Spatio-Temporal feature extraction techniques for isolated gesture recognition in Arabic sign language

T. Shanableh[1], K. Assaleh[2] and M. Al-rousan[3]
Department of Computer Science[1]
Department of Electrical Engineering[2]
American University of Sharjah[1,2]
Department of Computer Engineering[3]
Jordan University of Science and Technology[3]
tshanableh@aus.edu

Abstract

This paper presents various spatio-temporal feature extraction techniques with applications to online and offline recognition of isolated Arabic Sign Language (ArSL) gestures. The temporal features of a video-based gesture are extracted through forward, backward and bi-directional predictions. The prediction errors are thresholded and accumulated into one image that represents the sequence motion. The motion representation is then followed by spatial domain feature extractions. As such, the temporal dependencies are eliminated and the whole video sequence is represented by a few coefficients. The Gaussianity of the extracted features is assessed and its suitability for both parametric and non-parametric classification techniques is elaborated upon. The proposed feature extraction scheme was complemented by simple classification techniques, namely, KNN and Bayesian, i.e. likelihood ratio, classifiers. Experimental results showed classification performance ranging from 97% to 100% recognition rates.

To validate our proposed technique, we have conducted a series of experiments using the classical way of classifying data with temporal dependencies. Namely, Hidden Markov Models (HMMs). Experimental results revealed that the proposed feature extraction scheme combined with simple KNN or Bayesian classification yields comparable results to the classical HMM-based scheme. Moreover, since the proposed scheme compresses the motion information of an image sequence into a single image, it allows for using simple classifications techniques where the temporal dimension is eliminated. This is actually advantageous for both computational and storage requirements of the classifier.

Keywords:

Feature extraction, Visual languages, Motion analysis, Pattern classification

10.1109/TSMCB.2006.889630

1. Introduction

Deaf people use sign language as their prime means of communications with other deaf or hearing people. Developing a tool for interpreting sign language into text or speech enables hearing people to understand deaf people. The functionality of such a tool is referred to as Sign Language Recognition (SLR). The complement of such a tool is one that transforms text or speech back to sign language so as to make it possible for deaf people to understand hearing people. Interest in automatic SLR research has started in about twenty years ago particularly for American, Australian, and Japanese 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 Korean[1], Chinese[2], and to a lesser extent Arabic [3] .

Different levels of SLR have been proposed ranging from recognizing isolated alphabets [3-5] to recognizing continuous gestured sentences [6]. In terms of gesture data acquisition, SLR methods range from relying on instrumented gloves to vision based systems. Instrumented glove-based systems rely on electromechanical devices that capture the data of hand gestures via a set of motion sensors that are wired to a computer [1,7,8]. This data is then processed via signal processing and pattern matching to categorize and recognize the different gestures. Normally, such systems cause inconvenience to the signer and restrict his/her movements. On the other hand, vision-based systems capture the hand movement via video cameras in a non-restrictive way [6]. The captured video is then processed via image processing, and artificial intelligence to recognize and interpret hand gestures [9]. Some vision-based systems require that a signer wears a color-coded pair of gloves while others don't. While wearing these color-coded gloves may be inconvenient, it usually improves the recognition accuracy as it minimizes or eliminates the need for sophisticated image segmentation [5,10,11,12].

Regardless of the data acquisition method, gesture data needs to be processed, parameterized and modeled for the gesture to be recognized. Data processing ranges from simple geometric modeling of the hand and fingers for finger spelling [13] to more elaborate video and image processing techniques for full signing scenarios. For finger spelling, if it is done with instrumented gloves then the data processing needed is minimal and the focus is directed to the modeling and classification. On the other hand,

if the finger spelling is done via a vision system then some level of image processing will be needed especially if no color-coded gloves were used. For more sophisticated signing scenarios whether for limited isolated vocabulary or continuous large vocabulary more elaborate image processing techniques are required for feature extraction especially that the field of view includes the whole upper body of the signer. Extracted features need to characterize the position and the movement of the hands and sometimes the head and the face of the signer. Among the extracted features is the center-of-gravity of the hand blob relative to the face or body, relative to the first gesture frame, or relative to the previous frame [14,15,16,17] Other features characterize hand motion by tracking the hand pixels [18]. Statistical features were also used such as moment-based features [16]. Some researchers have gone beyond 2-D features by using multiple cameras to capture 3-D features of the signers' hands and other moving parts [15].

As in any recognition system, classification follows feature extraction. Once they are extracted, features have to be modeled in the training phase and recognized in the recognition phase. Many different classification methods have been used for both instrumented-gloves and vision-based SLR systems. Among these methods are hidden Markov models (HMM) [14,18], and various architectures of neural networks (NN) such as time-delay neural networks (TDNN) [20], Hopfield NN [19]. Other classifiers such as adaptive inference neuro-fuzzy systems (ANFIS) [4] and polynomial networks [5] have been used.

Finger spelling systems reported signer-independent recognition results in the 80 and 90 percentile ranges. For example, 80.1% recognition accuracy was achieved on Korean Sign Language recognition using fuzzy logic and fuzzy min-max neural networks [28]. More recently, a 93% recognition rate was obtained on signer independent Arabic sign language (ArSL) alphabet using polynomial networks [5]. For isolated word SLR previous results were reported on different international sign languages. For example, on isolated American Signal Language (ASL) words of vocabulary sizes of 28 to 40 words, recognition results were reported between 93% and 96% [16,20] using features based on pixel motion trajectories and classification based on TDNN for the 96% system. A comprehensive review of the latest advances in SLR can be found in [13].

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) has been published on ArSL. In this paper we present an isolated word ArSL system using a

variety of feature extraction and classification techniques. The introduced features are novel and, we believe, applicable to other sign languages as well.

2. Related work:

As mentioned in the introduction, no previous work has been reported on ArSL recognition of signs beyond static gestures. Our database, to the best of our knowledge, is the first of its kind. However, previous related work has been reported on other sign languages which mostly uses temporal features and HMM classification [29, 20, 6] as we have included in our experiments in Section 6. We would like to emphasize that we have, in our experimental results, used similar feature extraction and classification techniques on our database. We believe this gives an indicator to a comparison against the aforementioned work. Nonetheless, we summarize the reported related work and the obtained results as follows.

In [29] the authors proposed to extract spatial and temporal image features. The temporal features are based on the thresholded difference between two successive images. The spatial features are extracted from the skin color and edge information. A logical AND is then applied to combine the temporal and spatial features. The solution is further enhanced by applying Fourier Descriptors to extracted boundaries of hand shapes. Likewise temporal analysis is enhanced, albeit at a high computational cost, by the use of motion estimation. The temporal features are then extracted from the distribution of the magnitude and phase of the motion vectors. Hidden Markov Models (HMMs) are used for classification. Combining Fourier Descriptors with the motion analysis resulted in a classification accuracy of 93.5%. Classification based on Fourier Descriptors only resulted in a 90.5% accuracy.

In [20] feature extraction starts by breaking sentences with limited grammar into video gestures. Image segmentation is then used to segment out the hands. This task is very reasonable taking into account the cap-mounted camera pointed downwards towards the hands. The features are then extracted from the pixel-wise image differences, angle of the least inertia and the length of the associated eigenvector and lastly the ratio between the major axis the minor axis of the enclosing ellipse. Again HMMs are used for the classification. The reported classification accuracy is 91.9% for a restricted grammar.

In [6] similar Regions Of Interest (ROIs) across frames are tracked. ROIs are identified through skin color and geometric cues. Motion trajectories are then extracted from the concatenation of the affine transformations associated with these regions. Time-delay

neural networks are used for classification. The reported classification accuracy is 96.21% based on 40 ASL gestures.

3. Arabic sign language database description

Unlike other sign languages, Arabic does not yet have a standard database that can be purchased or publicly accessed. Therefore, we decided to collect our own ArSL database. We have collaborated with the Sharjah City for Humanitarian Services (SCHS) [31], UAE, and arranged for collecting data from there. In this first phase of our data collection, we have collected a database of 23 Arabic gestured words/phrases from 3 different signers. 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	عليكم السلام	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: Arabic sign language gestures and their English meanings[1].

Each signer was asked to repeat each gesture a total of 50 times over 3 different sessions resulting in a total of 150 repetitions of the 23 gestures. The signer was videotaped using an analog camcorder without imposing any restriction on clothing or image background. The video segments of each session were digitized and partitioned into short sequences representing each gesture individually. Note that the proposed feature extraction techniques to follow do not require any specific frame rate. An example of the sequence of frames of the Gesture 4 (Gift). is shown in Figure 1.

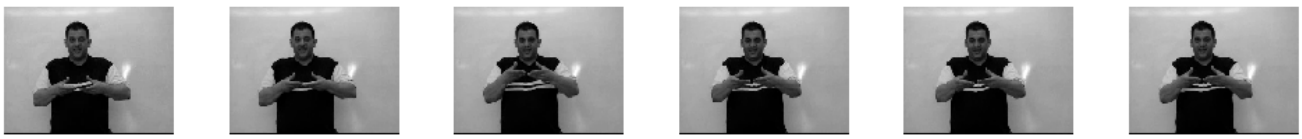


FIGURE1: Video sequence of Gesture 4 (Gift).

[1] The dataset can be made available upon request.

4. Feature extraction techniques

For video segments (i.e. image sequence) feature extraction is typically done in the temporal and spatial domains in order to capture the spatial and temporal information contents of the image sequence.

4.1. Temporal feature extractions:

The motion of the temporal sequence can be captured by removing temporal-domain redundancies. The motion can be accumulated into one image that represents the activity of the whole temporal sequence.

Temporal domain redundancy reduction techniques are well established in the video compression literature. Hybrid video compression standards employ backward and bi-directional prediction as specified by the ISO/IEC MPEG coders such as MPEG-4 part 10 [21]. On the other hand, wavelet based video coders employ sophisticated motion-compensated temporal filtering techniques as reported in [22,23].

To reduce the energy of prediction error, video coders employ motion estimation and motion compensation prediction on blocks of pixels referred to as macroblocks. The outcome of the motion estimation process is a two dimensional motion vector representing the relative displacement of a macroblock relative to a reference or anchor picture. The motion compensation prediction subtracts the macroblocks of the current picture from the best-matched location of the anchor picture as indicated by the relevant motion vector.

Content-based querying of video databases utilizes such motion vectors for video indexing and retrieval. For instance MPEG-7 visual motion descriptors use such motion information to identify object speed, motion trajectory, motion intensity, spatio-temporal distribution of motion activity and so forth [24]. The extraction of such descriptors is facilitated by the syntax of the coded video stream which includes the needed motion information. Thus neither further motion estimation nor compensation is needed in this case.

However in the proposed temporal motion activity extraction for sign language recognition, computation of motion vectors is computationally prohibitive. Thus we propose to capture the motion activity by examining the forward prediction error of successive pictures without motion compensation. Clearly, elimination of motion compensation for sign language recognition does not affect the preciseness of

classification; in fact reducing the energy of the prediction error through motion compensation prediction is undesired in this case.

Formally, the proposed forward prediction of successive frames is expressed as follows. Let $I_{g,i}^{(j)}$ denote image index 'j' of the i^{th} repetition of gesture number 'g'. The forward accumulated prediction can be computed by:

$$P_{g,i} = \sum_{j=1}^{n-1} \partial(I_{g,i}^{(j)} - I_{g,i}^{(j+1)}) \qquad (1)$$

Where n is the total number of images in the i^{th} repetition of gesture g.

∂ is a binary threshold function defined as:

$$\partial(x) = \begin{cases} 1 & if\ |x| \geq TH \\ 0 & if\ |x| < TH \end{cases} \qquad (2)$$

The TH threshold is empirically determined as illustrated in the experimental results section. Briefly, it can be set to either the mean intensity of motion pixels i.e. arithmetic mean of non-zero pixel differences, or the mean plus the standard deviation and so forth. Likewise, this section proposes the use of bi-directional accumulation of prediction errors which uses both previous and future images as a source of prediction. This can be expressed as follows:

$$B_{g,i} = \sum_{j=1}^{n} \partial(I_{g,i}^{(j)} - w_1 I_{g,i}^{(j-1)} - w_2 I_{g,i}^{(j+1)}) \qquad (3)$$

Where w_i is a bi-directional prediction weight inversely proportional to the time distance between images. Typically, equal weights of 0.5 are allocated to immediate previous and future reference images. Additionally, note that image index *0* and *n+1* are replicas of images at index *2* and *n-1* respectively. That is, two copied images are defined at the boundaries of the video sequences.

The steps of the proposed temporal feature extraction are illustration in the following diagram:

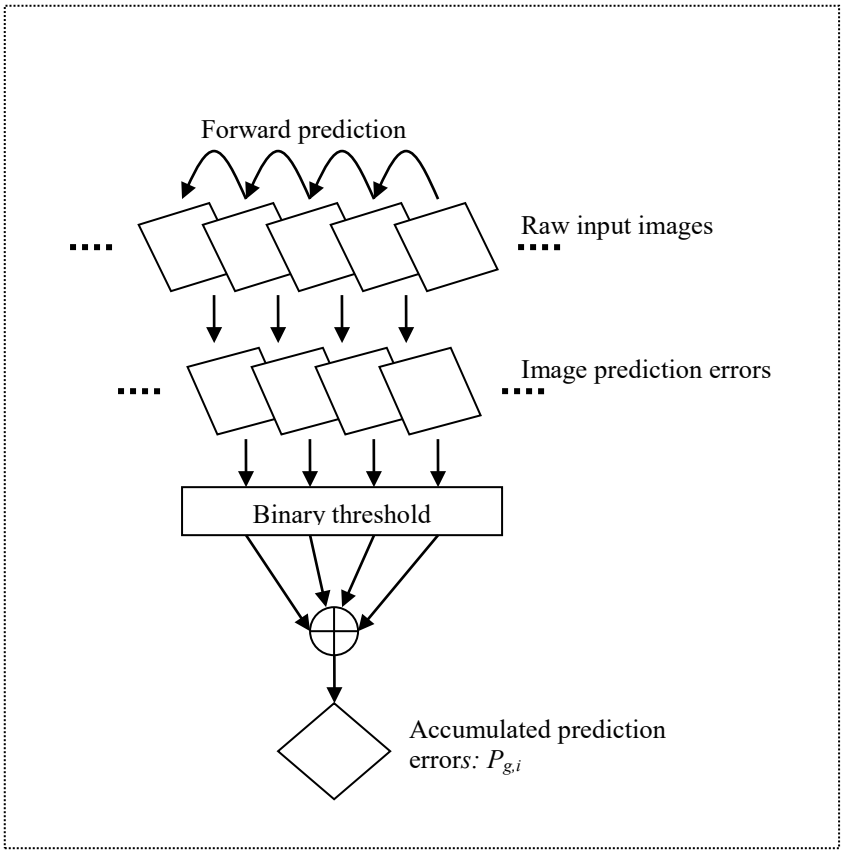
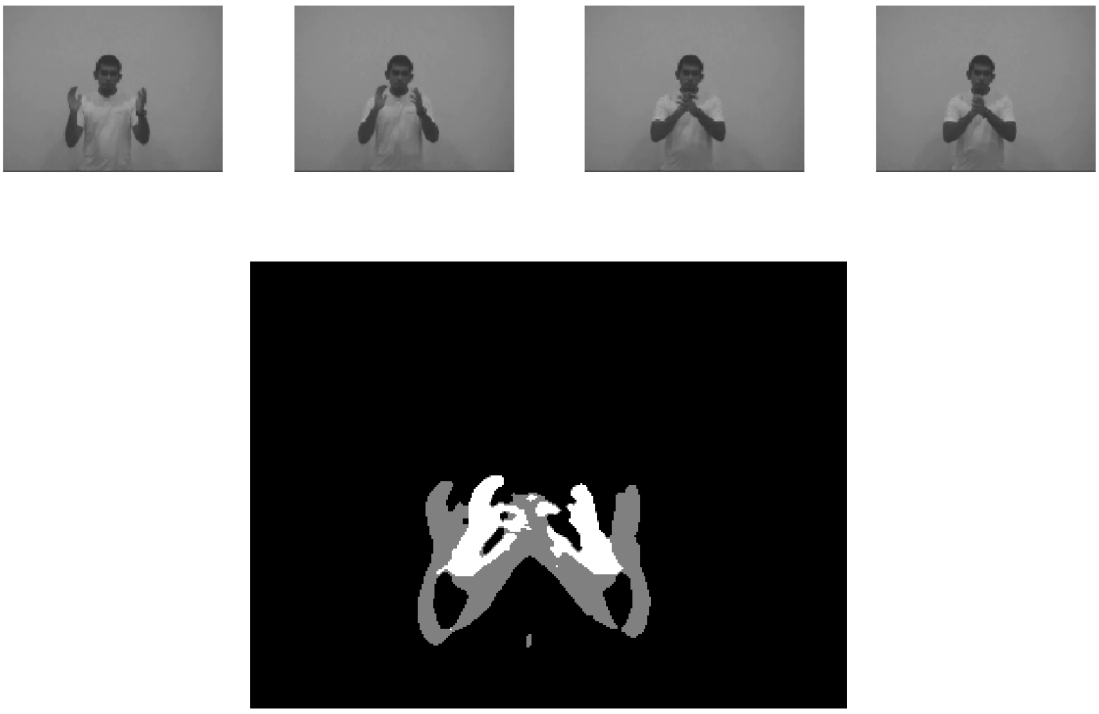


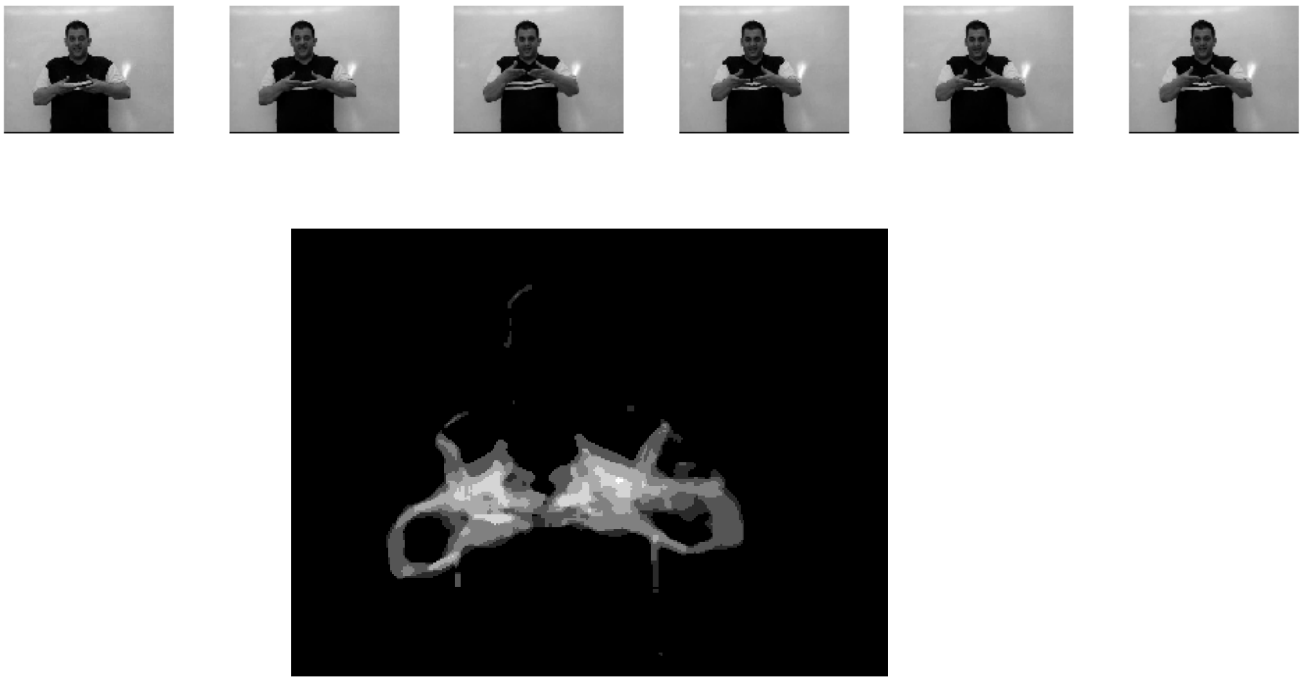
FIGURE 2. Illustration of the proposed temporal feature extraction technique.

The figure illustrates that successive images are subtracted without motion compensation. If bi-directional prediction is employed then the current image is subtracted from a weighted sum of previous and future images as mentioned previously. The resultant prediction error is thresholded into a binary image. All binary representations are then accumulated into one image which represents the temporal features of the input gesture.

Figure 3 shows an example of applying the above forward prediction technique to two different gestures.



(a)



(b)

FIGURE 3. Examples of accumulated prediction errors. Raw images and their temporal features for (a) Signer 3, Gesture 1, Repetition 1 of the test data (b) Signer 1, Gesture 4, Repetition 10 of the test data.

In the figure, the temporal features are scaled to the range of 0-255 for display purposes. Since the prediction error of successive images is subjected to a binary threshold, a median filter is applied to the temporal feature image in order to remove isolated prediction errors which can be treated as impulse noise. Additionally, the temporal feature images are normalized by the number of images in underlying image sequence. The figure shows that the motion of the image sequence is nicely captured by the proposed technique. Thus reducing the three dimensional image sequence into one image.

4.2. Spatial domain feature extractions

Having reduced the temporal sequence of a given gesture into an image of accumulated differences, we now turn our attention to spatial domain feature extractions.

This section proposes two different approaches namely: 2-D transformation followed by Zonal coding and Radon transformation followed by low pass filtering.

4.2.1 2-D transformation and Zonal coding

In the first approach, the accumulated temporal differences are transformed into the frequency domain using the 2-D Discrete Cosine Transformation (DCT) given by [30]:

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

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.

An attractive property of this transformation is its energy compaction. Low frequencies are concentrated in the top left corner of the transformed image. Thus the input accumulated temporal differences can be coarsely represented by discarding high frequencies. This can be realized through Zonal coding which was first employed for partitioning DCT coefficients into a number of video layers as specified in the MPEG-2 video codec [25]. In this work the DCT coefficients are zigzag scanned from the top left corner into an *n* dimensional vector. The dimensionality is empirically determined as illustrated in the experimental results section. The block diagram of the proposed spatial feature extraction is shown in Figure 4:

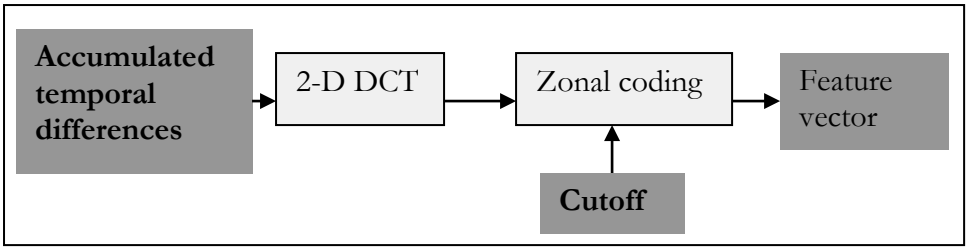


FIGURE 4. block diagram of the proposed spatial feature extraction technique.

This work also proposes the use of 2-D Walsh Hadamard (WH) transformation with the following kernel:

$$h(x,y,u,v) = \frac{1}{2^m}(-1)^{\sum_{i=0}^{m-1}\lfloor b_i(x)p_i(u)+b_i(y)p_i(v)\rfloor} \quad (5)$$

Where m is the number of bits needed to represent a pixel value, $b_i(x)$ is the i^{th} binary bit from right to left and $p_i(u) = b_{m-i}(u) + b_{m-i-1}(u)$. All sums are performed in modulo 2 arithmetic [26].

The WH transformation uses a binary transformation kernel composed of 1s and –1s. The transformation is known for its simplicity and suitability for transforming binary images. It makes more sense to decompose a binary image into square waves rather than smoothly varying sinusoids. This shall come in handy when successive input image differences are binarized without accumulation. Such an approach is needed for training HMM models as elaborated upon later.

4.2.2 Radon transformation and low pass filtering

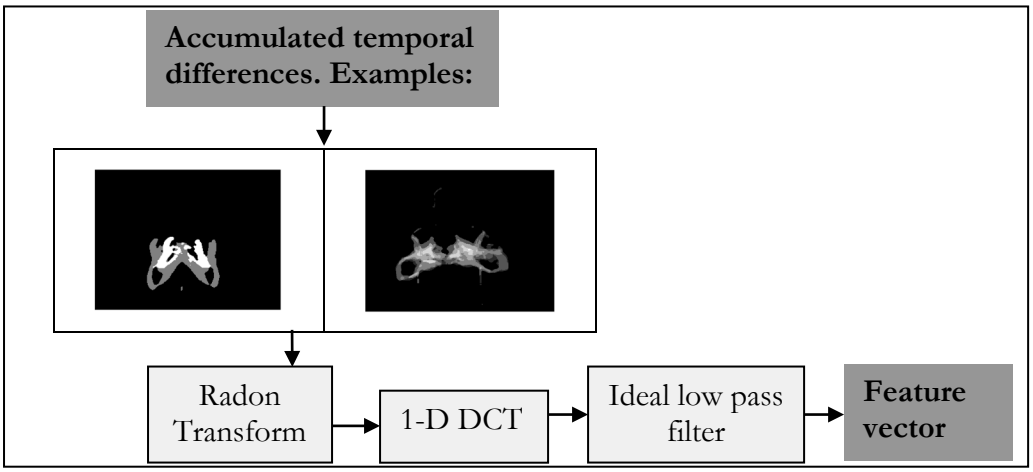
On the other hand, we also propose the use of image projections through Radon transformation. The pixel intensity of the accumulated temporal differences are projected at a given angle θ using the following equation:

$$R_\theta(x) = \int_{-\infty}^{+\infty} f(x'\cos\theta - y'\sin\theta, x'\sin\theta + y'\cos\theta)dy' \quad (6)$$

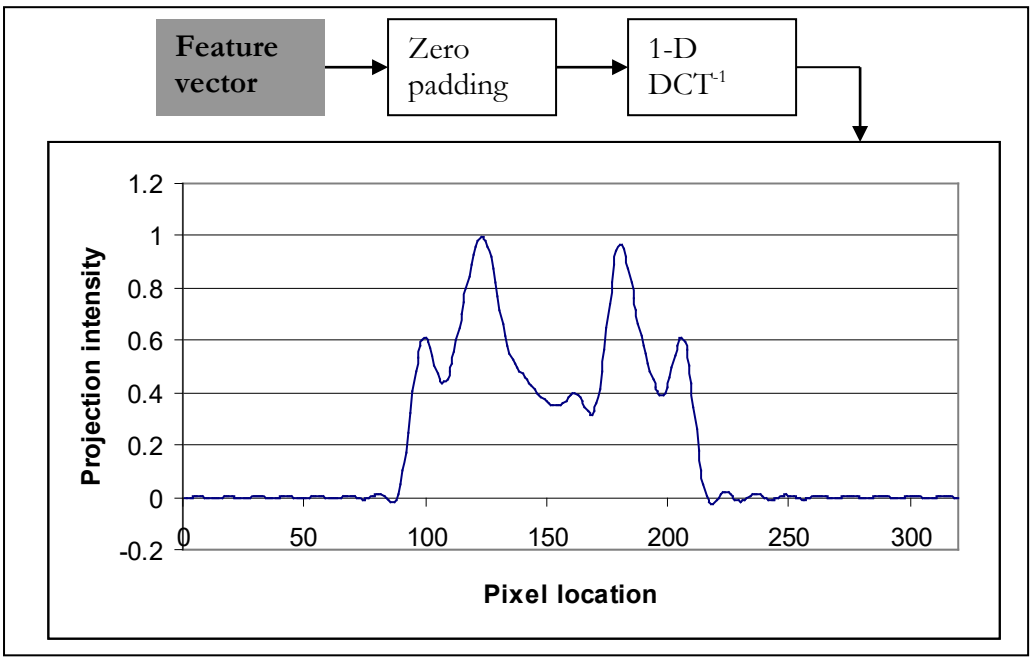
Where f is the input image and the line integral is parallel to the y' axis where x; and y' are given by:

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \quad (7)$$

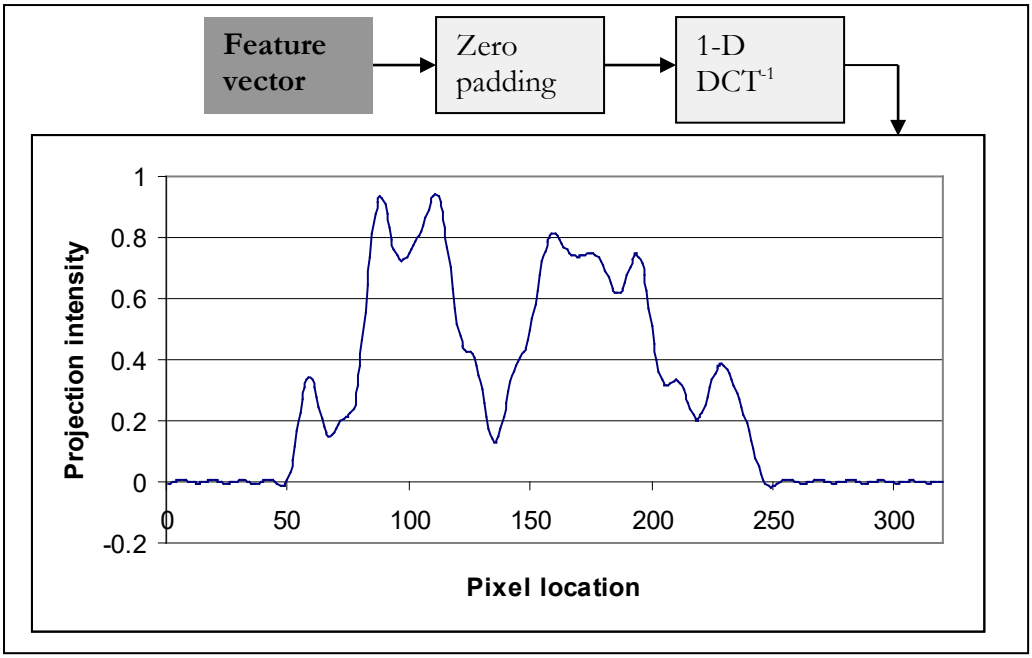
The resultant projection is then coarsely represented by transforming it into the frequency domain using a 1-D DCT followed by an ideal low pass filter. For completeness the spatial features of the accumulated temporal differences of Figure 3 above are illustrated in Figure 5:



(a) Block diagram of proposed solution.



(b) Reconstructed spatial features with the accumulated temporal differences of Figure3a.



(c) Reconstructed spatial features with the accumulated temporal differences of Figure3b.

FIGURE 5. Spatial feature extraction using image projections. Projection on the x-axis with *DCT cutoff = 50*

The projection angle in the figure is *0* degrees. The projected image are represented by the first 50 DCT coefficients in this example. The figure shows the reconstructed projection for illustration purposes. Various DCT cutoffs and projection angles are experimented with in the experimental results section.

5. Linear separability assessment of feature vectors:

To verify the suitability of the proposed feature extraction techniques, this section presents the recognition rates using linear discriminant functions such as

$$d(\mathbf{x}) = \mathbf{w}^t\mathbf{x} + w_0 \qquad (8).$$

Where **w** is a weight vector that determines the orientation of the linear decision hyperplane, w_0 is the bias and **x** is the feature vector.

Table 2 shows the results of applying the above linear discriminant function to the proposed temporal extraction followed by either the DCT zonal coding or the Radon transform approach with projections on the x-axis. For each of the 23 classes, we have used 70% of the data for training and the remaining 30% for testing.

	DCT zonal coding	Radon transform
Training error	1.49%	5.176%
Testing error	2.14%	11.95%

Table 2. Performance of linear discriminant functions for the proposed feature vector extraction techniques with 100 dimensional feature vectors.

Another linear classifier that is commonly used is the Fisher's linear discriminate. We have experimented with this classifier using the same data described above, and obtained comparable results to those of the linear discriminant functions. These results are shown in Table 3. The similarity of the results between the two classifiers is expected given that the data is linearly separable.

	DCT zonal coding	Radon transform
Training error	1.24%	4.6%
Testing error	2.82%	13.02%

Table 3. Performance of Fisher's linear discriminates for the proposed feature vector extraction techniques with 100 dimensional feature vectors.

The results obtained in the tables show that the features are reasonably linearly separable. This illustrates the efficiency of the proposed feature extraction method in projecting the complex sign videos into a compact, yet representative set of features. In the following section, this conclusion is further exhibited by the experimental results using relatively simple classification techniques such as K Nearest Neighbor (KNN).

6. Experimental results:

This section presents the experimental results for offline and online classification for both temporal and non-temporal classification. In the offline classification mode, training is done beforehand and model parameters are uploaded to the recognition stage. This is normally done when the training data is very large (due to large number of classes or excessive variability within each class) or the recognition is user-independent. On the other hand, some applications are geared towards user dependent mode with a limited number of classes. In this case training is required to be done as the user is enrolled to the system. Normally, this mode supports a limited number of classes with a few numbers of samples per class.

6.1. Offline classification:

The gesture database is divided into training and testing sets. As we mentioned in section 3, the database is composed of 50 repetitions for each of the 23 classes per signer. In this classification mode, we have used 70% of the data for training and the remaining 30% for testing. The training and testing sets contain mixed samples of all signers.

6.1.1 Bayesian and KNN based classification

This section experiments with the Bayesian i.e. likelihood ratio, and the KNN classifiers. As mentioned in section 4.1 above, the threshold of the temporal prediction error is empirically determined. Figure 6 employs a 1NN nonparametric classifier to examine three threshold values of 0, mean and one standards deviation above the mean of moving pixels.

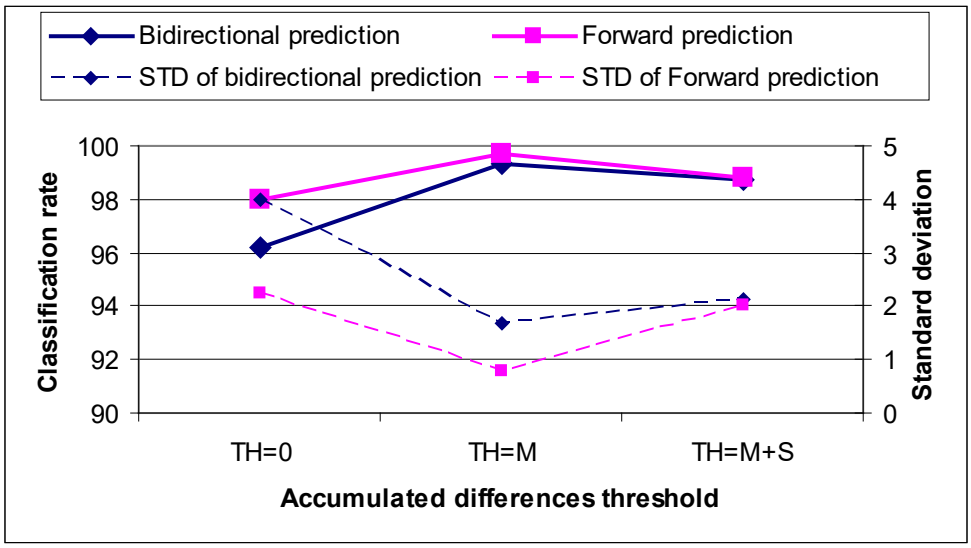


FIGURE 6. Threshold of the temporal prediction error. 1NN classifier with a DCT cutoff value of 50.

Forward and bi-directional temporal predictions are employed in the experiment. The figure shows that the forward prediction technique constantly outperforms its bi-directional counterpart. This can be justified as follows. The use of the previous and future images for prediction creates a temporally interpolated prediction that minimizes the motion residue; this defeats the purpose of motion detection.

The figure also shows that setting the threshold of the forward temporal prediction to the mean of moving pixels results in the highest classification results. Increasing the threshold results in further loss of motion information hence lower classification results.

Likewise, the DCT cutoff of the spatial domain feature extraction is empirically determined as pointed out in Section 4.2 above. Again Figure 7 employs the 1NN nonparametric classifier to examine 10 different values for the DCT cutoff.

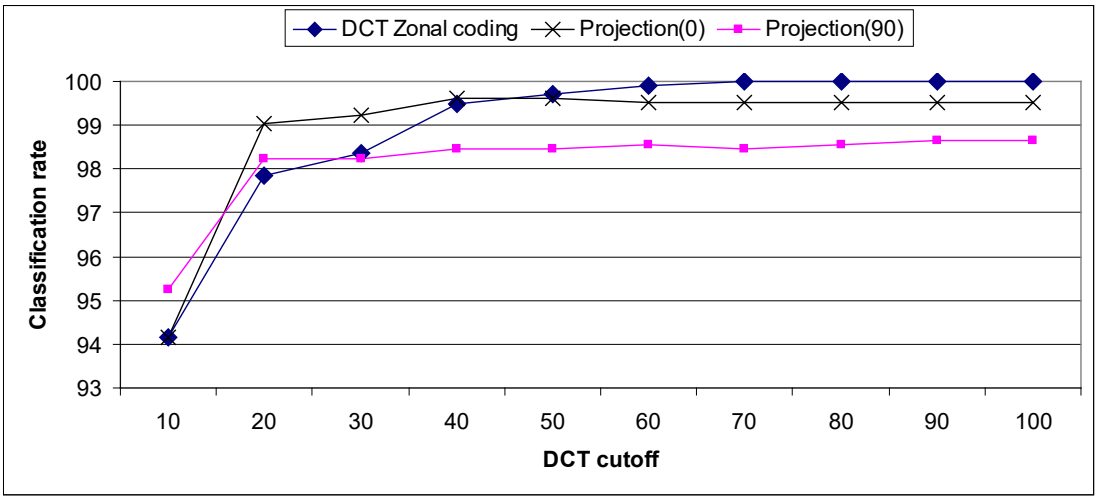


FIGURE 7. Determining DCT cutoff in spatial feature extractions. Forward temporal prediction with *TH=mean*. 1NN classifier used.

Additionally, Figure 7 experiments with 2-D DCT coding and image projections followed by 1-D DCT coding. The figure shows that coding the temporal features with 2-D DCT coding results in higher classification rates. At lower DCT cutoffs however, image projections with 1-D DCT coding are preferred. Clearly, coding 2-D DCT coefficients with such coarse zonal cutoffs under-represents the contents of the transformed temporal features, hence lower classification results. On the other hand, at higher cutoffs the 2-D DCT coding approach reaches a classification rate of 100%. This accuracy is not obtainable by the image projection technique because vertical projections result in losing vertical coordinates and vice versa. However, the vertical projections in this work scored classification results above 99% which is noteworthy.

The same experiment is repeated using a Bayesian classifier i.e. likelihood ratio, and the classification results are shown in Figure 8. The figure confirms the above discussion regarding the classification rates via 2-D DCT versus image projections. It is also important to note that the figure confirms the normality of the extracted feature vectors as indicated by the suitability of the Bayesian classifier. The figure also demonstrates that increasing the dimensionality of the feature vectors beyond a given DCT cutoff (60-70 in this case) adversely affects the classification rate. This is so because higher dimension multivariate normal distributions require more feature vectors to be parameterized

properly. This is related to the well known "curse of dimensionality problem" in parametric classifiers.

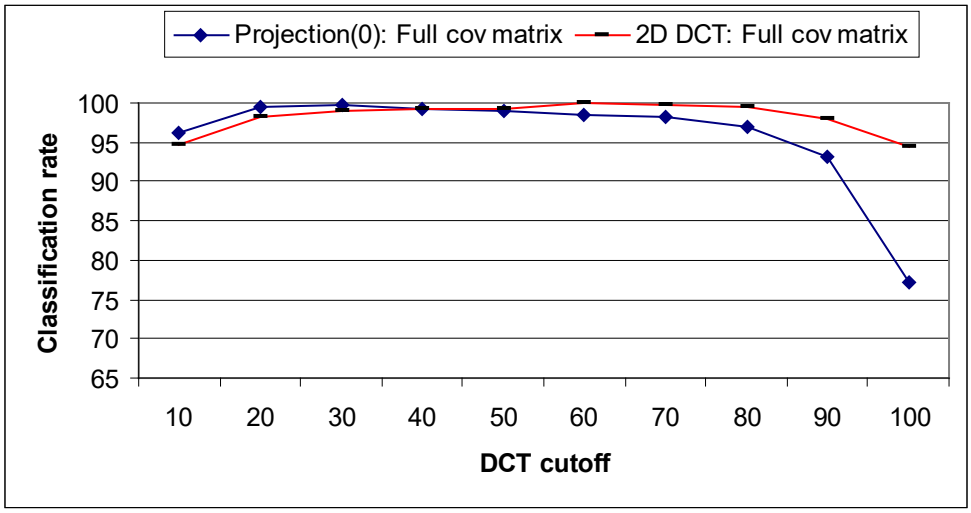


FIGURE 8. Bayesian (likelihood ratio) classification. DCT cutoff in spatial feature extractions. Forward temporal prediction with *TH=mean*.

However, this not the case when employing Fisher discriminate analysis which seems to be more robust as the DCT cutoff increases beyond 60 coefficients. This is illustrated in Figure 9 which compares the classification rates obtained by Bayesian (likelihood ratio) and Fisher linear discriminant classifiers.

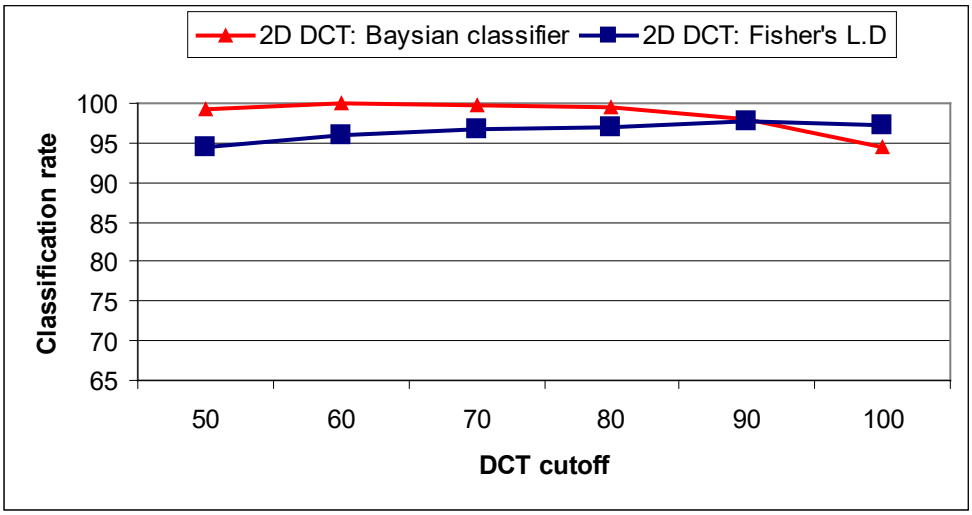


Figure 9. Comparison between Fisher's linear discriminant classification and Bayesian (likelihood ratio) classification as a function of 2-D DCT zonal cutoff.

6.1.2. HMM-based classification

This section classifies the sign language data using Hidden Markov Models (HMMs). Throughout the experiment we have used the left to right HMM architecture where a

state can only transit to its immediate right neighbor or stay in the same state. The training method applied is the Baum-Welch algorithm and the number of states is empirically determined to be 2-4 according to the complexity of the gesture.

In this approach the temporal domain information of the input image sequence is preserved. As pointed out previously, the feature extraction step preserves the absolute difference between successive images without accumulating them into one image. The absolute differences are then thresholded, binarized, transformed into the frequency domain and converted into a sequence of 1-D signals using zonal coding.

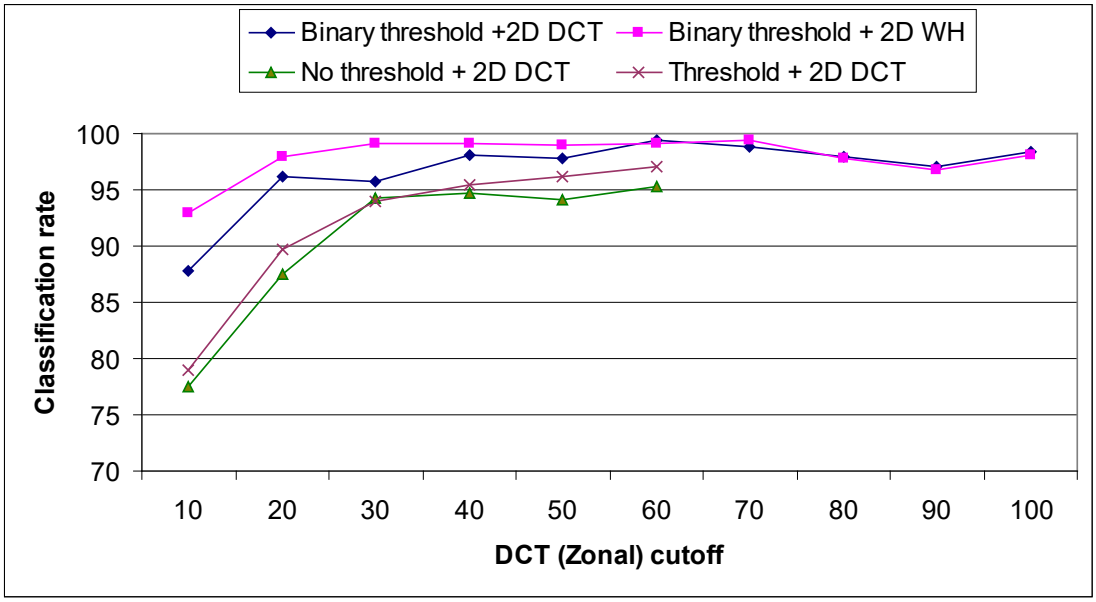


FIGURE 10. Classification using HMMs, Zonal cutoff versus classification rate. *TH=mean of moving pixels.*

In Figure 10, 4 different feature extraction approaches are employed. In the first approach successive absolute image differences are DCT zonal coded without thresholding (labeled 'No threshold + 2-D DCT'). Clearly, not all pixel difference comprise the actual gesture motion, some difference are considered noise and should not be part of the features. Therefore, this approach scored the lowest classification results as shown in the figure. In the second approach (labeled 'Threshold + 2-D DCT'), the absolute differences are thresholded, differences less than the mean of moving pixels are set to zero. This is followed by DCT zonal coding. Unlike the previous approach, noisy pixel differences are eliminated and therefore higher classification results are achieved.

Note that the modeling of both approaches did not converge when the feature vector dimensionality exceeded 60. The third approach (labeled 'Binary threshold + 2-D DCT') is similar to the aforementioned approach, however a binary threshold is employed. Thus

pixel values above the threshold are represented by the value one rather than the actual pixel difference. As such the intensity of pixel differences is decoupled from the gesture motion. The resultant binary data is then DCT zonal coded. The classification results are notably superior to the first two approaches. In the fourth and last approach (labeled 'Binary threshold + 2-D WH'), since the data is binarized it makes sense to use a binary transformation kernel rather than a weighted sum of cosine terms as provided by the DCT transformation. The transformation of choice is the Walsh-Hadamard with a transformation kernel consisting of 1s and -1s only. Such a kernel is more suitable than the smoothly varying basis vectors of DCT that resemble the intensity variation of natural images. The boost in the clarification rate is evident even at low zonal cutoffs as shown in Figure 10.

6.2 Online classification

To examine the suitability of the proposed solutions to online systems, we reduce the training data to the range of 1-6 repetitions per gesture per signer. More specifically, all the samples of a given gesture are merged into one set and divided into non-overlapping windows of 'n' samples, where 'n' is between 1 and 6 inclusive. Gestures of a signer are then classified based upon all the 'n' training samples in a round robin manner. That is, the first 'n' samples of all gestures are used for training and the resulting classification result is recorded. Then the second 'n' samples of all gestures are used for the training and so forth. All classification results are then averaged as shown in Figure 11. The mean and standard deviation of the minimum distance classifier (INN in this case) of the three signers are shown in figure.

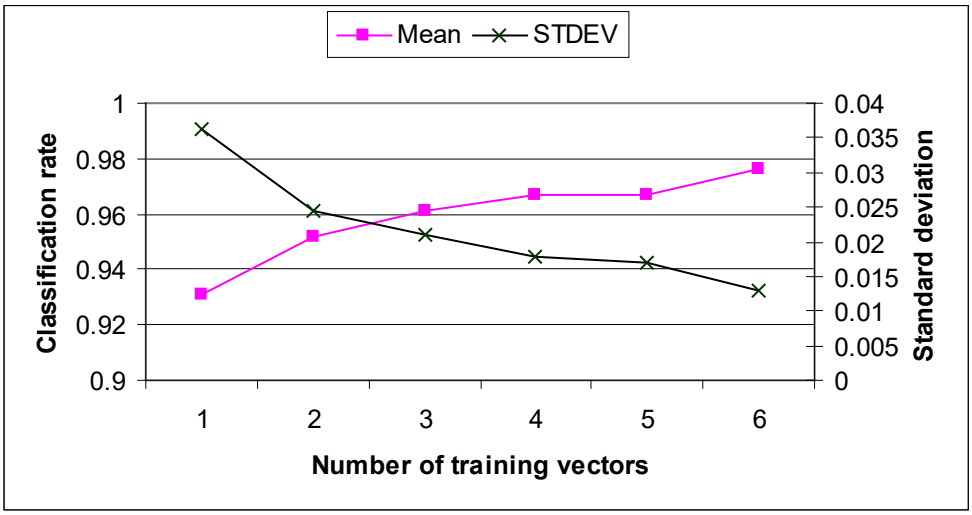
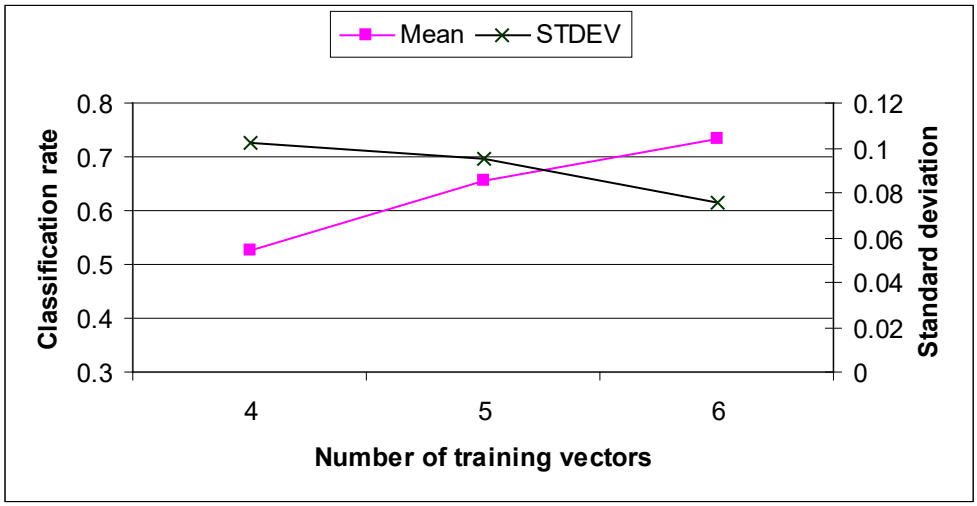
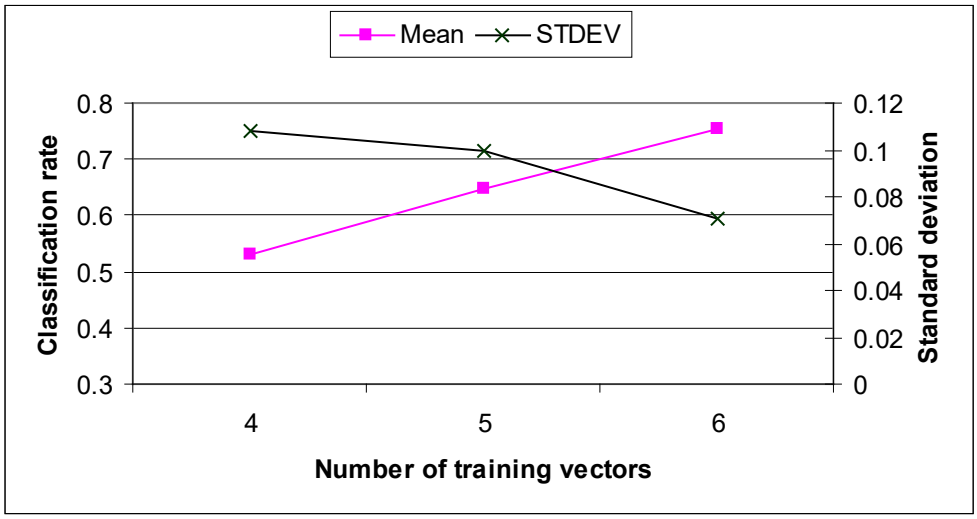


FIGURE 11. Online classification using 1NN. DCT zonal cutoff = 70, temporal *TH=mean of moving pixels*.

The figure shows that one training sample per gesture results in an average classification rate of 93%. Increasing the number of training samples to 3 results in a classification rate above 96% and so forth. On the other hand, the standard deviation of the classification rate is constantly decreasing which indicates marginal deviations from the plotted classification mean. In the worst case scenario of 1 training sample per gesture the standard deviation did not exceed 4%. Clearly, the classification accuracy at such low numbers of training samples is convenient and appreciated by users of online systems. Likewise, similar experiments are performed using the HMM classifier. The DCT zonal cutoffs are set according to the maximum classification rates of Figure 10 for both the DCT and WH transformation approaches. The classification results for the online HMM classifier are shown in Figure 12.



(a) 2-D WH with a binary threshold.



(b) 2-D DCT with a binary threshold.

FIGURE 12. Online HMM classifier.

The rather low classification results are expected in such a parametric classifier. The low number of training samples is inadequate for estimating the state transition probabilities and emissions. The figure shows that even with 6 training samples per gesture the classification rate does not exceed 75%. Increasing the number of training samples in this case defeats the purpose of online classification. Therefore the HMM classifier in this case is deemed impractical.

Lastly, it is worth mentioning that our future work will focus on recognition of continuous sign sentences. Such sentences can be segmented into isolated gestures by means of non-supervised learning for instance. Once performed, the proposed spatio-temporal solution can be applied to the segmented gestures. Another approach might be to adapt speech recognition techniques where sentences are divided into context-adaptive sub-gestures followed by Viterbi decoding and so forth. In both approaches the proposed spatio-temporal feature extraction can be applied to either the segmented sub-gestures or complete signs.

7. Conclusion

In this paper we have presented a variety of feature extraction methods for Arabic sign language recognition (ArSL). These techniques compress the motion information in a video segment into a single representative image. This was done based on the concept of temporal prediction and accumulated differences. The representative image is then

transformed into the frequency domain and parameterized into a precise and concise set of features. This process allowed for using simple classification techniques that are normally used with time-independent feature sets. To establish some comparison grounds with other classical techniques, we have conducted similar experiments using explicit temporal information with HMM classification. The proposed feature extraction scheme with KNN and Bayesian classifiers has been found to yield comparable results to the more elaborate HMM-based system.

In future work we are planning to conduct another phase of data collection where a larger number of signers will be used. As such, we will be able to focus on the user independence capability of the proposed system. Moreover, we plan to collect ArSL data for continuous signing.

Acknowledgement

The Authors acknowledge Mr. Salah Odeh of the Sharjah City for Humanitarian Services (SCHS) and Mr. W. Zouabi and F. Siam from AUS for their invaluable assistance in the facilitation of the ArSL data collection.

References

[1] 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.

[2] 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.

[3] O. Al-Jarrah and A. Halawani, "Recognition of Gestures in Arabic Sign Language Using Neuro-Fuzzy Systems," *Artificial Intelligence*, vol. 133, pp. 117-138, Dec. 2001.

[4] M. Al-Rousan and Mohammed Hussain, "Automatic Recognition of Arabic Sign Language Finger Spelling," *International Journal of Computers and Their Applications,* June 2001.

[5] 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.

[6] 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.
[7] 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," in *Proc. Int. Gesture Workshop*, 1997, pp. 273–284.
[8] S. S. Fels and G. E. Hinton, "Glove-talk: A neural network interface between a data-glove and a speech synthesizer," *IEEE Trans. Neural Networks*, vol. 4, pp. 2–8, Jan. 1993.
[9] M.W. Kadous, "Machine recognition of auslan signs using powergloves: Toward large-lexicon recognition of sign language," in *Proc. Workshop Integration Gesture Language Speech*, 1996, pp. 165–174.
[10] K. Grobel and M. Assan, "Isolated sign language recognition using hidden Markov models," in *Proc. Int. Conf. Systems, Man Cybernetics*, 1997, pp. 162–167.
[11] B. Bauer and H. Hienz, "Relevant features for video-based continuous sign language recognition," in *Proc. 4th Int. Conf. Automatic Face Gesture Recognition*, 2000, pp. 440–445.
[12] R. H. Liang and M. Ouhyoung, "A real-time continuous gesture recognition system for sign language," in *Proc. 3rd Int. Conf. Automatic Face Gesture Recognition*, 1998, pp. 558–565.
[13] C. Ong and S Ranganath, "Automatic Sign Language Analysis: A Survey and the Future beyond Lexical Meaning," *IEEE Transactions On Pattern Analysis And Machine Intelligence*, vol. 27, no. 6, pp.873-891, June 2005
[14] N. Tanibata, N. Shimada, and Y. Shirai, "Extraction of Hand Features for Recognition of Sign Language Words," *Proc. Int'l Conf. Vision Interface*, pp. 391-398, 2002.
[15] C. Vogler and D. Metaxas, "Adapting hidden Markov models for ASL recognition by using three-dimensional computer vision methods," in *Proc. IEEE Int. Conf. Systems, Man Cybernetics*, pp. 156–161, 1997
[16] Y. Cui and J. Weng, "Appearance-Based Hand Sign Recognition from Intensity Image Sequences," *Computer Vision Image Understanding*, vol. 78, no. 2, pp. 157-176, 2000.
[17] T. Kobayashi and S. Haruyama, "Partly-Hidden Markov Model and Its Application to Gesture Recognition," *Proc Int'l Conf. Acoustics, Speech and Signal Processing*, vol. 4, pp. 3081-3084, 1997.
[18] B. Bauer and K.F. Kraiss, "Video-Based Sign Recognition Using Self-Organizing Subunits," Proc. *Int'l Conf. Pattern Recognition*, vol. 2, pp. 434-437, 2002.

[10] C.-L. Huang and W.-Y. Huang, "Sign Language Recognition Using Model-Based Tracking and a 3D Hopfield Neural Network," *Machine Vision and Application*, vol. 10, pp. 292-307, 1998.

[20] M.-H. Yang, N. Ahuja, and M. Tabb, "Extraction of 2-D Motion Trajectories and Its Application to Hand Gesture Recognition," *IEEE Trans. Pattern Analysis Machine Intelligence*, vol. 24, no. 8, pp. 1061-1074, Aug. 2002.

[21] MPEG-4, "Information technology, Coding of audio-visual objects, Part 10: Advanced Video Coding," ISO/IEC 14496-10, 2005

[22] P. Chen, "Fully Scalable Subband/Wavelet Coding," PhD thesis, Rensselaer Polytechnic Institute, New York, May, 2003

[23] J.-R. Ohm, M. Schaar, J. Woods, "Interframe wavelet coding—motion picture representation for universal scalability," *Signal Processing: Image Communication*, vol. 19, no. 9, pp. 877-908, October, 2004

[24] S. Jeannin and A. Divakaran "MPEG-7 visual motion descriptors," *IEEE Transactions On Circuits And Systems For Video Technology*, vol. 11, no. 6, pp. 720-724, June 2001.

[25] MPEG-2, "Generic coding of moving pictures and associated audio information: Video" ISO/IEC 13818-2, November 1994.

[26] R. Gonzalez and R. Woods "Digital image processing," Prentice Hall, second edition, 2002.

[27] R. Duda, P. Hart and D. Stork, "Pattern classification," Wiley-Interscience; 2nd edition, October, 2000.

[28] C.-S. Lee, G. tae Park, J.-S. Kim, Z. Bien, W. Jang, and S.-K. Kim, "Real-time Recognition System of Korean Sign Language based on Elementary Components," in *Proceedings of the 6th IEEE International Conference on Fuzzy Systems*, (Barcelona, Spain), pp. 1463-1468, July 1997.

[29] 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, vol. 21, no. 8, pp. 745–758, 2003.

[30] K. R. Rao and P. Yip, "Discrete Cosine Transform: Algorithms, Advantages, Applications," Academic press, ISBN 012580203X, Aug 1990.

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