Journal of the ACS, Vol. 7, May 2013

Arabic Sign language Recognition using Radial Signature and Dynamic Time Warping

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

Ramzy Karam
Modern Academy,
Cairo, Egypt
Ramzy.karam92@Gmail.com

Abstract
Deaf communities face daily troubles communicating with others in society. A link between the communication with sign language and natural language is needed to facilitate the life of the deaf community. This paper, proposes a methodology based on image processing and dynamic time warping to translate the deaf signs to natural language. The technique based on radial signature as a feature extraction methodology and the dynamic time warping for the classification of features vectors. The experimental results of the proposed algorithm indicated eighty percent of real-time recognition on sets of 5 letters.

Keywords: Human Machine Interaction, Sign Language Recognition, Shape-Based Gesture Recognition, Radial Signature, Dynamic Time Warping

I. INTRODUCTION

Human-computer interfaces (HCI) evolved from text based interfaces through 2D graphical-based interfaces, multimedia-supported interfaces, to fully-fledged multi-participant Virtual Environment (VE) systems. While providing a new sophisticated paradigm for human communication, interaction, learning and training, VE systems also provide new challenges since they include many new types of representation and interaction [1].
Understanding sign language and gestures using computer programs is an active researched topic [1][5][12].
Researchers in the field of recognition, computer intelligence and machine intelligence provided several approaches to provide solutions to the problem of sign language recognition [13]. The developed techniques includes using sensors on the hands like cyber-gloves, image processing using restricted and unrestricted environment, and Neural networks based [5][12][13][14].

The image processing based methods in unrestricted environment is considered a complex task due the environment complexity. Environment complexity like skin-tone ranges, lightening effect, dynamic changes in the environment with motion both in lightening effects and objects positions, complex background elements, commonalties between face skin-tones and body skin-tones. Moreover, orientation, scaling, and the probability of mix with background objects. However a robust technique for extracting the hand required for the consequent steps to go with acceptable probability of success which is complicated as well.
The proposed approach uses a calibrated skin tone HSV values from the environment as a base for hand extraction. The extracted assumed hand passes through simple acceptance criteria to consider as a hand for further processing. The hand object goes through feature extraction and followed by the recognition algorithm to select the proper intended letter.
The rest of this paper is organized as following: Section two related work survey, section three is an overview of dynamic time warping, section four the preprocessing and feature extraction, section five is the recognition with dynamic time warping, section five includes tests and results, and finally, section six is the conclusion and future work.

II. RELATED WORK

Machine learning algorithms have been applied successfully to fields of research like, face recognition [4], automatic recognition of a musical gesture [5], classification of robotic soccer formations [6], classifying human physical activity from on-body accelerometers [7], and automatic road-sign detection [8, 9]. One simple classifier is K-Nearest Neighbor which was used in [4, 6]. This classifier represents each feature vector as a data in d–dimensional space, where d is the number of attributes/features extracted. The algorithm selects a set of representative using the training dataset. The computed proximity from the representatives' of the classes considered the measure of similarity or dissimilarity. In distance calculation, standard Euclidean distance is normally used, however other metrics can be used [10]. An artificial neural network is a mathematical/computational model that attempts to simulate the structure of biological neural systems. They accept features as inputs and produce decisions as outputs [14]. Maung et al [6, 9, and 12] used it in a gesture recognition system, Faria et al

[6] used it for the classification of robotic soccer formations, Vicen-Buéno [9] used in the problem of traffic sign recognition and Stephan et al used for static hand gesture recognition for human-computer interaction. Support Vector Machines (SVM's) is a technique based on statistical learning theory, which works well with high-dimensional data. The objective of this algorithm is to find the optimal separating hyper plane between classes by maximizing the margin between them. Faria et al. [4, 6] used to classify robotic soccer formations and the classification of facial expressions. Maldonado-Báscon [8] used for the recognition of road signs and Masaki et al used it in conjunction with SOM (Self-Organizing Map) for the automatic learning of a gesture recognition mode. Some examples feature extraction techniques which are used with neural network like hand-palm orientation and hand shapes [13], Orientation Histogram [14], Hu moments [15] [16]

Also, an early vision based system for the recognition of sign language is presented in [17]. Starner and Pentland use a single video camera and uniformly colored gloves to aid the segmentation and the feature extraction processes. Later they also showed that a user-calibrated skin color model delivers similar results in a known environment. Hienz et. al [18] use color coded gloves which allows detailed information about each finger of the hand. The environment is restricted to an empty white background. In arbitrary environments however, neither skin color nor any other color can be guaranteed to appear only within the object of interest, which is the hand. Thus, relying on color information only is not sufficient, not even with the aid of colored gloves.

A comprehensive study for Arabic Sign Language, ASL, based on pulse coupled neural network in [21-25]. In this study snapshots from two video cameras at different angles from the signer are used to provide images of resolution 160 X 120X24. The features extracted using the pulse coupled neural network, which are invariant to translation, rotation, and scaling, from the two sources are combined using weighting that depend on signature quality of each. The study also includes the use of natural language processing to aid the recognition process. The study was comprehended to the extend it left only two aspects of the problem: facial expression and non-uniform lightening environment.

III. DYNAMIC TIME WARPING

Dynamic time warping (DTW) is a well-known technique to find an optimal alignment between two given (time-dependent) sequences. Intuitively, the sequences are warped in a nonlinear fashion to match each other's [19][21]. Originally, DTW has been used to compare different speech patterns in automatic speech recognition. In fields such as data mining and information retrieval, DTW has been successfully applied to automatically cope with time deformations and different speeds associated with time-dependent data. The objective of DTW is to compare two (time-dependent) sequences $X := (x_1, x_2, \dots, x_N)$ of length $N \in \mathbb{N}$ and $Y := (y_1, y_2, \dots, y_M)$ of length $M \in \mathbb{N}$.

To compare two different features x, y ∈ F, one needs a local cost measure, sometimes also referred to as local distance measure, which is defined to be a function $c : F \times F \to R \geq 0$.

Typically, $c(x, y)$ is small (low cost) if x and y are similar to each other, and otherwise $c(x, y)$ is large (high cost). Evaluating the local cost measure for each pair of elements of the sequences X and Y, one obtains the cost matrix $C \in \mathbb{R}^{N \times M}$ defined by $C(n, m) := c(x_n, y_m)$. Then the goal is to find an alignment between X and Y having minimal overall cost. Intuitively, such an optimal alignment runs along a "valley" of low cost within the cost matrix C, for an illustration. The next definition formalizes the notion of an alignment.

An (N, M) −warping path (or simply referred to as warping path if N and M are clear from the context) is a sequence $p = (p_1, \dots, p_L)$ with $p_l = (n_l, m_l) \in [1{:}N] \times [1{:}M]$ $for\ l \in [1{:}L]$ satisfying the following conditions

1. Boundary condition: $p_1 = (1,1)$ and $p_L = (N, M)$.
2. Monotonicity condition: $n_1 \leq n_2 \leq \cdots \leq n_L$ and $m_1 \leq m_2 \leq \cdots \leq m_L$
3. Step size Condition: $p_{l+1} - p_l \in \{(1,0), (0,1), (1,1)\}$ $for\ l \in [1 : L - 1]$

The $total\ cost\ c_p\ (X, Y)$ of a warping path *p* between *X* and *Y* with respect to the local cost measure *c* is defined as

$$c_p(X,Y) := \sum_{l=1}^{L} c(x_{n_l}, y_{m_l})$$

Furthermore, an *optimal warping path* between X and Y is a warping path p^*having minimal total cost among all possible warping paths. The *DTW distance* DTW(X, Y) between X and Y is then defined as the total cost of $p*$:

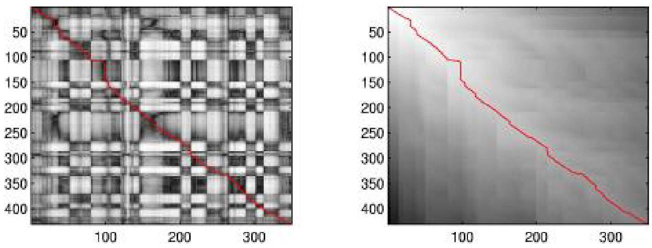
$$DTW(X,\ Y) := c_{p*}(X,\ Y)$$
$$= min\{cp(X,\ Y)\,|\, p\ is\ an\ (N,M) - warping\ path\}$$

To determine an optimal path $p*$, one could test every possible warping path between X and Y. Such a procedure, would lead to a computational complexity that is exponential in the lengths N and M.

Another way to reduce the redundant paths is calculating the Accumulated Cost Matrix

The accumulated cost matrix D satisfies the following identities:

- $D(n,\ 1) = \sum_{k=1}^{n} c(x_k, y_1)\ for\ n \in [1:N]$
- $D(1,\ m) = \sum_{k=1}^{m} c(x_1, y_k)\ for\ m \in [1:M]$
- $D(n,m) = min\{D(n-1,\ m-1), D(n-1,\ m), D(n,\ m-1)\} + c(x_n, y_m)\ for\ 1 < n \leq N\ and\ 1 < m \leq M$



(a) (b)

Figure (1) (a)Local Cost Matrix, (b) accumulated Cost Matrix

The optimal warping path $p^* = (p_1\ ,\ \ldots,\ p_L)$ can be computed in reverse order of the indices starting with $p_L = (N,M)$ using the accumulated cost matrix

Suppose that $p_l = (n,m)$ has been computed.

In case $(n,m) = (1,1)$, one must have $l = 1$ and we are finished. Otherwise,

$$p_{l-1} := \begin{cases} (1,\ m-1), & if\ n = 1 \\ (n-1,\ 1), & if\ m = 1 \\ argmin \begin{Bmatrix} D(n-1, m-1) \\ , D(n-1,\ m), \\ D(n,\ m-1) \end{Bmatrix}, & otherwise \end{cases}$$

IV. PRE-PROCESSING AND FEATURE EXTRACTION

Hand segmentation and feature extraction is a crucial step in computer vision applications for hand gesture recognition. The pre-processing stage prepares the input image and extracts features used later with the classification algorithms. Failure in this step means consequent wrong decisions will be made.

In this study, the hand extract from frames using HSV skin tone filtration. The filtration uses a calibrated HSV values from the user in the environment at the beginning of the program and once per session. The extracted assumed hand object(s) by HSV threshold maps the frame to a binary image that represents the hand. Figure (2) presents samples of extracted hand objects. The extracted objects go through filtration based on size and stability in consequent frames. The features then extracted from the images that passes the selection criteria.

Figure (2) Binary images extracted for our main focus signs,
Most left, sign of five “open palm”, middle, letter nūn ن , most right, letter ṣād ص

This process used twice in learning letters, and recognizing letters. In the learning process several shots for the same letter is used to find out a representative for the letter. In the recognition phase the extracted features go through the time warping process to select the best match.

The extracted hand object then goes through the feature extraction step. The feature extraction used in the study is the Radial Signature [3]. The base image used is set to size 320 x 240 pixels. The radial reference point used is the center point defined as:-

$$CenterPoint = \left(\frac{x_{max} - x_{min}}{2}, \frac{y_{max} - y_{min}}{2}\right)$$

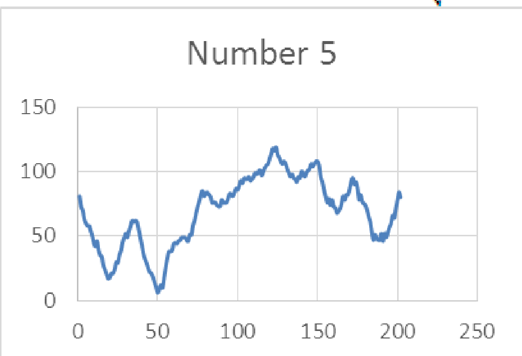
The extracted hand objects of contour over 400 pixels e down sampled to 400 points. A 100 equally spaced radial measures relative to the center is taken to form the feature vector. Figure (3) is samples of the radial contour for different letters.

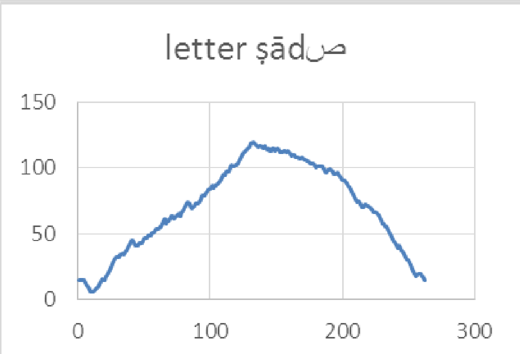
Get positional vector relative to the screen $(0,0)$ which is the upper left corner

$$Positional\ Vector = ((x_{contour} - x_{center}), (y_{contour} - y_{center}))$$

Get the Magnitude of each point on the contour

$$Value\ of\ Point\ on\ Contour = \sqrt[2]{(x_{positional})^2 + (y_{positional})^2}$$





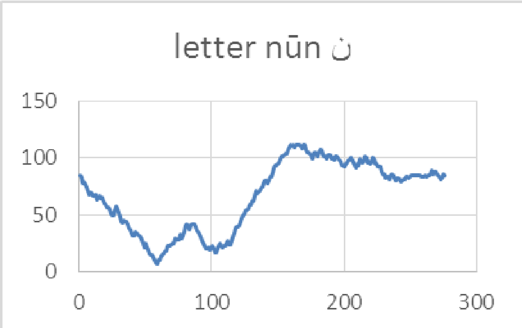


Figure (3) Examples of radial contour signature

V. Applying Dynamic Time Warping

During the learning phase the Radial signature from training images dataset categorized by char are save. These categorized radial signatures are recalled during the recognition phase to compute the cost matrices according to the Dynamic Time Warping Algorithm.

For two radial signatures,

One for sign five $F = \{ f_1, f_2, \dots, f_N \}$, the other for sign ṣād $S = \{ s_1, s_2, \dots, s_M \}$. Then we calculate cost matrix C, where $C \in \mathbb{R}^{N \times M}$ matrix, Where each cell c equals $c(n,m) = |f_n - s_m|$

In the letter recognition phase, the Accumulated Cost Matrix computed as following.

- $D(n,\ 1) = \sum_{k=1}^{n} c(x_k, y_1)\ for\ n \in [1:N]$
- $D(1,\ m) = \sum_{k=1}^{m} c(x_1, y_k)\ for\ m \in [1:M]$
- $D(n,m) = \min\{D(n-1,\ m-1), D(n-1,\ m), D(n,\ m-1)\} + c(x_n, y_m)$
- $for\ 1 < n \le N\ and\ 1 < m \le M$

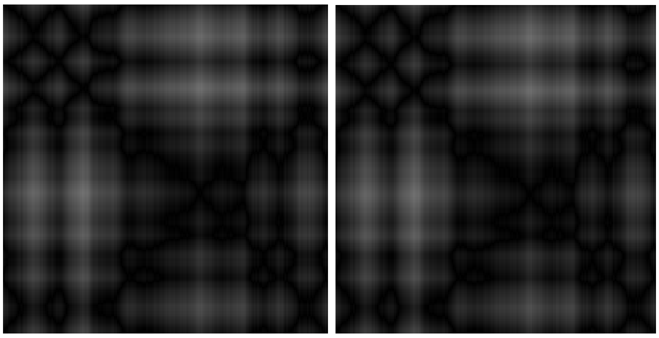
The best warping path for all accumulated matrices is selected. Other approach is found to be effective which is the count of black pixels in the accumulated cost matrices, and then pick the letter of matrix with the biggest count.

TESTS AND RESULTS

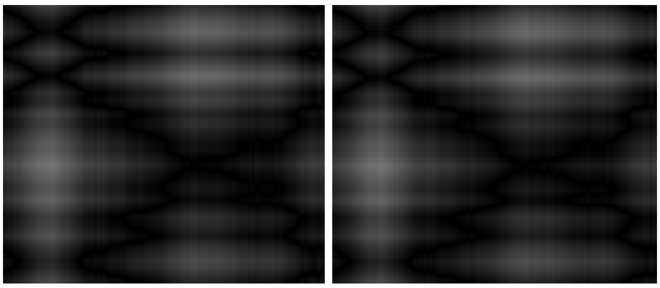
In testing the proposed algorithm, a sets of four letters is used. Samples of the detailed process in the following:-

First comparing sign of five “open palm” from frame with sign of five “open palm”, letter nūn ن, and letter ṣād ص from Database.

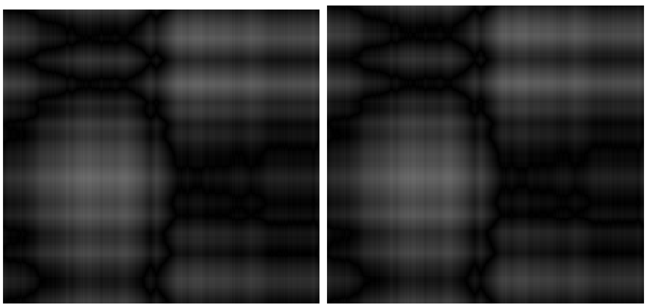
1) Sign of Five “open palm”



2) letter ṣādص

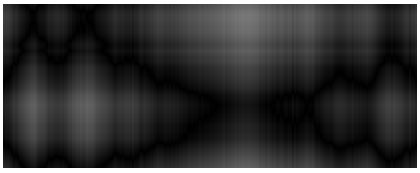


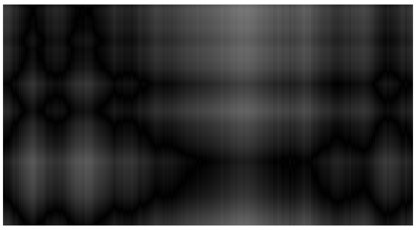
3) letter nūn ن



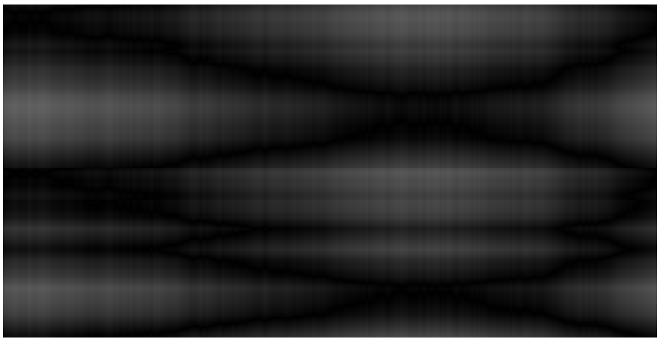
Second comparing sign of letter ṣād ص from frame with sign of five “open palm”, letter nūn ن, and letter ṣād ص from Database.

1) Sign of Five “open palm”

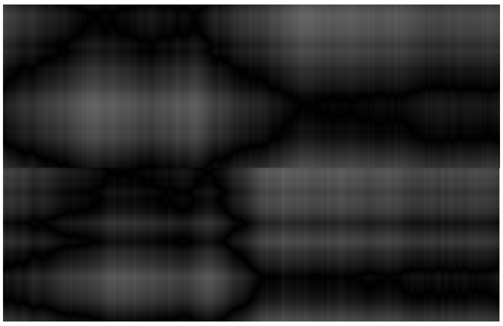




2) letter ṣādص

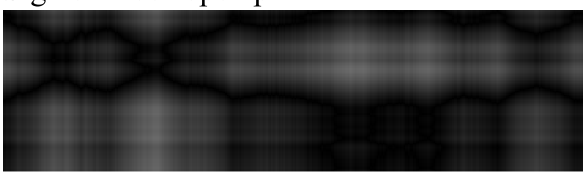


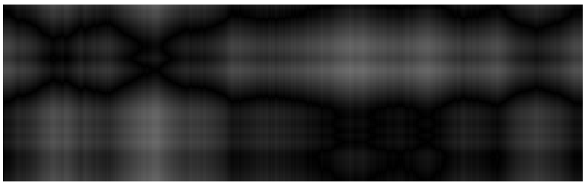
3) letter nūn ن



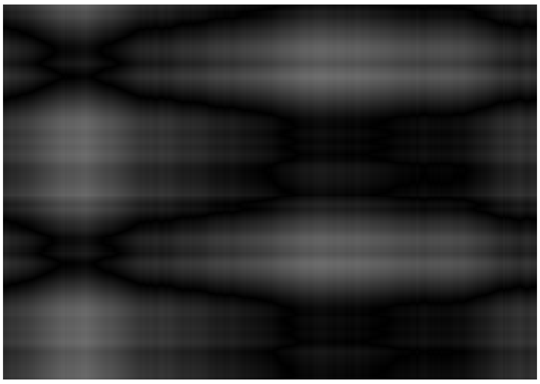
Third comparing sign of nūn ن from frame with sign of five “open palm”, letter nūn ن, and letter ṣād ص from Database.

1) Sign of Five “open palm”

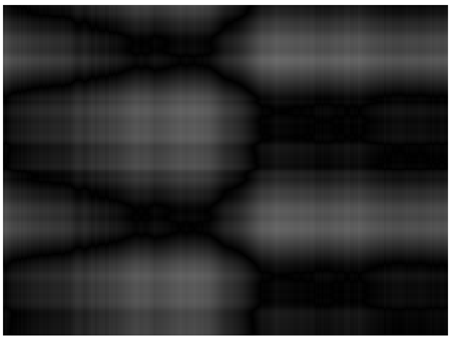




2) letter ṣādص



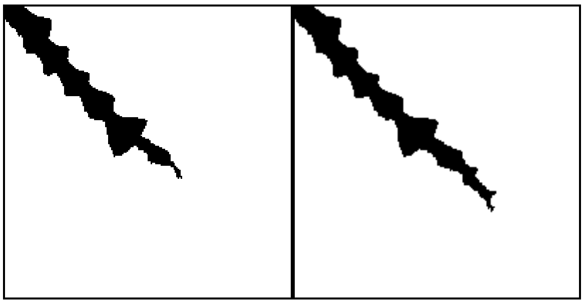
3) letter nūn ن



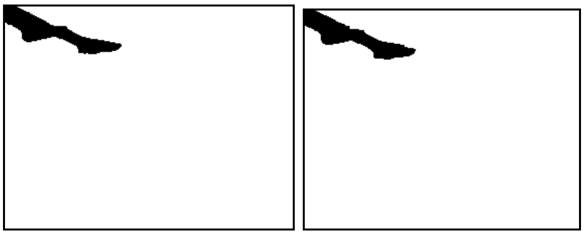
After viewing these images of Local Cost Matrices, we can see that there might be diagonal of the valley of low cost, but the warping path is not unique, so we can calculate the Accumulated Cost Matrix for the same signs First comparing sign of five "open palm" from frame with sign of five "open palm", letter nūn ن, and letter ṣād ص from Database.

We framed them with black line to be visible in the white background of the document

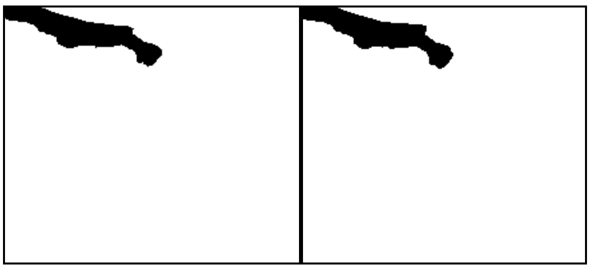
1) Sign of Five "open palm"



2) letter ṣād ص

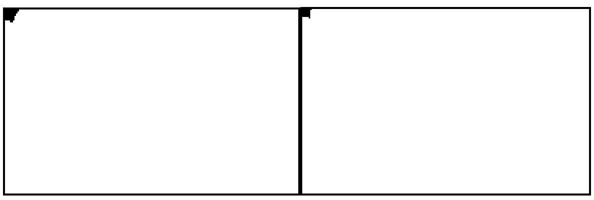


3) letter nūn ن



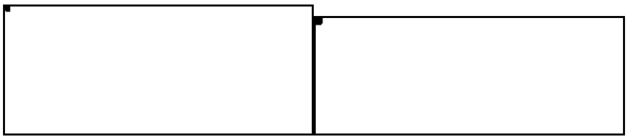
Second comparing sign of letter ṣād ص from frame with sign of five “open palm”, letter nūn ن, and letter ṣād ص from Database.

1) Sign of Five “open palm”



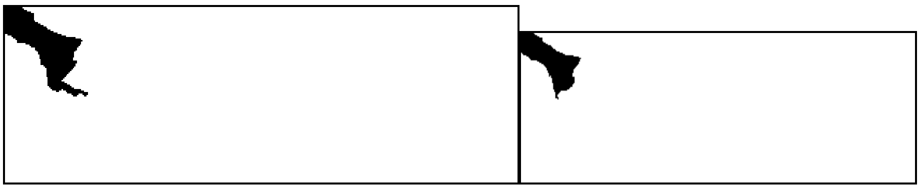
2) letter ṣādص

3) letter nūn ن

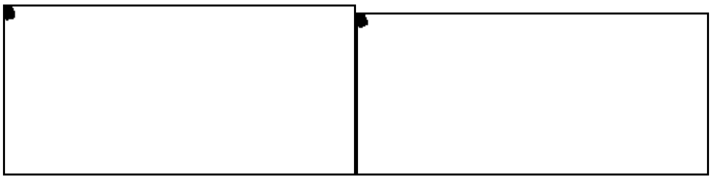


Third comparing sign of nūn ن from frame with sign of five “open palm”, letter nūn ن, and letter ṣād ص from Database.

1) Sign of Five “open palm”



2) letter ṣādص

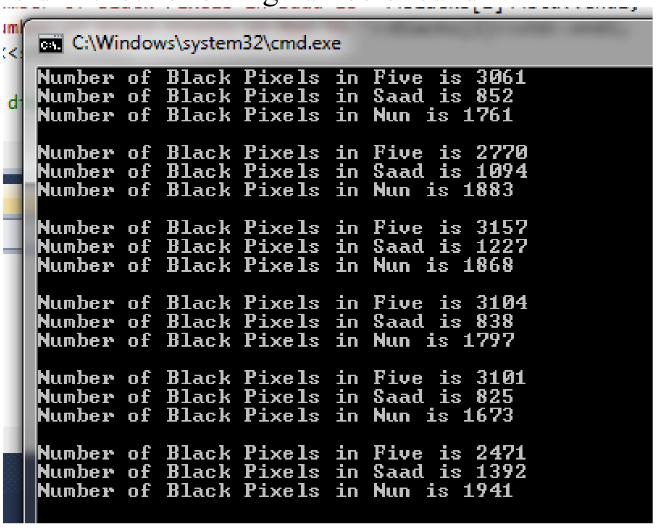


3) letter nūn ن

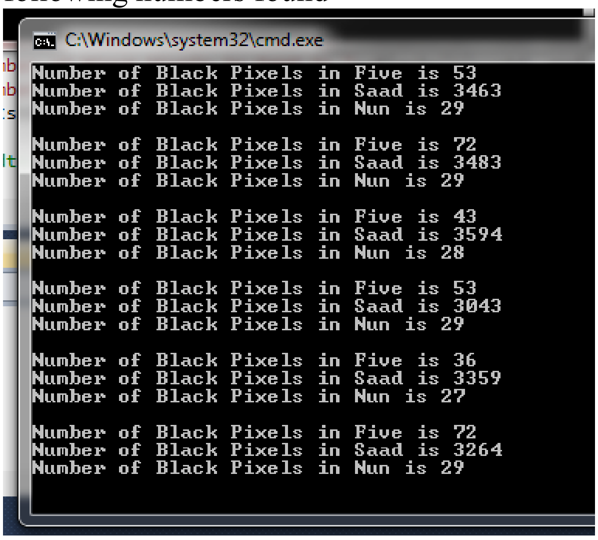
- those images are resized to fit the document

The idea of black pixels counting comes from the fact that black pixels are points of coincide between the two objects. Presenting that fact through experimental results are as following:-

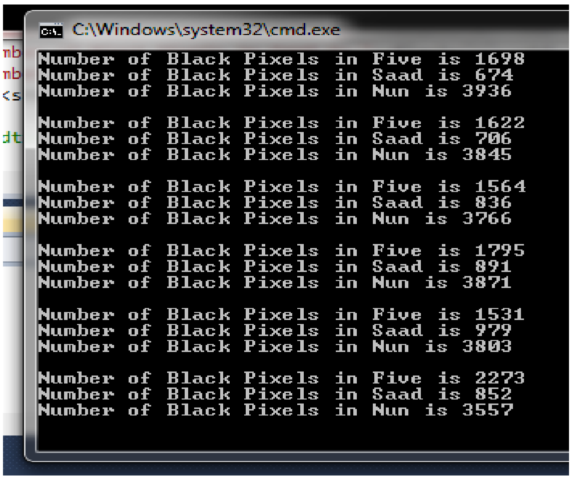
When frame radial signature representing open palm “five sign” compared to database following numbers found



Frame radial signature representing letter ṣādص compared to database following numbers found



Frame radial signature representing letter nūn ن compared to database following numbers found



So we calculated the following average data

	Database Five	Database Saad	Database Nun
Five from frames	3061	852	1761
	2770	1094	1883
	3157	1227	1868
	3104	838	1797
	3101	825	1673
	2471	1392	1941
average	2944	1038	1820.5
highest average	highest		

	Database Five	Database Saad	Database Nun
Saad from frames	53	3463	29
	72	3483	29
	43	3594	28
	53	3043	29
	36	3359	27
	72	3264	29
average	54.83333333	3367.666667	28.5
highest average		highest	

	Database Five	Database Saad	Database Nun
Nun from frames	1698	674	3936
	1622	706	3845
	1564	836	3766
	1795	891	3871
	1531	979	3803
	2273	852	3557
average	1747.166667	823	3796.333333
highest average			highest

The proposed methodology is implemented and is working near real-time recognition on system ‘Core 2 Duo (2.80 GHz) and 4 GB of memory on a personal computer with Microsoft Windows 7’ platform. The probability of recognition was around 80% for the 4 four sets of letters.

VI. CONCLUSIONS AND FUTUREWORK

Hand gesture recognition is a difficult problem and the current work is only a small step towards achieving real-time system working system.

The proposed methodology composed of four phases of processing: calibration to set proper thresholds, hand extraction using the environment thresholds, feature extraction using radial signature, and recognition based on dynamic time warping.

Currently the methodology recognizes sets of 4 or 5 signs efficiently very close to real-time with probability of success around 80%. Adding more signs the system response time was not real time. Also, the probability of recognition degrades. The proposed methodology is a simple out of the box approach which could be aided with more features, natural language processing as correction aid and cameras at different view angles as [21-25], more adaption to the extracted contour (smoothing and normalization), and finer radial sampling .

REFERENCES :

[1] Chen, Q., N.D. Georganas, and E.M. Petriu, Real-time Vision-based Hand Gesture Recognition Using Haar-like Features, in Instrumentation and Measurement Technology Conference, IEEE, Editor 2007: Warsaw, Poland.

[2] Mitra, S. and T. Acharya, Gesture recognition: A Survey, in IEEE Transactions on Systems, Man and Cybernetics2007, IEEE. p. 311-324.

[3] Murthy, G.R.S. and R.S. Jadon, A Review of Vision Based Hand Gestures Recognition. International Journal of Information Technology and Knowledge Management, 2009. 2(2): p. 405-410.

[4] Faria, B.M., N. Lau, and L.P. Reis. Classification of Facial Expressions Using Data Mining and machine Learning Algorithms. in 4ª Conferência Ibérica de Sistemas e Tecnologias de Informação. 2009. Póvoa de Varim, Portugal.

[5] Gillian, N.E., Gesture Recognition for Musician Computer Interaction, in Music Department2011, Faculty of Arts, Humanities and Social Sciences: Belfast. p. 206.

[6] Faria, B.M., et al., Machine Learning Algorithms applied to the Classification of Robotic Soccer Formations ans Opponent Teams, in IEEE Conference on Cybernetics and Intelligent Systems (CIS)2010: Singapore. p. 344 – 349

[7] Mannini, A. and A.M. Sabatini, Machine learning methods for classifying human physical activity from on-body accelerometers. Sensors, 2010. 10(2): p. 1154-75.

[8] Maldonado-Báscon, S., et al., Road-Sign detection and Recognition Based on Support Vector Machines, in IEEE Transactions on Intelligent Transportation Systems2007. p. 264-278.

[9] Vicen-Bueno, R., et al., Complexity Reduction in Neural Networks Appplied to Traffic Sign Recognition Tasks, 2004.

[10] Witten, I.H., E. Frank, and M.A. Hall, Data Mining - Pratical Machine Learning Tools and Techniques. Third Edition ed2011: Elsevier.

[11] Snyder, W.E. and H. Qi, Machine Vision2004: Cambridge University Press.

[12] Stephan, J.J. and S. Khudayer, Gesture Recognition for Human-Computer Interaction (HCI). International Journal of Advancements in Computing Technology, 2010. 2(4): p. 30-35.

[13] Phadtare, L.K.; Kushalnagar, R.S.; Cahill, N.D., "Detecting hand-palm orientation and hand shapes for sign language gesture recognition using 3D images," *Image Processing Workshop (WNYIPW), 2012 Western New York* , vol., no., pp.29,32, 9-9 Nov. 2012

[14] Hyung-Ji Lee; Jae-Ho Chung, "Hand gesture recognition using orientation histogram," *TENCON 99. Proceedings of the IEEE Region 10 Conference*, vol.2, no., pp.1355, 1358 vol.2, Dec 1999

[15] Yun Liu; Zhijie Gan; Yu Sun, "Static Hand Gesture Recognition and its Application based on Support Vector Machines," *Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing, 2008. SNPD '08. Ninth ACIS International Conference on*, vol., no., pp.517, 521, 6-8 Aug. 2008

[16] Liu Yun; Zhang Peng, "An Automatic Hand Gesture Recognition System Based on Viola-Jones Method and SVMs," *Computer Science and Engineering, 2009. WCSE '09. Second International Workshop on*, vol.2, no., pp.72, 76, 28-30 Oct. 2009

[17]T. Starner, J. Weaver, A. Pentland, Real-Time American
Sign Language Recognition from Video Using Hidden
Markov Modells, M.I.T. Media Laboratory Perceptual
Computing Section Technical Report No. 375, 1996.

[18] H. Hienz, K. Grobel, Automatic Estimation of BodyRegions from Video Images, in I. Wachsmuth and M.Fröhlich (Eds.) *Gesture and Sign Language in Human-Computer Interaction* (Berlin: Springer-Verlag,1998), 135-145.

[19] Meinard Müller, Information Retrieval for Music and Motion Springer-Verlag New York, Inc. Secaucus, NJ, USA ©2007

[20]Keogh, E. & Pazzani, M. (2001). Derivative Dynamic Time Warping. In First SIAM International Conference on Data Mining (SDM'2001), Chicago, USA

[21] Source: Internet web site of Columbia university, http://www.ee.columbia.edu/~dpwe/resources/matlab/dtw/

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

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

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

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