Journal of Computer Science

Original Research Paper

Arabic Sign Language Translator

[1]Abdelmoty M. Ahmed, [2]Reda Abo Alez, [3]Gamal Tharwat,
[4]Muhammad Taha, [5]B. Belgacem, [6]Ahmad M.J. Al Moustafa and [7]Wade Ghribi

[1,5,7]Department of Computer Engineering, College of Computer Science, King Khalid University Abha, Saudi Arabia
[1,2,3]Department of Systems and Computer Engineering, Faculty of Engineering, Al Azhar University, Cairo, Egypt
[4]Department of Mathematics, Faculty of Science, Al Azhar University, Cairo, Egypt
[6]Department Computer Science, College of Computer Science, King Khalid University Abha, Saudi Arabia

Article history:
Received: 30-03-2019
Revised: 27-05-2019
Accepted: 24-10-2019

Corresponding Author:
Abdelmoty M. Ahmed
Department of Computer Engineering, College of Computer Science, King Khalid University Abha, Saudi Arabia
Email: abd2005moty@yahoo.com
amoate@kku.edu.sa

Abstract: Development of systems that can recognize the gestures of Arabic Sign language (ArSL) provides a method for hearing impaired to easily integrate into society. This paper aims to develop a computational structure for an intelligent translator to recognize the isolated dynamic gestures of the ArSL. In our proposed system we build a datasets for ArSL from scratch of, we used 100-sign vocabulary from ArSL, we have applied 1500 video files for these signs. These signs were divided into five types of signs, recognizing a sign language gestures from dynamic gestures could be a difficult analysis issue. This paper solves the problem using gradient based key frame extraction technique. These key frames are useful for splitting continuous language gestures into sequence of signs for removing uninformative frames. After splitting of gestures every sign has been treated as isolated gesture. Then features of pre-processed gestures are extracted using Intensity Histogram by integrating with Gray Level Co-occurrence Matrix (GLCM) features. Experiments are performed on our own ArSL dataset and the matching between the ArSL and Arabic text is tested by Euclidian distance. The evaluation of the proposed system for the automatic recognition and translation for isolated dynamic ArSL gestures has proven to be effective and highly accurate. The experimental results show that the proposed system recognizes signs with a precision of 95.8%.

Keywords: Arabic Sign Language, Intelligent Translator, Isolated Gestures, Dynamic Gestures

Introduction

Deaf, dumb and also hearing impaired cannot speak as common persons; so they have to depend upon another way of communication using vision or gestures during their life. Persons with hearing loss and speech are deprived of normal contact with the rest of the community.

Sign language is not universal. It varies by country, geographic region or even by the interests of the hearing impaired. Sign language is a combination of descriptive and non-descriptive signs as well as alphabets (fingers spelling) signs.

There is no uniform format for ArSL, which make education a difficult challenge for the hearing impaired persons, so we need to have education to be bilingual for reading and writing, also the shortage of skilled teachers who know ArSL makes education difficult for the hearing impaired.

The presence of a machine intelligent system that can recognize ArSL and translate it into Arabic and vice versa will help to bridge the gap between hearing impaired and normal people, it will also help to integrate the hearing impaired at different levels of education and give them access to science using their mother tongue.

Recently, ArSL has been registered and documented in Arabic and many efforts have been made to establish a unified sign language in Arab countries.

Egypt, Tunisia, Jordan and the Gulf States are trying to unite the language and disseminate it among members of the deaf community and people interested in it.

Generally researchers in the field of individual deaf education are looking for automatic ways and methods for education to raise from the deaf academic level, enables them to self-learn and helps them to read and write in a manner consistent with sign language. The developments in machine intelligent technology, video processing technology and machine learning will help provide systems that have the ability to support and service deaf and dumb through computer technology to automate systems that help to communication between

them and make their life easier.

This paper is one of the efforts made to serve the hearing impaired and people interested in the ArSL; in this study we designed a system for automatic translation and recognition of isolated dynamic gestures for descriptive and non-descriptive signs in ArSL.

Our paper is formatted as follows:

Section 2 presents the structure of the proposed automatic translation system; section 3 provides and explains the main experimental results of the proposed system; Section 4 provides the conclusions and Section 5 presents future works.

System Methodology

Developments in pattern recognition techniques, video processing techniques, computer vision, automated learning technologies and statistical methods are some of the new developments in the computer's ability to predict and interpret automatic translation systems that provide communication to deaf and dumb people using computers. In this study, we introduce one of these systems.

We aim to build an intelligent translation system to recognize the isolated dynamic gestures of the ArSL to the corresponding Arabic spoken language. In Fig. 1 we have proposed to build a number of modules (sub-systems) required for this as follows:

1. Building and collection datasets subsystem
2. Video processing and understanding subsystem
3. Features extraction subsystem
4. Mapping between ArSL and Arabic text subsystem
5. Arabic text generation (transformation) subsystem

Building and Collection of Datasets Subsystem

The proposed system depends on building two datasets of video gestures from three resources: Standard ArSL dictionary, Collection of ArSL Videos from websites and video gestures from different human experts.

The proposed system includes two different data sets; one isolated gesture and another dataset for corresponding and matching Arabic words for these gestures.

No standard dataset is available for Arabic Sign Language. Therefore, in the proposed system we create a new dataset of ArSL video gestures, this dataset comprising of 100 dynamic and isolated signs of ArSL using one and two hands. The 100 Arabic sign were done by 3 different persons (5 gestures per every sign).

In this dataset, many gesture operators perform different isolated dynamic signs in more than one way. This is done to ensure that many users can use our system and are it is not restricted to one user; in addition, we keep in mind the different gestures background which is a uniform or un uniform background as well as the variability in age of gesture operators, the Fig. 2 Show some of the gestures operators.

The dataset of signs consists of two parts, the first part is used for the training set and the second is used for testing set. Users repeated the signs multiple times with different variable speed and movement, orientation, etc.

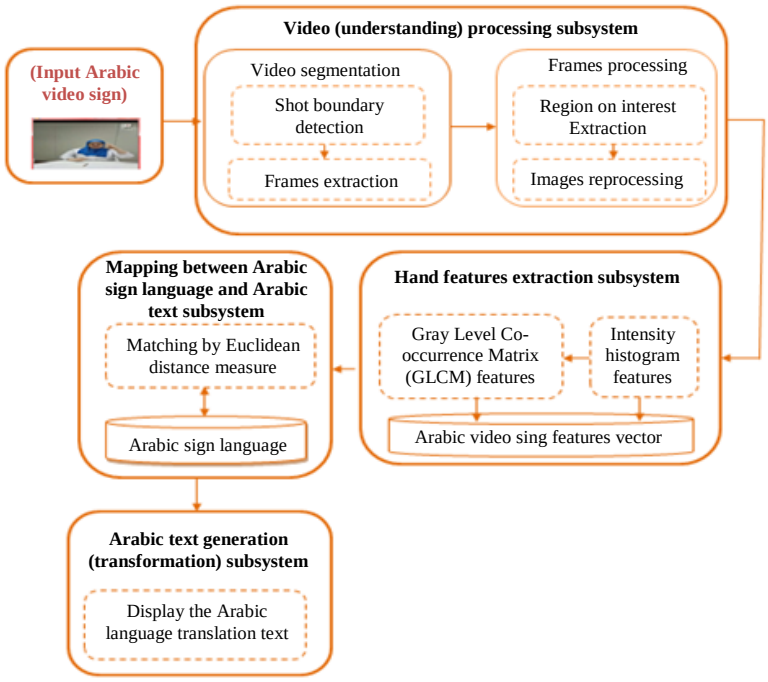

Fig. 1: Proposed system structure

Fig. 2: Different signers for our ArSL dataset of the proposed system

Table 1: Types of gestures classes that used in proposed system

Type of classes	No. of classes	No. of video files
Alphabets	30	530
Arabic life Expressions	10	115
Numbers	10	130
Prepositions, pronouns and question words	10	120
Common nouns and verbs	40	605
Total	100	1500

The ArSL Isolated Gesture Dataset

Our dataset collection contains 100 Arabic signs applied by 1500 dynamic isolated gestures for ArSL videos. This dataset is consisting of 30 video for Arabic finger spelling alphabet: from "أ (ألف)" to "ى (ياء)" plus two addition alphabets("ء","لا"), Arabic Numbers from one (pronounced wahed in Arabic) to ten (pronounced ashra in Arabic), 10 isolated Arabic Hand-shapes and Signs to show different motions (Prepositions, pronouns and question words), 10 video for Arabic life expressions and finally 40 common nouns and verbs used in the different aspects of life.

These 100 Arabic signs (1500 video files) are organized in five types of classes that used in our proposed system. In our work, the types of videos are AVI or MP4 and the extracted frames being RGB, each video is 30 frames per second. Table 1 describe types of classes for gestures that used in the proposed system.

Video (Understanding) Processing Subsystem

This part of the proposed system works to understand and analyse the videos of the Arabic sign language. This subsystem is the process of converting or dividing the gestures of the Arabic sign videos into their frames, to extract important information from video gestures of the ArSL without loss of information.

Video processing subsystem is using the Key Frames Extraction (KFE) algorithm (Kaur, 2017).

This subsystem is consist of two models the first is video segmentation and the second is frames processing, in the first model we segmented the input Arabic video sign into the sample of frames of Arabic video signs, the Fig. 3 shows a sample of the input Arabic video sign, the sections 2.2.1 and 2.2.2 we illustrate these models by details.

Video Segmentation Model

The main model in the video processing subsystem is video segmentation where the process of splitting video into a frameset is one of the most important processes for understanding video, KFE technique is very useful in this model, where they extract a small and useful number of frames, which can summarize video content. Video segmentation techniques, especially the KFE technology have become critical to the development of advanced digital video systems and here we propose some modifications and enhancements to this technique (El-Alfi *et al*., 2014).

In the video segmentation model we use two main algorithms: KFE and the Shot Boundary Detection (SBD). The steps of these algorithms will be explained in the following points.

DOI: 10.3844/jcssp.2019.1522.1537

a. Shot Boundary Detection Algorithm

SBD is an important component of video analysis, used in many video applications such as auto-detection, face recognition, gesture recognition and manual video editing and analysis, SBD is aim to divided an edited video into successive frames which show a continuous gradation of video, SBD is a main element of our proposed system, we have set out to improve SBD in accuracy and speed.

SBD also is considering an essential stage for most of the video application such as our proposed system containing the following features: Understanding, indexing, characterization of video segmentation, etc. SBD algorithm is achieved according to following five steps:

SBD algorithm

Input video frames sequence: $F(K) = F_1,F_2,..,F_v$ Let F(k) be the K[th] frame in video sequence, k = 1,2,..,*v* (v denotes the total number of video frames).

Step1: Partitioning a frame into blocks with m rows and n columns and *B* (*i, j, k*) stands for the block at (*i, j*) in the kth frame.

Step2: Computing x^2 histogram matching difference between the two corresponding blocks of consecutive frames in video sequence. *H*(*i, j, k*) and *H*(*i, j, k* +1) stand for the histogram of blocks at (*i, j*) in the k^{th} and (*k* +1)[th] frame respectively. Block's difference is measured by the following equation:

$$D(k,k+1,i,j)=\sum_{i=0}^{L-1}\frac{\left[H(i,j,k)-H(i,j,k+1)\right]^2}{H(i,j,k)} \quad (1)$$

where, *L* is the number of gray level in an image.

Step3: Computing the Equation (2) measured the histogram difference between frames *i* and *i*+1.

$$D(k,k+1)=\sum_{i=1}^{m}\sum_{i=1}^{n}W_{ij}D_B(k,k+1+i,j) \quad (2)$$

where, W_{ij} stands for the weight of block at (*i, j*).

Step4: Computing Threshold Automatically: Computing the Mean and standard variance of x^2 histogram difference over the whole video sequence. Mean and standard variance are defined as follows:

$$MD=\sum_{k=1}^{F_{v-1}}\frac{D(k,k+1)}{F_{v-1}} \quad (3)$$

$$STD=\sqrt{\sum_{k=1}^{F_{v-1}}\frac{\left(D(k,k+1)-MD\right)^2}{F_{v-1}}} \quad (4)$$

Step5: Shot boundary detection Let threshold $T = MD + a \times STD$. Where a is the constant. Say $a = 1$

$$T=MD+(a\times STD) \quad (5)$$

b. Key Frame Extraction Algorithm

Key frame plays a main concept for video abstracting. Key frames are a collection of distinctive frames obtained from video gradients. It provides an easy but effective method to summarize the video content for viewing, querying and retrieval of the key frames and video data to make processing amount up to the minimum. (Kathiriya, 2013).

Here are the key steps to implement this algorithm:

KFE algorithm

Step1: For finding a KEY frame from video, take first frame of each shot is reference frame and all other frames within shots are general frames.

Step2: Compute the difference between all the general frames and reference frame in each shot.

$$\mathbf{D_c}(1,\mathbf{k})=\sum_{i=1}^{m}\sum_{j=1}^{n}\mathbf{W_{ij}D_{CB}}(1,\ \mathbf{k},\mathbf{i},\mathbf{j})=2,3,\ldots\mathbf{F_{CN}} \quad (6)$$

Where, FC (k): The kth frame within the current shot, *k* = 1, 2, 3… F_{CN} (*k*) (F_{CN} (*k*) is the total number of the current shot)

Step3: Searching for the maximum difference within a shot:

$$\mathbf{Max(i)}=\left\{\mathbf{D_c}(1,\mathbf{k})\right\}_{\mathbf{max}},\mathbf{k}=2,3,4,\ldots.\mathbf{N} \quad (7)$$

Step4: To determine the Shot Type ‖ by the following Dynamic Shot or by relation between max(i) and MD: Static Shot(0)

$$\mathbf{shot\ type_c}=\begin{cases}1 & if\ Max(\mathbf{i})\geq MD\\ 0 & others\end{cases} \quad (8)$$

Step5: Determining the position of key frame.

Now if the Max (i) > MD, then shot type$_c$ = 1so the frame with the maximum difference is declared as key frame.

The videos for isolated dynamic signs in our dataset contain a large number of frames. It is not necessary for all these frames to specify the meaning of the video sign. Only a few of these frames are important for understanding the video. These most important frames are therefore known as key frames. In our system, we propose an algorithm to extract key frames to understand video (Zare and Zahiri, 2018; Kagalkar and Gumaste, 2016).

The selected key frame might change in position but not on shape, the hardness is the advantage that can distinguish between those frames that have a variation in shape.

Figure 4 and 5 explained the implementation of KFE algorithm steps on a sign video sample to extract general frames and key frames for the boy sign video, in Table 2. We present some experimental data and implementation results of an algorithm KFE that used in proposed system.

Figure 6 illustrate graph shows the percentage between key frames to general frames.

Frames Processing Model

Frames processing model is depend on several steps shown in Fig. 7.

After extract Key frames from the input video sign in the previous model, this key frames are processed, Each key frame which has been extracted contains a lot of details; we need only the Region on Interested

(ROI) which represent gesture in sign language. We are considered more steps but we can summarized this phase in two main steps, firstly is skin filtering and the second is hand cropping (Abdelmoty *et al*., 2015; Phu and Tay, 2014).

a. Skin Filter Technology

Firstly, skin filtering of the input image, which separates the skin colored pixels from the non-colored pixels. This way is very useful for hand detection. The skin filter is a technique for detecting areas of colored pixels in the skin and extracting them from the background, for detecting one or both hands.

The skin filter technology applied to the key frames extracted in the first model according to the following functions:

RGB image is converted to HSV (Hue, Saturation and Value)

Given RGB values, find the max and min.

max = maximum of RGB (9)

min = minimum of RGB (10)

$$V = \max \quad (11)$$

$$S = (\max - \min) / \max \quad (12)$$

If Saturation = 0, Hue is undefined

Else

$$\textbf{delta} = \textbf{max} - \textbf{min} \quad (13)$$

$$\textbf{Step}\,3\text{: if } \mathbf{R} = \max, \mathbf{H} = \frac{\mathbf{G} - \mathbf{B}}{\textbf{delta}} \quad (14)$$

$$\textbf{if } \mathbf{G} = \max, \mathbf{H} = 2 + \frac{\mathbf{R} - \mathbf{G}}{\textbf{delta}} \quad (15)$$

$$\textbf{if } \mathbf{B} = \max, \mathbf{H} = 4 + \frac{\mathbf{B} - \mathbf{R}}{\textbf{delta}} \quad (16)$$

$$\mathbf{H} = \mathbf{H} * 60 \quad (17)$$

$$\textbf{if } \mathbf{H} < 0, \mathbf{H} = \mathbf{H} + 360$$

Where S= Saturation, H=Hue, V=Value

b. Hand Crop Technology

Secondly, using the hand crop technology to recognize of different gestures for ArSL. In this step we need to know the area of the hand or hands to the wrist and remove the unnecessary part. The steps of hand cropping technology are summarized as follows (Rautaray and Agrawal, 2015):

1. The filtered image is scanned from all directions to separate the wrist from hand and then its location can be detected
2. Find the minimum and maximum positions of the skin pixels for all orientations in the key frame images

So we can get images that include X_{min}, Y_{min}, X_{max}, Y_{max}, one of which is hand and wrist mode. Figure 8 represents a sample of images after cropping.

After getting the desired part of the image, the image can be resized into 201×201 pixels and after that feature extraction of subsystem is execute (Raheja *et al*., 2012).

Hand Features Extraction Subsystem

Feature extraction is defined as the first stage of intelligent image analysis. It is a necessary step for any classification task.

Feature extraction plays a primary role for most image analysis tasks. But they are critical in pattern recognition. Many techniques and algorithms have been proposed for the extracted features have a significant role and importance so they must be carefully and accurately identified. The method of texture analysis is successful method to extract features (Abdelmoty *et al*., 2017), Texture is a main component of human visual perception also it is low level features of images, which is considered an essential feature when querying and retrieving private data in the image. It can be used to describe the contents of an image or a region in additional to colour features as colour features are not sufficient to identify the image since different images may have similar histograms. Texture property of an image has a very important aspect in the human visual system of recognition (Dinnes *et al*., 2018; Tian, 2013).

The method of texture analysis is principally divided into two approaches: Statistical and structural (Howarth *et al*., 204). The statistical approach is appropriate for our gesture frames that were collected and performed for our proposed system, because the frames that extracted from video is normally not periodical.

Fig. 3: Sample of input Arabic video signs

DOI: 10.3844/jcssp.2019.1522.1537

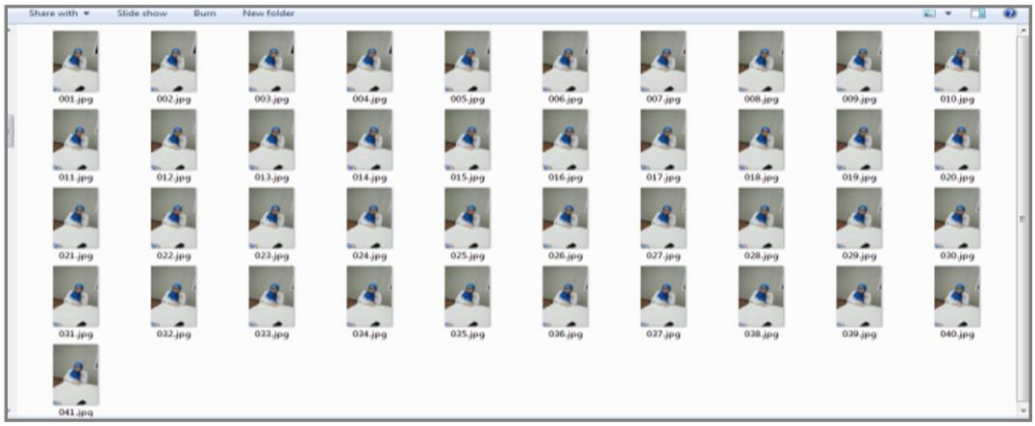

Fig. 4: Extracted general frames for the boy sign video

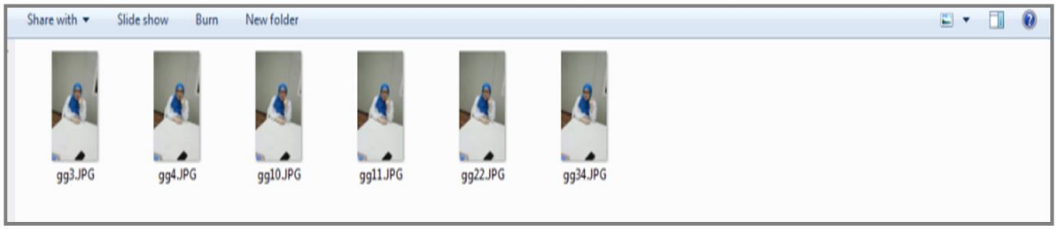

Fig. 5: Extracted key frames for the boy sign video

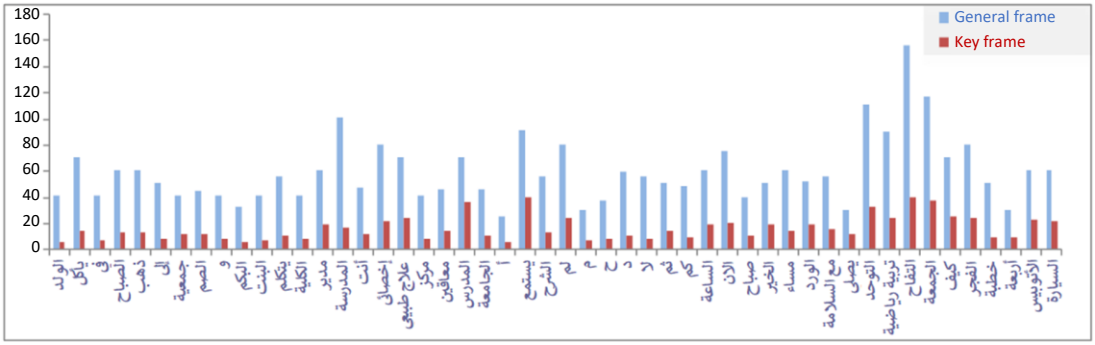

Fig. 6: The graphical showing average no. of key frames extracted from video file

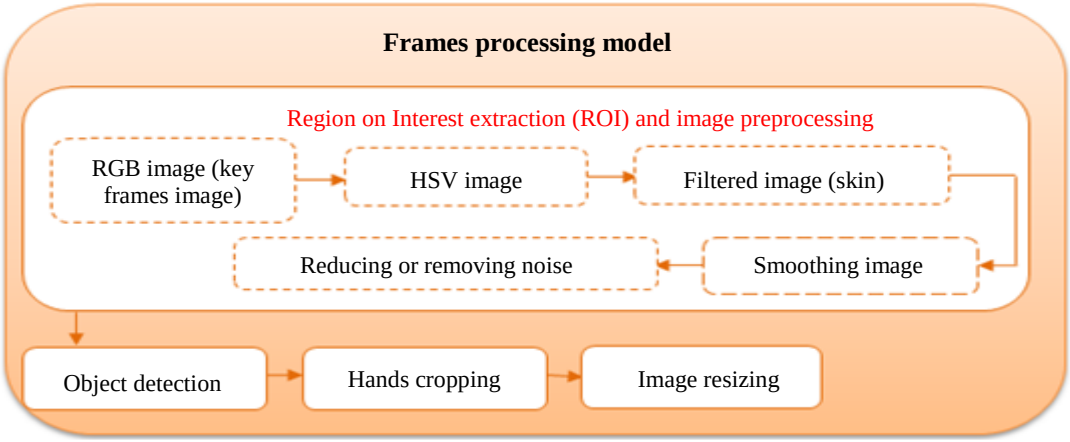

Fig. 7: Frames processing model block diagram

Fig. 8: Extracted region of interest from the input key frames

Table 2: Experimental Data that used in proposed system

Gesture No.	Gesture name in Arabic	Video Duration (in seconds)	General Frames	Key Frames	Percentage	Gesture No.	Gesture name in Arabic	Video Duration (in seconds)	General Frames	Key Frames	Percentage
G01	الولد	1.366700	41	06	14%	G26	لم	2.70000	81	24	30%
G02	يأكل	2.366667	71	15	21%	G27	م	1.03333	31	07	23%
G03	في	1.366670	41	07	17%	G28	ح	1.26667	38	09	24%
G04	الصباح	2.033300	61	13	21%	G29	د	2.00000	60	11	18%
G05	ذهب	2.033300	61	13	21%	G30	لا	1.86667	56	08	14%
G06	إلى	1.700000	51	09	18%	G31	ثم	1.70000	51	14	27%
G07	جمعية	1.366670	41	12	29%	G32	كم	1.63333	49	10	20%
G08	الصم	1.500000	45	12	27%	G33	الساعة	2.03333	61	19	31%
G09	و	1.366670	41	08	20%	G34	الان	2.53333	76	21	28%
G10	البكم	1.100000	33	06	18%	G35	صباح	1.33333	40	11	28%
G11	البنت	1.366670	41	07	17%	G36	الخير	1.70000	51	19	37%
G12	يتكلم	1.866670	56	11	20%	G37	مساء	2.03333	61	15	25%
G13	الكلية	1.366670	41	09	22%	G38	الورد	1.76667	53	19	36%
G14	مدير	2.033300	61	20	33%	G39	مع السلامة	1.86670	56	16	29%
G15	المدرسة	3.366670	101	17	17%	G40	يصلي	1.03333	31	12	39%
G16	أنت	1.600000	48	12	25%	G41	التوحد	3.70000	111	33	30%
G17	إخصائي	2.700000	81	22	27%	G42	تربية رياضية	3.00000	90	24	27%
G18	علاج طبيعي	2.366670	71	24	34%	G43	التفاح	5.20000	156	40	26%
G19	مركز	1.366670	41	09	22%	G44	الجمعة	3.90000	117	38	32%
G20	معاقين	1.533330	46	15	33%	G45	كيف	2.36670	71	26	37%
G21	المدرس	2.366670	71	36	51%	G46	الفجر	2.70000	81	24	30%
G22	الجامعة	1.533330	46	11	24%	G47	خطية	1.70000	51	10	20%
G23	أ	0.866670	26	06	23%	G48	أربعة	1.03300	31	10	32%
G24	يستمع	3.033330	91	40	44%	G49	الأتوبيس	2.03300	61	23	38%
G25	الشرح	1.866670	56	13	23%	G50	السيارة	2.03300	61	22	36%

In the statistical approach, there are various methods to measure the features of the texture such as: Spatial Gray-Level Dependence Method (SGLDM), Gray-Level Difference Method (GLDM), Gray-Level Run Length Method (GLRLM), Power Spectrum Method (PSM), Gray Level Co-occurrence Matrix (GLCM), Intensity histogram. In our proposed system we use the Intensity histogram features and GLCM features to make integrated between them to get the require features (Kumar *et al*., 2017).

a. Intensity Histogram Features

Histogram-based approach to structure analysis is dependent on the strength value concentrations upon all or component of an image represented like a histogram.

First order texture measures are statistics calculated from the original image values, like variance and do not consider pixel neighbourhood relationships histogram based approach to texture analysis is based on the intensity value concentrations on all or part of an image represented as a histogram. Common features include moments such as mean, variance, dispersion, mean, skewness, kurtosis Energy and Entropy.

Table 3 represent some statistical measures used in analyzing video frames based on the intensity histogram features.

b. Gray Level Co-occurrence Matrix

In this phase we use a statistical approach such as co-occurrence matrix to help providing valuable information about the relative position of the neighbouring pixels in an image where GLCM or gray-level spatial dependence matrix based calculations of second-order statistics.

GLCM texture considers the relation between two pixels at a time, called the reference and the neighbor pixels; we have proposed a set of 23 textual features extracted from the co-occurrence matrix such as Contrast, Homogeneity, Dissimilarity, Angular Second Moment, Energy and Entropy.

A level of a gray of co-occurrence matrix (GLCM) Includes information for pixels with similar gray values.

Table 4 represent some statistical measures used in analyzing video frames based on GLCM features.

DOI: 10.3844/jcssp.2019.1522.1537

c. Features Integration

This phase includes some processes: collect the features process of interest for both Intensity histogram features and GLCM and normalization process is applied on the collected key frames to overcome mainly three problems.

Firstly the variation of user position and the position of the camera is close or far away, secondly the variation of user's sizes and thirdly similarities in the features of the frames of some words and features integration process for fusing the hand features with the Intensity histogram features and GLCM features in order to form the final features vector.

Intensity Histogram Features vector of key frames (F_h) contain 6 features is represent the first-order statistical information about the image. Features derived through this approach consist of moment such as mean, variance, skwness, kurtosis, Energy and Entropy.

GLCM Features vector of key frames (F_g) contain 23 features is represent the second-order statistical information about the image features.

The features integration will occurred on the level of both vectors Intensity Histogram Features vector F_h and GLCM Features vector F_g to generate the integrated features vector $F_i = \{F_h, F_g\}$. The Intensity Histogram features vector has 6 features and the GLCM features vector has 23 features, so the resultant integrated features vector has of 26 because three features as same in both as in Fig. 9.

Mapping Between Arabic Sign Language and Arabic Text Subsystem

At this subsystem, samples of the features extracted from the previous subsystem are taken and defined in terms of the corresponding Arabic spoken language in order to be recognized, five samples were selected for this purpose.

Those selected samples based on the knowledge of experts in the field of ArSL translation. After calculating the previous statistics, we create the integrated features vector (F_i) which contains 26 features values (Mean, Standard deviation, kurtosis, Skewness, ASM, Energy, Correlation, Homogeneity Contrast…, etc.) for each frame.

Table 3: Texture features extracted based on the intensity histogram features

Feature Name	Feature equation	
Mean	$S_M = \mu_i = \frac{1}{N}\sum_{j=1}^{N}\left(f_{ij}\right)$	(18)
Standard deviation	$S_D = \sigma_i = \frac{1}{N}\left[\sum_{j=1}^{N}\left(f_{ij}-\mu_i\right)^2\right]^{1/2}$	(19)
Skewness	$S_S = \gamma_i = \frac{1}{N}\left[\sum_{j=1}^{N}\left(f_{ij}-\mu_i\right)^3\right]^{1/3}$	(20)

Table 4: Texture features extracted from GLCM

Feature Name and equation	
$Contrast = \sum_{n=0}^{Ng-1} n^2 \left\{\sum_{i=1}^{Ng}\sum_{j=1}^{Ng} P(i,j), \lvert i-1\rvert = n\right\}$	(21)
$Correlation = \sum_{i=1}^{Ng}\sum_{j=1}^{Ng} \frac{(i,j)P(i,j)-\mu_x\mu_y}{\sigma_x\sigma_y}$	(22)
$Energy = \sum_{i}^{Ng}\sum_{j}^{Ng} P(i,j)^2$	(23)
$Homogeneity = \sum_{i}^{Ng}\sum_{j}^{Ng} \frac{P(i,j)}{1+\lvert i-1\rvert}$	(24)

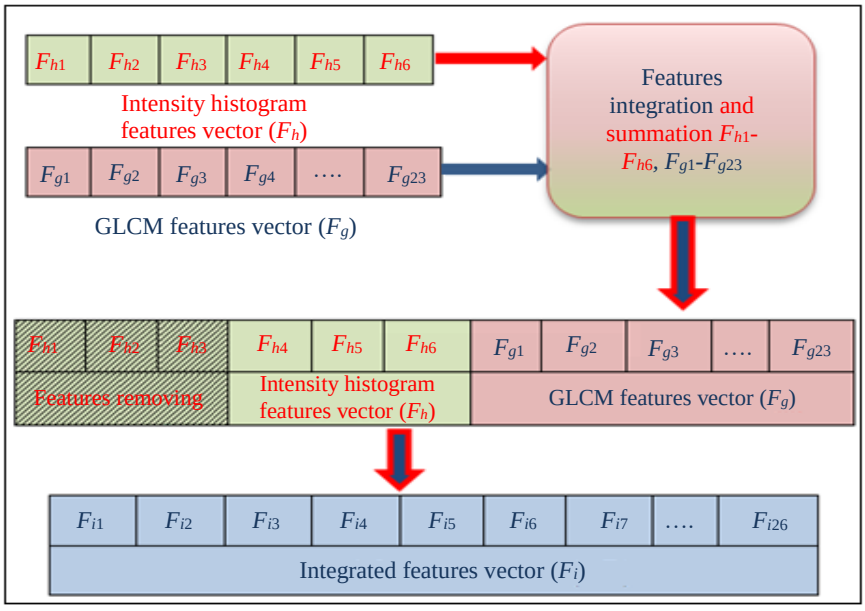

Fig. 9: Features integration phase block diagram

Now the matching process will be considered, where the obtained pattern containing the features vector will be compared with the built in dataset patterns.

The feature vectors corresponding to an image *k* can be denoted by:

$$fi^{(k)} = \left\{ f_{i1}{}^{(k)}, f_{i2}{}^{(k)}, f_{i3}{}^{(k)}, \ldots\ldots, f_{in}{}^{(k)} \right\} \quad (25)$$

where, each component $fi^{(k)}$ typically an invariant moment function of the video frame. The set of all $fi^{(k)}$ constitute the reference library of the features vectors. The video frame for which the reference vectors are computed and stored is dataset.

The images for which the reference vectors are computed and stored is a set of patterns used for pattern recognition. The problem considered here is to match a features vector $fi^{(M)}$:

$$fi^{(M)} = \left\{ f_{i1}{}^{(M)}, f_{i2}{}^{(M)}, f_{i3}{}^{(M)}, \ldots\ldots, f_{in}{}^{(M)} \right\} \quad (26)$$

where, each component $fi^{(M)}$ typically an invariant moment of the match image.

The retrieval is performed by the similarity measure, which using to compute distance between stored images classes in the dataset and the match image. In our proposed system we use weighted Euclidean distance measure to compute distance between stored feature vectors in the database and the feature vector of match image. The formula of weighted Euclidean distance measure between vectors can be written as follows (Shijin and Dharun, 2017; Ibrahim *et al*., 2018):

$$d\left(fi^{(M)}, fi^{(k)} \right) = \sqrt{\sum_{i=1}^{n} W_i \left(fi^{(M)} - fi^{(k)} \right)^2} \quad (27)$$

where, W_i denotes the weight added to the component V_i to balance the variations in the dynamic range.

The value of *k* for which the function *d* is minimum, is selected as the matched image index. The value of n denotes the dimension of the features vector and the *N* value denotes the number of images in database.

Where the weight W_i is given by the following equation:

$$W_i = \frac{N}{\sum_{k=1}^{n} \left(fi^{(k)} - \overline{fi^{(k)}} \right)^2} \quad (28)$$

where, the $\overline{V_i}$ is given by the following equation:

$$\overline{fi^{(k)}} = \frac{\sum_{i=1}^{n} \left(fi^{(k)} \right)}{N} \quad (29)$$

Euclidean Distance Algorithm

The length of the line segment connecting the *X* and *Y* points is the Euclidean distance between these two points $\overline{xy}$ as describe in equation (Nagarajan and Subashini, 2015):

$$d1 = \left(\left(Xi_1 - Xj_1 \right)^2 + \ldots + \left(Xi_n - Xj_n \right)^2 \right)^{\frac{1}{2}} \quad (30)$$

In this study, Euclidean distance Algorithm is used to compute distance between stored images classes in the datasets and the match image:

Euclidean distance algorithm

Step 1: Measure the distance between the new match images M and training images in database by weighted Euclidean distance according to the above Equation 30
Step 2: Sort the distance values as $d_1 \leq d_1+1$
Step 3: Select the smallest distance ratio.
Step 4: Select Texture class which will retrieved results on them.
Step 5: App1y Arabic Text generation (transformation) Subsystem

Arabic Text Generation Subsystem

After selected the texture class which contain the match frames as discussed in the previous subsystems, the result of translation will transfer the sign language image into corresponding Arabic text.

According to databases that included alphabets, words, numbers and Arabic life expressions and their pattern description. Depending on the result matching pattern the word will be specified. This step will be repeated for each input video key frame to allow the integration between their descriptions and this leads to displaying the words representing the translation of the input video pattern.

The descriptions obtained from the gestures of ArSL dataset for every key frame are concatenated to transform the video into a word.

Results and Discussion

The proposed system is designed and implemented to translate the ArSL to Arabic text. Our system translates and recognizes gestures using one hand or both hands. The signers are not required to wear any gloves or to use any devices to interact with the system. The Graphical User Interface (GUI) for the proposed system was implemented by MATLAB, Fig. 10 Illustrate GUI of proposed system.

DOI: 10.3844/jcssp.2019.1522.1537

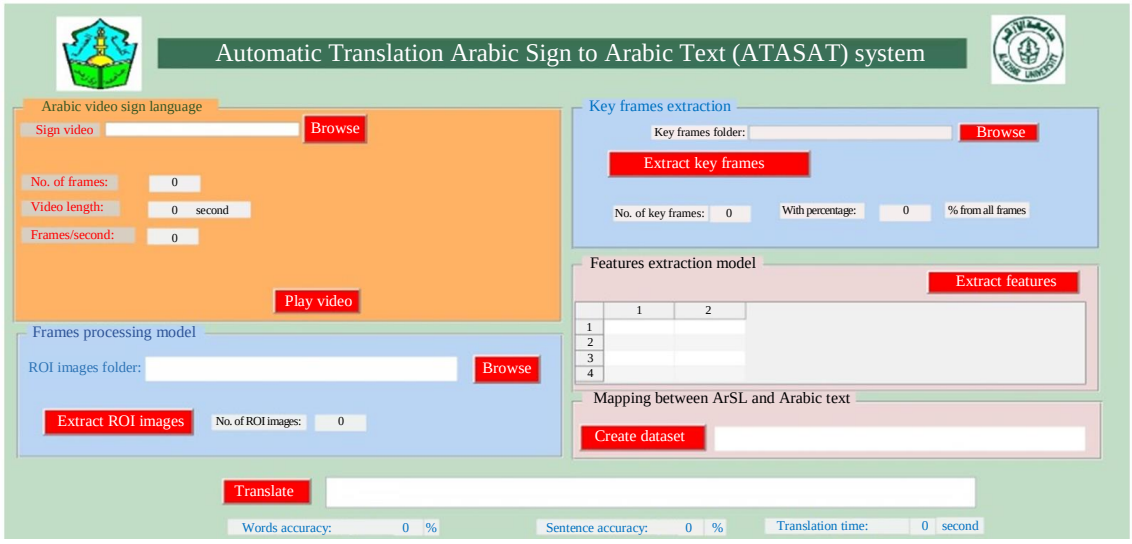

Fig. 10: Snapshot of the GUI for proposed system

Experiment 1: Arabic Alphabets Recognition

This section is illustrating some of the experimental results of the proposed system. Here we will discuss the results of the first experiment, which deals with the recognition of the alphabet only and is captured with one hand.

Table 5 gives the details of average recognition rate obtained by the proposed system using integration feature vector to recognize alphabets, also Fig. 11. Shows the total number of samples correctly identified in the alphabet class.

The recognition rate in our proposed system is including signer dependent and signer independent. The number of training samples considered is 5 for each alphabet sign. After that a number of sign samples are then tested from multiple users, a number of system signs operators (signers) were shown in Fig. 2 the total no. of different classes of each type of gestures for dynamic isolated dataset were shown in Table 1.

Experiment 2: Arabic Life Expressions Recognition

In this section we will discuss the results of the second experiment, which deals with the recognition of the Arabic life expressions and was captured by one or both hands and use face expressions.

Table 6 gives the details of average recognition rate obtained by the proposed system using integration feature vector to recognize Arabic life expressions, also Fig. 12 shows the total number of samples correctly recognized in the Arabic life expressions class.

The rate of recognition in this experiment is 93.91%, where 108 out of 115 gestures were recognized in this experiment.

Experiment 3: Arabic Numbers Recognition

Here we will discuss the results of the third experiment, which deals with the recognition of the Arabic numbers and was captured with one hand.

Table 7 gives the details of average recognition rate obtained by the proposed system using integration feature vector to recognize Arabic numbers, also Fig. 13. Shows the total number of samples correctly recognized in the Arabic numbers class.

The rate of recognition in this experiment is 96.15%, where 125 out of 130 gestures were recognized in this experiment.

Experiment 4: Prepositions, Pronouns and Question Words Recognition

The experimental results of the fourth experiment in our proposed system will be discussed in this part, which deals with the recognition of the prepositions, pronouns and question words and was captured with one or both hands and face expressions.

Table 8 gives the details of average recognition rate obtained by the proposed system using integration feature vector to recognize prepositions, pronouns and question words, also Fig. 14. Shows the total number of samples correctly recognized in the repositions, pronouns and question words class.

The rate of recognition in this experiment is 91.67%, where 110 out of 130 gestures were recognized in this experiment.

Experiment 5: Common Nouns and Verbs Recognition

The experimental results of the fifth experiment in our proposed system we will discuss in this part, which deals with the recognition of the common nouns and verbs and was captured with one or both hands and face expressions.

The rate of recognition in this experiment is

96.86%, where 586 out of 605 gestures were recognized in this experiment

Performance Evaluation

In this section we shows the performance evaluation from the above experimental results, Firstly; the proposed system was able to perform a recognition and translation of dynamic isolated signs from ArSL into Arabic text with recognition rate of 95.80%, the number of recognized gestures are 1437 pattern. Secondly the recognition rate of 4.20% and the number of unrecognized gestures are 63 patterns. These unrecognized gestures are used to update the system dataset, beside some more unrecognized gestures by domain experts. Finally the process of updating or adding for gestures will give us increase in the percentage of accuracy and decrease in error rate. Thus, the proposed system can be used on a great scale to increase the communication between hearing impaired and normal people. We summarize all the performance evaluations of the proposed system according to the different types of classes are tabulated in Table 9, also Fig. 15 and 16 are showing more details about performance evaluation results.

Table 5: Matching results of the alphabets classes

No.	Alphabets (class)	No. of signs (video files)	No. of correctly signs	Accuracy percentage
1	أ	17	16	94.12%
2	ب	14	13	92.86%
3	ت	14	14	100.00%
4	ث	18	18	100.00%
5	ج	13	13	100.00%
6	ح	14	14	100.00%
7	خ	18	17	94.44%
8	د	18	17	94.44%
9	ذ	18	16	88.89%
10	ر	18	18	100.00%
11	ز	24	22	91.67%
12	س	17	17	100.00%
13	ش	18	16	88.89%
14	ص	18	18	100.00%
15	ض	17	16	94.12%
16	ط	17	17	100.00%
17	ظ	16	15	93.75%
18	ع	17	16	94.12%
19	غ	12	11	91.67%
20	ف	11	10	90.91%
21	ق	17	17	100.00%
22	ك	28	26	92.86%
23	ل	18	17	94.44%
24	م	25	24	96.00%
25	ن	18	18	100.00%
26	هـ	17	16	94.12%
27	و	26	26	100.00%
28	لا	21	20	95.24%
29	ى	17	17	100.00%
30	ء	14	13	92.86%
Total		**530**	**508**	**95.85%**

Table 6: Matching results of Arabic life expressions classes

No	Arabic life expressions (class)	No. of Signs (video files)	No. of Signs correctly	Accuracy percentage
1	صباح الخير	10	10	100.00%
2	مساء الخير	10	10	100.00%
3	السلام عليكم	10	9	90.00%
4	تفضل	12	12	100.00%
5	مع السلامة	8	8	100.00%
6	أهلا وسهلا	10	7	70.00%
7	شكراً	15	14	93.33%
8	صديقي	10	9	90.00%
9	وحشتنى	15	14	93.33%
10	صباح الورد	15	15	100.00%
Total		**115**	**108**	**93.91%**

Table 7: Matching results of Arabic numbers classes

No.	Arabic numbers (class)	No. of Signs (video files)	No. of Signs correctly	Accuracy percentage
1	واحد	10	10	100.00%
2	إثنين	10	10	100.00%
3	ثلاثة	10	10	100.00%
4	أربعة	11	11	100.00%
5	خمسة	14	10	71.43%
6	ستة	16	16	100.00%
7	سبعة	17	14	82.35%
8	ثمانية	17	17	100.00%
9	تسعة	12	12	108.33%
10	عشرة	13	13	107.69%
Total		**130**	**125**	**96.15%**

Table 8: Matching results of prepositions, pronouns and question words classes

No.	Prepositions, pronouns and question words (class)	No. of Signs (video files)	No. of Signs correctly	Accuracy percentage
1	في	7	7	100.00%
2	إلى	14	12	85.71%
3	و	17	16	94.12%
4	ثم	14	14	100.00%
5	لا	7	6	85.71%
6	لم	12	11	91.67%
7	كم	15	13	86.67%
8	كيف	14	13	92.86%
9	أنت	10	9	90.00%
10	أنا	10	9	90.00%
Total		**120**	**110**	**91.67%**

Table 9: The results of the proposed system according to the different types of classes

Type of classes	No. of classes	No. of Signs (video files)	No. of Signs correctly	Accuracy percentage
Alphabets	30	530	508	95.85%
Arabic life Expressions	10	115	108	93.91%
Numbers	10	130	125	96.15%
Prepositions, pronouns and question words	10	120	110	91.67%
common nouns and verbs	40	605	586	96.86%
Total	**100**	**1500**	**1437**	**95.80%**

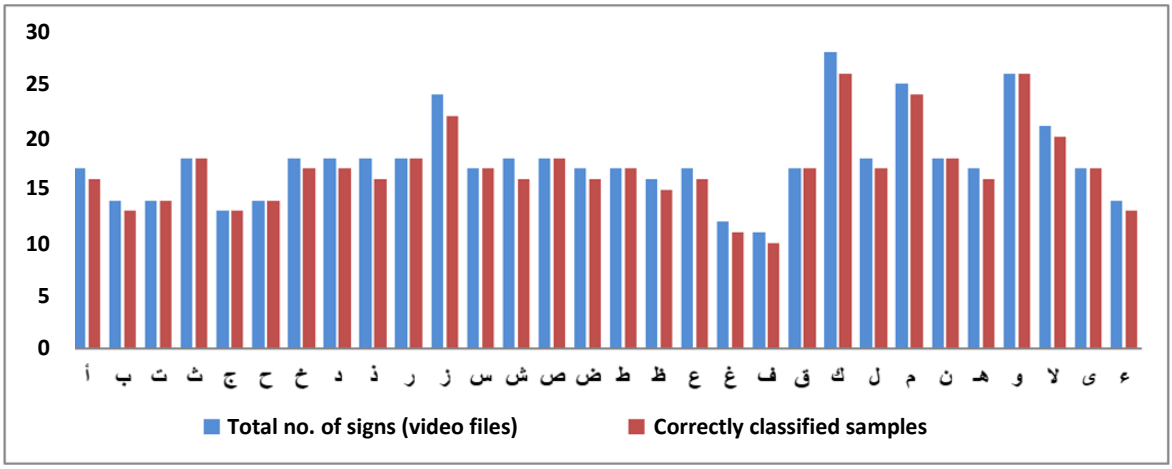

Fig. 11: The graph shows the total number of samples correctly identified in the alphabetical categories

DOI: 10.3844/jcssp.2019.1522.1537

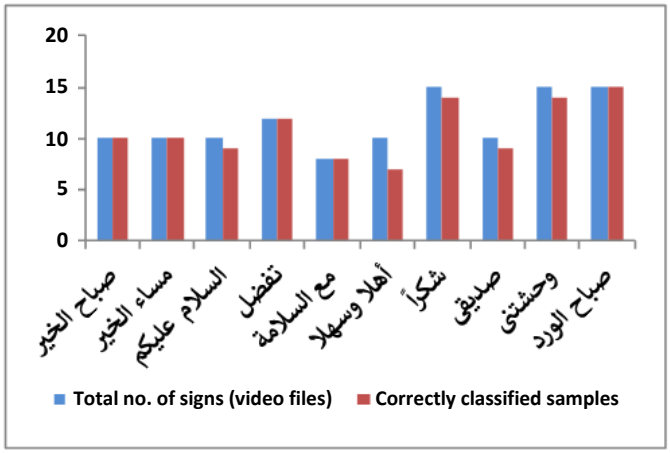

Fig. 12: The graphical contains the total number of samples identified in the Arabic life expression categories

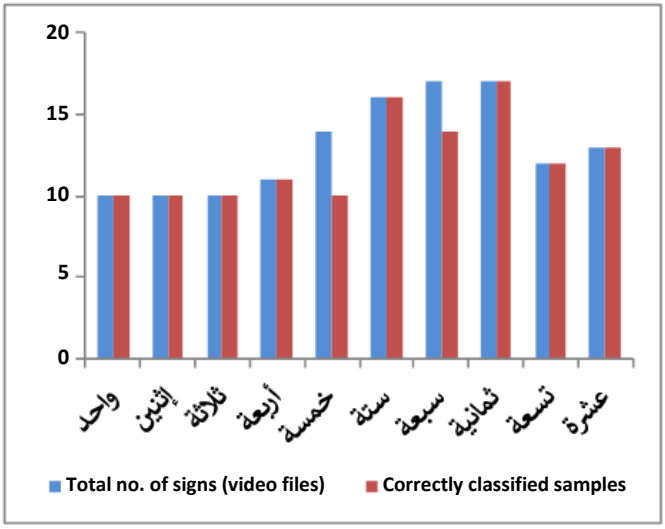

Fig. 13: The graphical showing the total number of samples and correctly classified samples in Arabic numbers classes

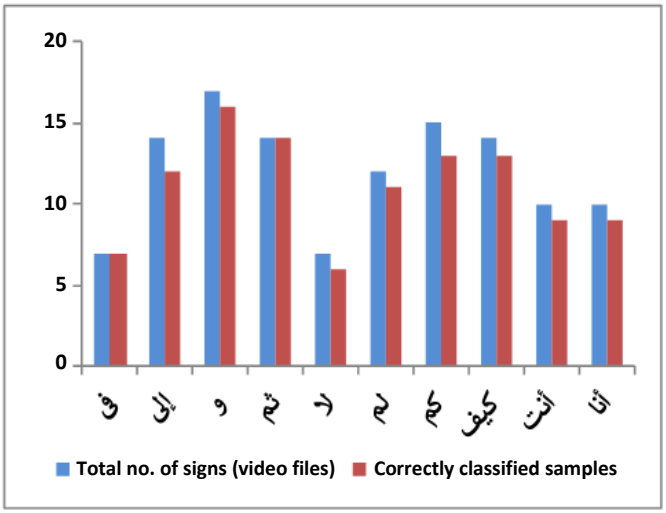

Fig. 14: The graphical showing the total No. of samples and correctly classified samples in prepositions, pronouns and question words classes

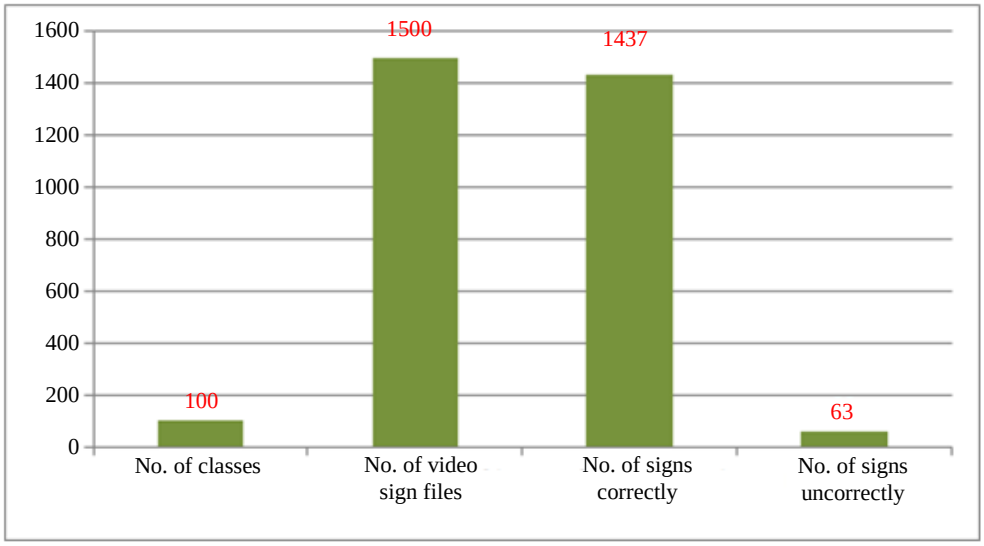

Fig. 15: The graphical showing the total No. of samples, No. of samples classified correctly in the proposed system

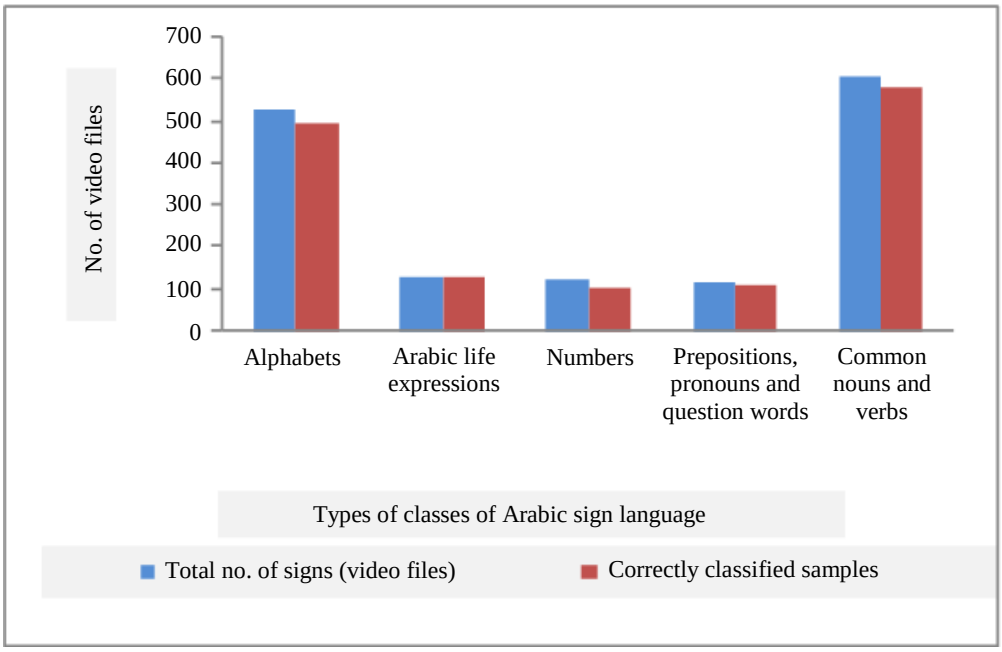

Fig. 16: The graphical showing the total No. of samples, No. of samples classified correctly in each type class of proposed system

Conclusion

The research achieved the desired goal to build a system to recognize and translate isolated dynamic gestures, which are performed using one or both hands with facial expressions. Many efforts have been made to establish ArSL translation systems and this work is one of these efforts. It presents a proposed system to support the communication between deaf, dumb and normal people by translating ArSL video to its corresponding text, our system was applied on 100 Arabic video signs. We found that the designed system was able to perform a translation of ArSL into Arabic text with recognition rate of 95.8%. Through the addition of more unrecognized gestures by domain experts to the system dataset, we will obtain a higher recognition rate. With more efforts we can obtain standard Arabic sign language translator.

Recommendations and Future Work

In this research, an effective system was developed to recognize and video translate dynamic isolated signs for ArSL by using an optimization algorithm to extract the Key frame to speed up and enhance the recognition function. There is no standard dataset obtainable and ready for ArSL signs; therefore, we created our own dataset from the scratch, that consisting of 1,500 videos representing 100 signs.

We proposed a system for recognition and translation of ArSL dynamic isolated sign by using one or both handed sign (manual features) without considering non-manual features. Our future work will focus on following issues: Adding non-manual features for the complete ArSL recognition, adding more signs to dataset and fusion of classifiers for better recognition rate. As well as work on recognition and translating continuous dynamic gestures (sentences).

Author's Contributions

Abdelmoty M. Ahmed: Designed the research plan, organized and ran the experiments, contributed to the presentation, analysis and interpretation of the results, added and reviewed genuine content where applicable.

Reda Abo Alez: Supervised the study and made considerable contributions to this research by critically reviewing the manuscript for significant intellectual content.

Gamal Tharwat: Supervised the study and made considerable contributions to this research by critically reviewing the manuscript for significant intellectual content.

Muhammad Taha: Presented idea and developed the theory and performed the computations and verified the analytical methods.

B. Belgacem: Made considerable contributions to this research by critically reviewing the literature review and the manuscript for significant intellectual content.

Ahmad M.J. Al Moustafa: Took the lead in review writing the manuscript and provided critical feedback in manuscript.

Wade Ghribi: took the lead in review writing the manuscript and provided critical feedback in manuscript.

Conflict of Interest

The authors declare that they have no Conflict of Interest.

References

Abdelmoty, M.A., R.A. Alez, M. Taha and G. Tharwat, 2015. Propose a new method for extracting hand using in the Arabic Sign Language Recognition (ArSLR) system. Int. J. Eng. Res. Technol.

Abdelmoty, M.A., W. Ghribi, R.A. Alez, G. Tharwat and M. Taha *et al*., 2017. Towards the design of automatic translation system from Arabic sign language to Arabic text. Proceedings of the International Conference on Inventive Computing and Informatics, Nov. 23-24, IEEE Xplore Press, Coimbatore, India, pp: 653-637. DOI: 10.1109/ICICI.2017.8365365

Dinnes, J., J.J. Deeks, N. Chuchu, D.R.L. Ferrante and R.N. Matin *et al*., 2018. Dermoscopy, with and without visual inspection, for diagnosing melanoma in adults. Cochrane Database Syst. Rev., 12: CD011902- CD011902.
DOI: 10.1002/14651858.CD011902.pub2

El-Alfi, A., A. El-Gamal and R. El-Adly, 2014. Real time Arabic sign language to Arabic text and sound translation system. Int. J. Eng.

Howarth, P. and S. Rüger, 2004. Evaluation of Texture Features for Content-Based Image Retrieval. In: Image and Video Retrieval, Enser, P., Y. Kompatsiaris, N.E. O'Connor, A.F. Smeaton and A.W.M. Smeulders (Eds.), Springer, Berlin, Heidelberg,
ISBN-13: 978-3-540-22539-3, pp: 326-334.

Ibrahim, N.B., M.M. Selim and H.H. Zayed, 2018. An Automatic Arabic Sign Language Recognition System (ArSLRS). J. King Saud Uni. Comput. Inform. Sci., 30: 470-477.

Kagalkar, R.M. and S. Gumaste, 2016. Gradient based key frame extraction for continuous Indian sign language gesture recognition and sentence formation in Kannada language: A comparative study of classifiers. Int. J. Comput. Sci. Eng., 4: 1-11.

Kathiriya, P.V., 2013. $\chi 2$ (chi-square) based shot boundary detection and key frame extraction for video. Int. J. Eng. Sci., 2: 17-21.

Kaur, S., 2017. Review of purposed method for key frame extraction from videos. Int. J. Adv. Res. Comput. Sci.

Kumar, S., M.K. Bhuyan and B.K. Chakraborty, 2017. Extraction of texture and geometrical features from informative facial regions for sign language recognition. J. Multimodal User Interfaces, 11: 227-239.

Nagarajan, S. and T. Subashini, 2015. Weighted Euclidean Distance Based Sign Language Recognition Using Shape Features. In: Artificial Intelligence and Evolutionary Algorithms in Engineering Systems, Suresh, L., S. Dash and B. Panigrahi (Eds.), Springer, New Delhi,
ISBN-13: 978-81-322-2134-0, pp: 149-156.

Phu, J.J. and Y.H. Tay, 2014. Computer vision based hand gesture recognition using artificial neural network. University Tunku Abdul Rahman (UTAR), Malaysia.

Raheja, J.L., K. Das and A. Chaudhary, 2012. Fingertip detection: A fast method with natural hand. Int. J. Embedded Syst. Comput. Eng.

Rautaray, S.S. and A. Agrawal, 2015. Vision-based hand gesture recognition for human computer interaction: A survey. Artificial Intell. Rev., 43: 1-54.

Shijin, K.P.S. and V. Dharun, 2017. Extraction of texture features using GLCM and shape features using connected regions. Int. J. Eng. Technol., 8: 2926-2930. DOI: 10.21817/ijet/2016/v8i6/160806254

Tian, P.D., 2013. A review on image feature extraction and representation techniques. Int. J. Multimedia Ubiquitous Eng., 8: 385-396.

Zare, A.A. and S.H. Zahiri, 2018. Recognition of a real-time signer-independent static Farsi sign language based on Fourier coefficients amplitude. Int. J. Machine Learn. Cybernet., 9: 727-741.