of repetitions, intra-class variations are modeled more accurately resulting in higher classification rates.

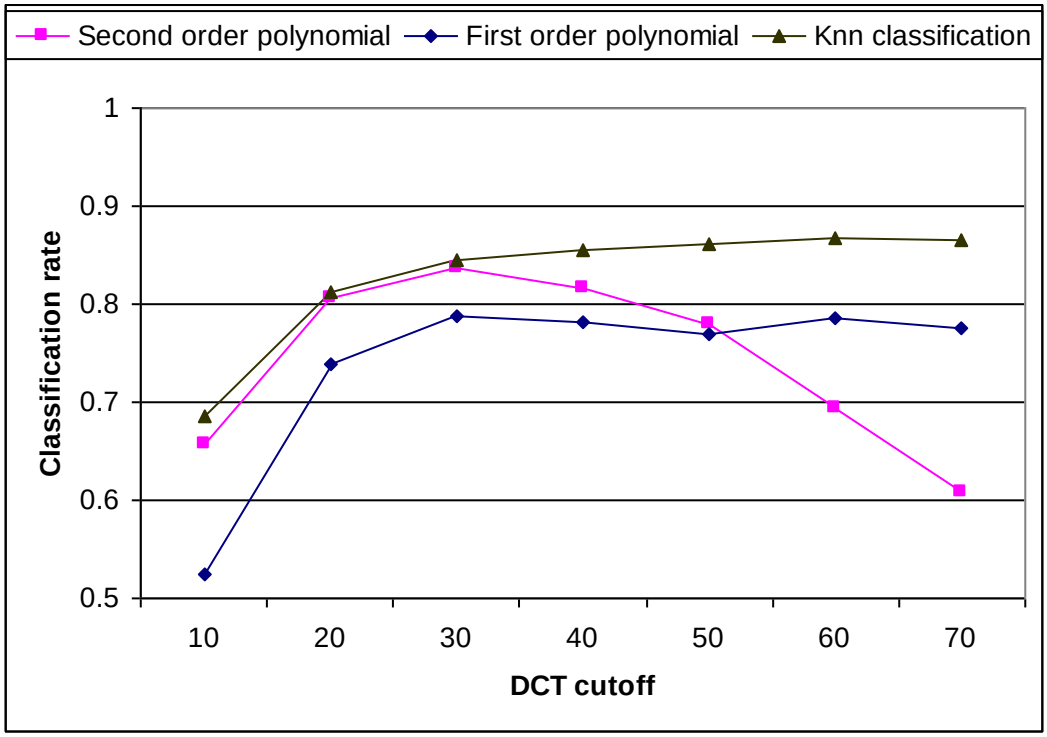

Figure 5. User-independent classification of proposed solution using KNN, 1st order polynomial classifier and 2nd order polynomial classifier.

Figure 5 shows a comparison in recognition rates using three different classifiers. Namely, KNN, 1st order polynomial classifier which is equivalent to linear discriminant function classification, and 2nd order polynomial classifier. It is found that the KNN classifier offers the best recognition rates amongst the three classifiers where the recognition rate consistently improves as the DCT cutoff increases. However, this comes at the expense of storing the entire training feature vector set. Obviously, this scenario becomes impractical as the number of users increases. In such cases, one has to resort to other intelligent techniques to reduce the number of the training feature vectors via clustering methods such as k-means.

The figure also shows the results of using 1st and 2nd order polynomial classifiers. The former classifier is equivalent to the linear discriminant functions and it yields significantly lower recognition rates than the KNN classifier. Note that this classifier models each gesture/class by a weight vector with the same dimensionality of a feature vector which is far smaller in size than the required storage in the KNN case. The difference in recognition rates between KNN and 1st order polynomial classifier suggests that the data is not linearly separable. To make the data linearly separable we project the feature vectors into a higher dimensionality space via 2nd order polynomial expansion.

It is shown in Figure 5 that using a 2nd order polynomial classifier the recognition rates significantly improve over the 1st order classifier approaching the rates obtained by KNN. However, this improvement ceases to exist after a DCT cutoff of 40 due to the exponential increase of the dimensionality of the expanded feature vectors which may results in numerical instability. For instance, at a DCT cutoff of 70, the expanded feature vector length jumps to 2556. As mentioned in Section 5, the computation of the weight vector involves inverting a matrix formed from the expanded feature vectors as illustrated in Equation 10. As the dimensionality of the expanded vector increases the condition number of such a matrix also increases leading to an ill-conditioned matrix that produces an unreliable set of weight vectors. This is illustrated in Figure 6 which shows the increase of the condition number as a function of the DCT cutoff for both 1st and 2nd order polynomial classifiers. The figure shows that the relative condition (in a logarithmic scale) number for the 1st order polynomial classifier is increasing at a much lower rate than that of the 2nd order. Hence, the inconsistency (decrease) of the recognition rates of the 2nd order polynomial at higher DCT cutoffs as shown in Figure 5.

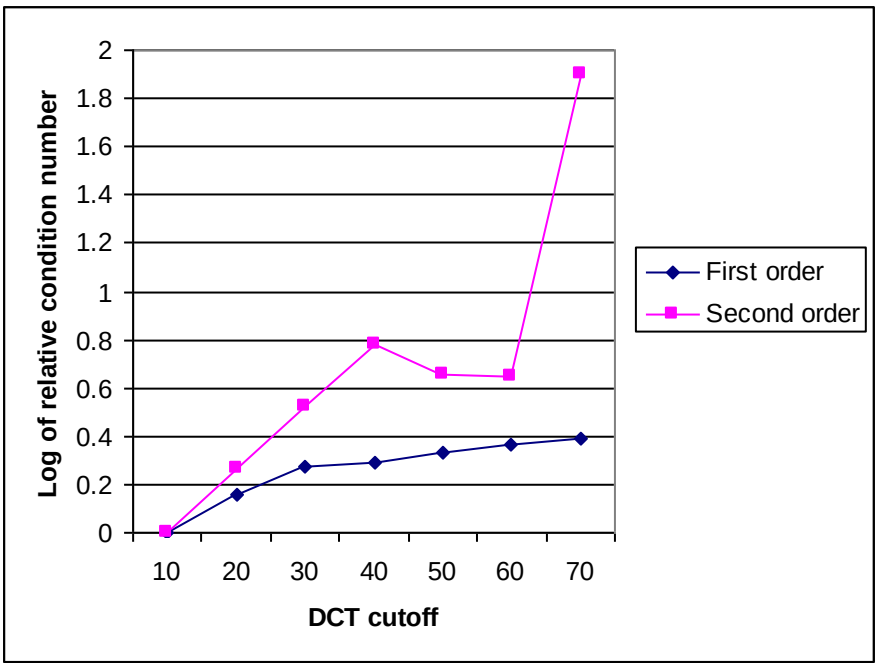

Figure 6. Log of relative condition number of matrix **R** in Equation 10 for 1st t and 2nd order polynomial expansions.

For completeness, we apply the proposed solution in the context of user-dependent rather than independent classification and compare it to previously reported work [7]. Similar to the data arrangement followed in [7], 70% of the dataset comprising all signers is used for training and the remaining 30% is allocated for validation. Figure 7. shows that the previous work which employs absolute ADs and does not rely on hand segmentation is more effective for user-dependent classification. The proposed solution achieves classification results of 95% at a DCT cutoff of 70 which is inferior to the previous solution [7] in terms of user-dependent classification. Recall that the proposed solution includes steps that smear the use-dependent information such as hand segmentation, bounding box and weighted directional accumulated prediction errors. For example, potentially useful user-dependent information from the motion residuals of the body and head movement is being filtered out in the proposed solution. Hence the 5% reduction in user-dependent classification rate when compared to the previous work as shown in the figure.

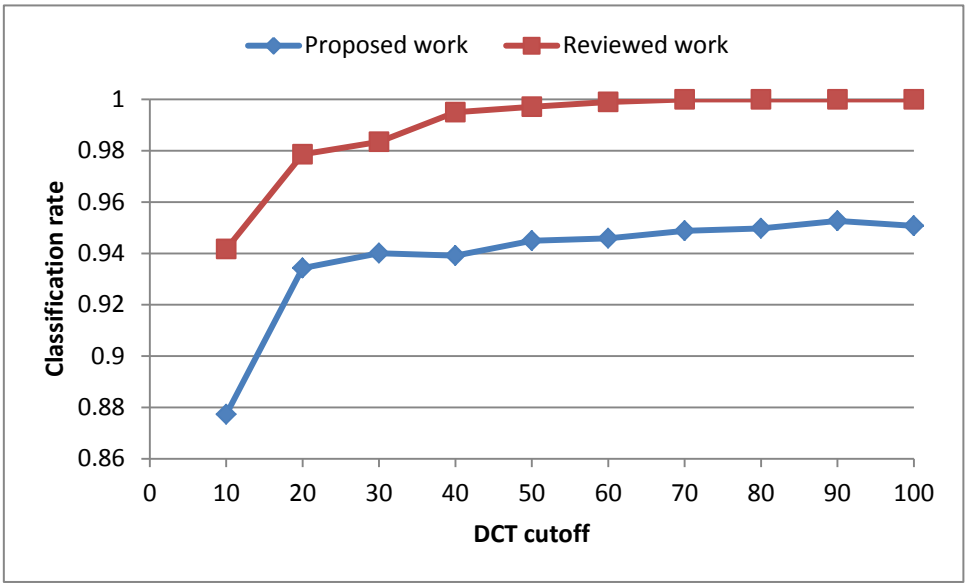

Figure 7. User-dependent classification of the proposed solution versus existing work using KNN classifier.

7. Discussion and conclusion

This paper proposed a set of feature extraction schemes suitable for user-independent classification of isolated Arabic sign language recognition. The eventual application is meant to be for mobile and hand-held devices that can be used by the deaf society. Such devices normally impose constraints on storage and computational resources. This

necessitates the development of efficient feature extraction and classification algorithms that are less demanding in terms of storage requirements and computational complexity. Video-based gestures are preprocessed to segment out the hands of the signer. The prediction errors of successive segmented images are then accumulated into one image. The paper proposed the use of directional and weighted accumulation differences techniques. Bounding boxes around the accumulated differences of the segmented hands are identified. This results in a feature extraction scheme that is insensitive to translation and scaling. The bounded images are then transformed into the frequency domain via DCT. Feature vectors are formed by means of Zonal coding. The proposed solution was validated for user-independent classification using KNN and polynomial classifiers. The classification rate reached 87%. Experiments showed that 2^{nd} order polynomial classifiers and KNN classifiers yield comparable results for feature vector dimensionality up to 30. However, due to numerical considerations, KNN outperforms 2^{nd} order polynomial classifiers for feature vector dimensionality beyond 30. Lastly, it was shown that the proposed feature extraction schemes have a relatively small adverse effect on classification rates when used for user-dependent classification.

References:

[1] T. Starner, J. Weaver, and A. Pentland, "Real-Time American Sign Language Recognition Using Desk and Wearable Computer Based Video," *IEEE Trans. Pattern Analysis Machine Intelligence*, vol. 20, no. 12, pp. 1371-1375, Dec. 1998.

[2] M.W. Kadous, "Machine recognition of Australian signs using power gloves: Toward large-lexicon recognition of sign language," *Proc. Workshop Integration Gesture Language Speech*, pp. 165–174, 1996

[3] J. S. Kim,W. Jang, and Z. Bien, "A dynamic gesture recognition system for the Korean sign language (KSL*)," IEEE Trans. Syst., Man, Cybern. B*, vol. 26, pp. 354–359, Apr. 1996.

[4] H. Matsuo, S. Igi, S. Lu, Y. Nagashima, Y. Takata, and T. Teshima, "The recognition algorithm with noncontact for Japanese sign language using morphological analysis," *Proc. Int. Gesture Workshop*, 1997, pp. 273–284.

[5] C. Wang, W. Gao, and Z. Xuan, "A Real-Time Large Vocabulary Continuous Recognition System for Chinese Sign Language," *Proc. IEEE Pacific Rim Conf. Multimedia*, pp. 150-157, 2001.

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

[7] T. Shanableh, K. Assaleh and M. Al-Rousan, "Spatio-Temporal feature extraction techniques for isolated gesture recognition in Arabic sign language *," IEEE Trans. Syst., Man, Cybern. B*, 37(3), June 2007

[8] T. Shanableh and K. Assaleh, "Telescopic Vector Composition and polar accumulated motion residuals for feature extraction in Arabic Sign Language recognition," *EURASIP Journal on Image and Video Processing, vol. 2007, Article ID 87929, 10 pages, 2007. doi:10.1155/2007/87929.*

[9] H. M. Cooper and R. Bowden, "Large Lexicon Detection of Sign Language," In Proc. Human Computer Interaction Workshop, ICCV, Rio de Janeiro, October 2007.

[10] J. Zieren and K. Kraiss, "Robust Person-Independent Visual Sign Language Recognition," In Proc. of the 2nd Iberian Conference on Pattern Recognition and Image Analysis IbPRIA, 2005

[11] U. Von Agris, D. Schneider, J. Zieren and K. Kraiss, "Rapid Signer Adaptation for Isolated Sign Language Recognition," Proc. of IEEE Workshop on Vision for Human-Computer Interaction (V4HCI), 2006.

[12] G. Fang, W. Gao, X. Chen, C. Wang, J. Ma, I. Wachsmuth and T. Sowa, "Signer-independent continuous sign language recognition based on SRN/HMM Gesture and Sign Language in Human-Computer Interaction," International Gesture Workshop, Springer, 2298, pp. 76-85, 2001.

[13] Sharjah City for Humanitarian Services (SCHS), website: http://www.sharjah-welcome.com/schs/about/

[14] R. Gonzalez, R. Woods and S. Eddins, "Digital image processing using Matlab," Prentice Hall, first edition, 2002.

[15] N. Ahmed, T. Natarajan, and K. R. Rao, "Discrete cosine transform," IEEE Transactions on Computers, vol. C-32, pp. 90-93, Jan. 1974.

[16] G. H. Golub and C. F. Van Loan, Matrix Computations. John Hopkins, 1989.

[17] W. M. Campbell,K. T. Assaleh, and C. C. Broun, "Speaker recognition with polynomial classifiers," *IEEE Transactions on Speech and Audio Processing* , 10(4), pp 205 - 212, 2002.

[18] K. T. Assaleh and W. M. Campbell, "Speaker Identification Using a Polynomial-based Classifier," *Proc. of the fourth International Symposium on Signal Processing and its Applications ISSPA'99*, Brisbane, Australia, August 1999.

[19] W. M. Campbell and K. T. Assaleh, "Low-Complexity Small-Vocabulary Speech Recognition for Portable Devices," *Proc. of the fourth International Symposium on Signal Processing and its Applications ISSPA'99*, Brisbane, Australia, August 1999.