TWO TIER FEATURE EXTRACTIONS FOR RECOGNITION OF ISOLATED ARABIC SIGN LANGUAGE USING FISHER'S LINEAR DISCRIMINANTS

Tamer Shanableh[1] and Khaled Assaleh[2]
Computer Science Department[1]
Department of Electrical Engineering[2]
American University of Sharjah
Sharjah, UAE
{tshanableh or kassaleh} @aus.edu

Abstract
This paper proposes a two tier feature extraction approach for the recognition of video-based isolated Arabic sign language gestures. In the first tier, the prediction error of the image sequence is binarized and collapsed into two unidirectional *accumulated differences* images. In the second tier of feature extractions, two approaches are applied to the accumulated differences images: frequency domain transformation, and radon transformation. We apply such feature extractions on each of the *accumulated differences* images and then concatenate the resultant feature vectors. Alternatively, the *accumulated differences* images are concatenated prior to the second tier of feature extractions. The paper reports on the classification results of both solutions using Fisher's linear discriminants. Comparisons with existing work reveal that up to 39% of the misclassifications have been corrected.
Keywords: Pattern recognition, Image motion analysis, video signal processing.

1. Introduction

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 sign language recognition research. To date, only small number of research papers has been published on ArSL. Signer independent recognition of Arabic sign language alphabet using polynomial networks was reported in [1]. More recently, the authors reported the recognition of isolated gestures [2]. This work extends the latter application by introducing polar Accumulated Differences (ADs) as explained in Section 3.

Related work on recognition of non-Arabic using temporal-domain feature extractions mainly reply on computationally expensive motion analysis approaches such as motion estimation. Additionally since the time domain is preserved, Hidden Markov Models (HMM) are usually used for classification. For instance in [3] 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. Combining Fourier Descriptors with the motion analysis using HMM classifier resulted in a classification accuracy of 93.5%. Classification based on Fourier Descriptors only resulted in a 90.5% accuracy. In [4], 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 [5] similar regions of interest across frame are tracked. ROIs are identified thru 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. This paper is organized as follows. Section2 describes the dataset used in this study. Various features extraction schemes are described in section 3. In section 4, we present the experimental results of this study. Finally, a conclusion is given in section 5.

2. Arabic sign language database description

As the authors reported in [2], Arabic Sign Language 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) [6], UAE, and arranged for collecting ArSL data. 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	صـــديق	Friend	13	يأكل	To Eat
2	جار	Neighbor	14	ينام	To sleep
3	ضـــيف	Guest	15	يشـــرب	To Drink
4	هدية	Gift	16	يســـــتيقظ	To wake up
5	عدو	Enemy	17	يســـمع	To listen
6	عليكـــم السـلام	Peace upon you	18	يســـكت	To stop talking
7	اهلا وسهلا	Welcome	19	يشـــم	To smell
8	شـــكرا	Thank you	20	يســـاعد	To help
9	تفضـــــل	Come in	21	امس	Yesterday
10	عيـب	Shame	22	يدهب	To go
11	بيـــت	House	23	يأتي	To come
12	انـا	I/me			

Table 1: Arabic sign language gestures and their English meanings.

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; a total of 3450 video segments. 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 do not require any specific frame rate. An example of the sequence of frames of the Gesture 3 (Guest) is shown in Figure 1 part a.

3. Feature extractions

This section introduces the use of polar ADs in the first tier of feature extractions. The section also reviews two solutions for a second tier of feature extractions. Lastly, we propose a two tier feature extraction solution that combines the aforementioned solutions.

3.1 First tier of feature extractions:

The first tier of feature extractions extracts motion information from the temporal domain of the input image sequence. For instance, successive image differencing is an approach that compares images on pixel basis to detect the motion information. Let $I_{g,i}^{(j)}$ denote image index j of the i^{th} repetition of letter index g. The ADs image can be computed by:

$$AD_{g,i} = \sum_{j=1}^{n-1} \partial(I_{g,i}^{(j)} - I_{g,i}^{(j-1)}) \quad (1)$$

Where n is the total number of images in the i^{th} repetition of letter at index g. ∂ is a binary threshold function defined as:

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

The TH can be set to the mean intensity of motion pixels or the mean plus the standard deviation and so forth. Polar ADs of an image sequence preserves the directionality of hand movements. Such differences can be categorized into three types: Absolute (|AD|), Positive (AD_+) and Negative (AD_-). Thus the name polar ADs. Formally polar ADs are described as:

$$|AD|(x,y) = \begin{cases} AD+1 & if\ \ |f(x,y,t_k) - f(x,y,t_{k-1})| \geq Th_{(k,k-1)} \\ AD & otherwise \end{cases} \quad (3)$$

$$AD_+(x,y) = \begin{cases} AD_+ +1 & if\ \ (f(x,y,t_k) - f(x,y,t_{k-1})) \geq Th_{(k,k-1)} \\ AD_+ & otherwise \end{cases} \quad (4)$$

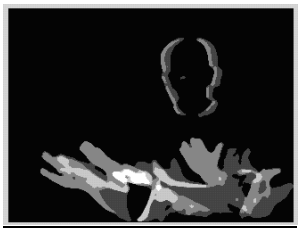
$$AD_-(x,y) = \begin{cases} AD_- +1 & if\ \ (f(x,y,t_k) - f(x,y,t_{k-1})) \leq -Th_{(k,k-1)} \\ AD_- & otherwise \end{cases} \quad (5)$$

Where (x,y) are the pixel coordinates of the ADs image. The absolute ADs approach was proposed for sign language recognition by the authors in [2]. Here we extend this work by experimenting with polar ADs as well. Note that the latter ADs have been successfully used in the recognition of video-based recognition of Arabic handwritten alphabets as reported by the authors in [7].

Figure 1 shows an example of applying the above ADs approaches to gesture 3 of Table 1 above.

(a) Input image sequence, gesture 3 (Guest).

(b) Polar ADs images.



(c) Absolute ADs image.

Figure 1. Examples of ADs images.

3.2 Second tier of feature extractions

Once the ADs images are computed, a second tier of feature extractions is applied. Two different approaches are employed: a) 2-D Discrete Cosine Transformation (DCT) followed by Zonal coding, and b) Radon transformation followed by low pass filtering.

An attractive property of the DCT transformation is its energy compaction. Thus, the input ADs can be represented by Zonal coding of the DCT coefficients via a zigzag scanned path into an *n*-dimensional vector [8]. The dimensionality is empirically determined as illustrated upon in section 4.

On the other hand, we also experiment with image projections through Radon transformation. The pixel intensities of the ADs 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.

3.3 Two tier feature extractions

The aforementioned first and second tiers of feature extraction techniques are merged using one of the following approaches as described in subsections 3.3.1 and 3.3.2.

3.3.1 Polar Accumulated Differences

In this approach, the positive and negative ADs are concatenated together into one image prior to the second tier of feature extractions as shown in Figure 2.

In case of 2-D DCT, the transformed image is Zonal coded with different cutoffs. On the other hand, if radon transformation is applied then the projected image is 1-D DCT transformed followed by ideal low pass filtering. This approach is illustrated in Figure 2 below.

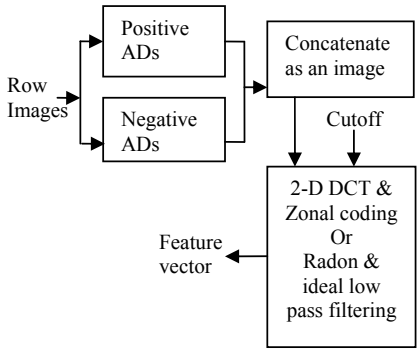


Figure 2. Polar accumulated differences.

3.3.2 Vectorized Polar Accumulated Differences

Here the positive and negative ADs are computed. A second tier of feature extractions is then applied to each of the ADs images. The concatenation is thereafter applied to the resultant feature vectors. This approach is illustrated in Figure 3 below.

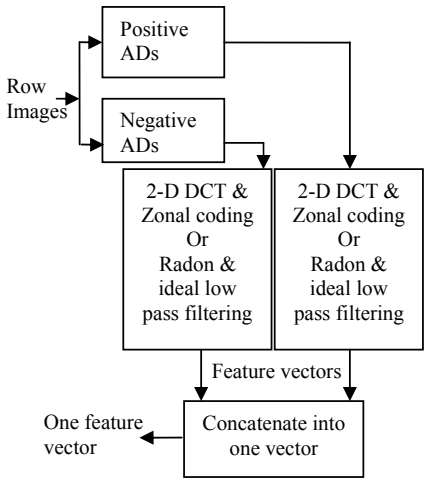


Figure 3: Vectorized accumulated differences with 2-D transformation.

4. Experimental results

This section presents the experimental results for the various feature extraction schemes described above. Training is done in an offline mode and model parameters are uploaded to the recognition stage. Offline training mode is normally done when the training data is large (due to large number of classes or excessive variability within each class) or the recognition is in user-independent mode. The gesture database is divided into training and testing sets. As we mentioned in section 2, 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.

In the following classification experiments, Fisher's linear discrimination is employed. The proposed polar ADs approaches are compared against the work reported in [2] (thereafter referred to as 'Absolute ADs'). In Figure 4, 2-D transformations and Zonal coding are used for the second tier of feature extractions. The proposed vectorized ADs of Figure 3 above outperforms the absolute ADs. The figure also shows that results of concatenating the positive and negative ADs images prior to the second tier of feature extraction (as proposed in Figure 2 above) is comparable to the results of absolute ADs up to a DCT cutoff of 90 coefficients. In all cases the figure shows that a cutoff of 90 coefficients minimizes the classification error rate.

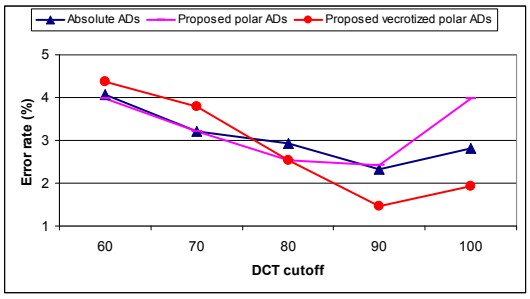


Figure 4. Fisher's linear discrimination with 2-D transformation and Zonal coding.

On the other hand, the classification gain of the proposed solution is more pronounced with Radon transformation and ideal low pass filtering. Figure 5 shows that both approaches of concatenating ADs images and concatenating the feature vectors outperform the Absolute ADs for all values of DCT cutoff. For instance, at a cutoff of 60 the misclassifications is reduced by 39.4%. The figure also shows that the proposed polar ADs approach maintains stable linear separability even at low DCT cutoffs.

5. Conclusion

This work proposed a two tier feature extraction approach for the recognition of video-based Arabic Sign language gestures. The proposed approach is combined with a second round of feature extractions in two different arrangements. In one, the positive and negative ADs images are concatenated prior to a second tier of feature extractions. In the other, the second tier of feature extractions are applied to each of the ADs images and the resultant feature vectors are concatenated into one concise representation. Experimental results revealed the linear separability of the proposed approaches and illustrated a boost in classification results when compared to existing work.

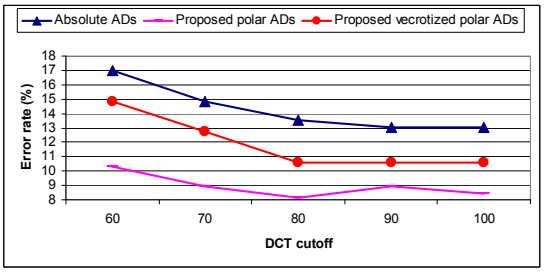


Figure 5. Fisher's linear discrimination with vertical Radon transformation and ideal low pass filtering.

Acknowledgments

We acknowledge the help of Feras Siam, Wasseim Al Zouabi, and Salah Odeh from Humanitarian City for data collection. The authors would also like to thank the American University of Sharjah for a research grant in support of this work (2006-2007).

References

[1] 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.

[2] T. Shanableh, A. Assaleh and M. Al-Rousan, "Spatio-Temporal feature extraction techniques for isolated gesture recognition in Arabic sign language ," IEEE Transactions on Systems, Man and Cybernetics, Part B (to appear).

[3] 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.

[4] 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.

[5] 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.

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

[7] K. Assaleh, T. Shanableh and H. Hajjaj, "Online Video-Based Handwritten Arabic Alphabet Recognition," Third AUS International Symposium on Mechatronics, AUS-ISM06, Sharjah, UAE, April 2006.

[8] C. Gonzalez and R.E. Woods, "Digital Image Processing," 2nd Edition, Prentice Hall, 2002.