Journal of King Saud University – Computer and Information Sciences 30 (2018) 470–477

Contents lists available at ScienceDirect

Journal of King Saud University – Computer and Information Sciences

journal homepage: www.sciencedirect.com

An Automatic Arabic Sign Language Recognition System (ArSLRS)

Nada B. Ibrahim *, Mazen M. Selim, Hala H. Zayed

Department of Computer Science, Faculty of Computers and Informatics, Benha University, Benha, Egypt

ARTICLE INFO

Article history:
Received 5 June 2017
Revised 12 September 2017
Accepted 26 September 2017
Available online 5 October 2017

Keywords:
Gesture recognition
Arabic sign language recognition
Isolated word recognition
Image-based recognition

ABSTRACT

Sign language recognition system (SLRS) is one of the application areas of human computer interaction (HCI) where signs of hearing impaired people are converted to text or voice of the oral language. This paper presents an automatic visual SLRS that translates isolated Arabic words signs into text. The proposed system has four main stages: hand segmentation, tracking, feature extraction and classification. A dynamic skin detector based on the face color tone is used for hand segmentation. Then, a proposed skin-blob tracking technique is used to identify and track the hands. A dataset of 30 isolated words that used in the daily school life of the hearing impaired children was developed for evaluating the proposed system, taking into consideration that 83% of the words have different occlusion states. Experimental results indicate that the proposed system has a recognition rate of 97% in signer-independent mode. In addition to, the proposed occlusion resolving technique can outperform other methods by accurately specify the position of the hands and the head with an improvement of 2.57% at $\tau = 5$ that aid in differentiating between similar gestures.

1. Introduction

Hearing impairment is a board term referred to partial or complete loss of hearing in one or both ears. The level of impairment varies between mild, moderate, serve or profound.

Granting to the world health organization (WHO) in the year of 2017, Over 5% of the world's population – 360 million people' has disabling hearing loss (328 million adults and 32 million children). Roughly one-third of people over 65 years of historic period are affected by disabling hearing loss. The majority of people with disabling hearing loss live in low and middle income countries (Center, 2017).

SLRS is one of the application areas of HCI. The main goal of SLRS is to recognize signs of hearing impaired people and converting them to text or voice of the oral language and vice versa. These systems use either isolated or continuous signs. The performer of the isolated systems signs only one letter or word at a time while in continuous systems the performer signs one or more complete sentences. Further, SLRS can be categorized as a signer-dependent or signer-independent. Systems rely on the same signers to perform in both training and testing phases are signer-dependent and this affects the recognition rate positively. On the other hand, in signer-independent systems singers performed the training stage is not admitted in the testing stage and this adds a challenge of adapting the system to accept any signer. The goal of SLRS can be achieved by either a sensor-based or an image-based system.

The sensor-based system employs just about variety of electromechanical devices that are incorporated with many sensors to recognize signs, e.g.: Data gloves (Shukor et al., 2015), power gloves (Mohandes et al., 2004), cyber gloves (Mohandes, 2013), and Dexterous master gloves (Hoshino, 2006). Sadek et al. (2017) designed a smart glove using a few sensors depending on a statistical analysis of the anatomical shape of the hands when performing the 1300 words of the Arabic sign language (ArSL). The glove costs about \$65 which is according to the author 5% less than the cost of commercial smart gloves. The high cost and less normality of this method contributes to the appearance of the image-based way, where one or more cameras are employed to capture the signs. Classification can be done either by marker-based or visual-based techniques.

In marker-based techniques, markers with predefined color or colored gloves are placed on the fingertips and wrist. These

* Corresponding author at: Department of Computer Science, Faculty of Computers and Informatics, Benha University, Benha Mansoura Road, next to Holding Company for Water Supply and Sanitation Benha, Qalyubia Governorate, Egypt.

E-mail addresses: Nada.Ahmed@fci.bu.edu.eg (N.B. Ibrahim), Selimm@bu.edu.eg (M.M. Selim), Hala.zayed@fci.bu.edu.eg (H.H. Zayed).

Peer review under responsibility of King Saud University.

Production and hosting by Elsevier

https://doi.org/10.1016/j.jksuci.2017.09.007
1319-1578/

predefined colors are then detected and segmented from an image captured by a 2D camera using image processing methods, but these techniques also lack for normality (Wang and Popović, 2009; El-Bendary et al., 2010). On the other hand, visual-based techniques use bare hands without any markers. These techniques have high normality and higher mobility than any other types of the SLRS. Visual -based SLRS have low cost as one camera can be used. But these techniques suffer from changing in illumination. Hand occlusion with either each other or the face is another drawback as 2D images lack the depth information that aid in solving occlusion. This paves the way to the depth sensors that depends on RGB-D image technique giving the depth of each pixel in the image helping in constructing a 3D model of the objects in the scene. Till now it's still an open field of research. In most of the researches vision-based refers to visual-based vision system. A further discussion and detailed overview on related work in the field of SLR is given at (Cooper et al., 2011; Mohandes et al., 2014; Rautaray and Agrawal, 2015; Agrawal et al., 2016). This paper will focus on ArSL. Recent isolated vision ArSLR systems are pointed out.

Al-Rousan et al. (2009) developed a system that automatically recognizes 30 isolated ArSL words using discrete cosine transform (DCT) to extract features and Hidden Markov Model (HMM) as recognition method. The system obtained a word recognition rate of 94.2% for signer-independent off-line mode. Due to the nature of DCT, the observation features produced by the DCT algorithm misclassified similar gestures. Also, the system did not concern with working out the occlusion problem.

To overcome the misclassification of similar gesture, Al-Rousan et al. (2010) developed a system that used two-level scheme of HMM classifier. The system overcomes the occlusion state by treating the occluded objects as one object or by taking the preceding features of the objects before occlusion. In the real situation, this is not the case.

Another technique for solving the occlusion states was developed by El-Jaber et al. (2010) where a stereo vision is applied to estimate and segment out the signer's body using its depth information to recognize 23 isolated gestures in signer-dependent mode. It aches from its high cost as more than one camera is needed to construct the stereo vision. Disparity maps are computationally expensive as any change in the distance between the two cameras and the object will affect the performance of solving the correspondence problem.

In Elons et al. (2013), a 3D model of the hand posture is generated from two 2D images from two perspectives that are weighted and linearly combined to produce single 3D features trying to classify 50 isolated ArsL words using Hybrid pulse-coupled neural network (PCNN) as feature generator technique followed by non-deterministic finite automaton (NFA). Then, "Best-match" algorithm is used to find the most probable meaning of a gesture. The recognition accuracy reaches 96%. The misclassification comes from the fact that the NFA of some gestures may be wholly included in another gesture NFA.

Ahmed and Aly (2014) uses a combination of local binary patterns (LBP) and principal component analysis (PCA) to extract features that are fed into a HMM to recognize a lexical of 23 isolated ArSL words. Occlusion is not resolved as any occlusion state is handled as one object and recognition goes on. The system achieves a recognition rate of 99.97% in signer- dependent mode. But LBP may not work properly on the areas of constant gray-level because of the thresholding schemes of the operator (Ahmed and Aly, 2014).

Obviously, most vision systems suffer from two main problems: confusing similar gestures in motion, and curing the occlusion problem. The aim of the research documented in this paper is to decrease the misclassification rate for similar gestures and resolves all occlusion states using only one camera and without any complicated environment to compute the disparity map.

This paper presents an automatic visual SLRS that translates isolated Arabic word signs into text. The proposed system has four primary stages: hand segmentation, tracking, feature extraction and classification. Hand segmentation is performed utilizing a dynamic skin detector based on the color of the face (Ibrahim et al., 2012). Then, the segmented skin blobs are used in identifying and tracking of the hands with the help of the head. Geometric features of the hands are employed to formulate the feature vector. Finally, Euclidean distance classifier is applied for classification stage. A dataset of 30 isolated words used in the daily school life of the hearing impaired children was developed. The experimental results indicate that the proposed system has a recognition rate of 97%. Taking into consideration that 83% of the words mainly cover all the occlusion states to prove the robustness of the system.

The upcoming sections are arranged as follows: Dataset description is illustrated in Section 2. The proposed approach including a novel identifying and tracking method is described in Section 3. Results and evaluation are outlined in Section 4. Finally, a conclusion is given in Section 5.

2. ArSLRS dataset

A unified Arabic sign language dictionary was published in two editions in 2008. Despite of that, there are no common databases available for researchers in the area of Arabic sign language recognition. Thus, each researcher has to establish his own database with reasonable size.

The dataset used is an ArSL database videos which was collected at Benha University. The database consists of 450 colored ArSL videos captured at a rate of 30 fps. These videos represent 30 Arabic words which were selected as the daily common used words in school. 300 videos are used for training while 150 are for testing. The signers performing the testing clips are different from whom performed the training clips to guarantee the signer -independency of the system designed. The videos are gathered in different illumination, backgrounds, and clothing. The signer is asked to face the camera with no orientation, then starts signing from silence state where both hands are placed beside the body and then ends again with a silence state.

It was considered that the database contains words that have variety of using one hand or both hands with occlusion with each other or with the face to test the validity of the system in solving different occlusion states.

The list of the used words and their description is given in the Table 1. The occlusion column identifies that the sign performed has an occlusion state with either one of the hands or both hands and the face. RH and LH columns show whether the sign is executed with the right hand or left hand, respectively. R-L H column indicates that the sign is performed with both hands. The last row illustrates the estimated percentage of occlusion states in the built database.

3. The proposed ArSLRS

As shown in Fig. 1, the vision-based SLRS has two modes. The first mode is from the hearing-impaired people to the vocal people where a video of the sign language (SL) is translated into oral language either in the form of text or voice. This mode is called vision-based SLRS. On the other hand, the second mode is from vocal people to the hearing-impaired people where the oral language voice record is converted into SL video. The vision-based SLRS mode was the interest in this paper. Each stage is illustrated in detail in the upcoming sub-section.

Table 1
List of dataset words and their description.

Words	Occlusion	R-L H	LH	RH
Peace be upon you	√			√
Thank you	√		√	
Telephone				√
I			√	
Eat	√		√	
Sleep	√		√	
Drink	√		√	
Prayer	√	√		
To go	√			√
Bathroom	√			√
Ablution	√	√		
Tomorrow				√
Today	√		√	
Food	√	√		
Water	√			√
To love	√	√		
To hate	√			√
Money	√			√
Where're you going?	√	√		
Where	√	√		
Why	√	√		
How much			√	
Yes	√			√
No	√		√	
Want				√
School	√			√
Teacher	√	√		
Help	√	√		
Sick	√		√	√
Friend	√	√		
Percentage	83%	33%	27%	40%

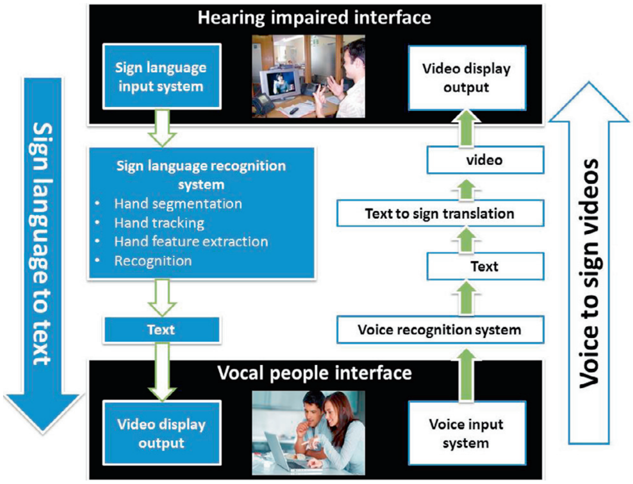


Fig. 1. A diagram of vision-based SLRS.

3.1. Hand segmentation

This term refers to the extraction of hands from the frames through the entire video sequence. The video sequence may contain just the hands or the whole body of the signer. In the first instance, either a background removing technique or skin detection technique is employed to segment hands. Merely in the second case, background removing technique followed by skin detection may be used or a skin detection technique is applied directly to the frames. Accumulated difference image (AD) is applied to extract the hands if they were the only moving object in the video (Assaleh et al., 2010).

There has appeared many skin detectors that rely on the face to detect the skin regions. In Kawulok (2008), a face region that includes the eyes and the mouth which are non-skin non-smooth regions is used to calculate the probability of a pixel to be skin or non-skin pixel. This has affected the results of this approach by detecting the mouth, the eyes and the brows as skin region which is not true. In Bilal et al. (2015), a $10 * 10$ window around the center pixel of the face is used to distinguish the skin tone pixels, which in most of the cases is the nose tip. But, this region suffers from the effect of illumination and may give wrong indications.

In this paper, a dynamic skin detector based on face skin tone color is used in segmenting the hands (Ibrahim et al., 2012). YCbCr color space is used after discarding the luminance channel. Face detector is applied to the first frame. The probability distribution function (PDF) histogram bins are calculated and trimmed at 0.005. To avoid eyes and mouth regions to be recognized as the skin, a threshold is applied to remaining PDF values after trimming. The pixels along the major and minor axes of the bounding rectangle of detecting face are used to calculate a dynamic threshold. This threshold is applied to the face image to identify skin pixels. And then, the threshold is updated by increasing the pixels around the axes until 95% from face pixels are recognized as a skin. Finally, this threshold is applied to the entire image. This method is employed due to its adaptive nature which make it applicable for different races. In addition to, using the YCbCr color space reduce dramatically the effect of illumination on the segmentation. The outcome of this phase is a binary image that holds the hands and the face with white pixels and other objects with dark.

3.2. Tracking

Tracking is defined as the problem of estimating the trajectory of an object in the image plane as it moves around a scene (Yilmaz et al., 2006). Numerous approaches for tracking have been proposed. Some of these approaches are: detecting motion with an active camera (Lee et al., 2012), skin blob tracking (Zaki and Shaheen, 2011), active contour (Holden et al., 2005), camshift (Li et al., 2011), particle filter (Gianni et al., 2007) and Kalman filter (Asaari and Suandi, 2010). A review study on recent advances and trends in tracking is given in Yang et al. (2011), Baskaran and Subban (2014).

Dreuw et al. (2006) developed a dynamic programming tracking (DPT) technique that relies on two paths to decide the correct tracking path for the hands. A forward path is used to calculate an overall score function for all the frames of a sequence. A backward path that went from the final frame is applied to compute the best route for the tracked hand. The overall scoring function was utilized in calculating the best path with respect to a specific score function. This technique is a model-dependent and a signer-independent technique. Taking the tracking decision at the end of the sequence improves the ability of the DPT algorithm to prevent wrong local decisions. By combining this approach with Viola and Jones method for tracking (Viola and Jones, 2004) the results were improved to reach 0% tracking error rate (TER) at tolerance $(\tau) = 20$ (SIGNSPEAK, 2012). Only the two-path method needs a bunch of computations and the score functions needs some modifications to make the required results.

A proposed technique that relies on tracking skin blobs by using the Euclidean distance between the skin blobs in two consecutive frames is developed to tackle using two paths and using scoring functions. The Viola and Jones method first identify the head. Then, hands are identified by the distance between their centers and the centroid of the head. Euclidean distance is used to keep track of the head and the hands. An occlusion is detected when two or more tracked objects pointed to the same skin blob. Resolving of occlusion states depends on computing the deviation in elevation between the former and the current positions of the head and the hands. In this technique, it is estimated that the hand shape changes are small. A translation of the former position of the head and the hands is done to occupy the bounding rectangle of the occluded objects. The TER at $\tau = 20$ is 0.08%. This technique is signer-independent and model-free technique. This technique uses forward tracking path in addition to the preceding information about the tracking object to decide its next location.

3.2.1. Head tracking

The head can be easily localized by using a cascade boosting algorithm (Viola and Jones, 2004) but it is computationally very expensive to apply this algorithm on all frames to detect head especially if this application is a real-time one. Consequently, the cascade boosting algorithm is applied to the first frame only to obtain the bounding rectangle of the head. Since the position of the head during the signing process mainly doesn't change and nearly has the same location, Euclidean distance applies to the preceding frames to recognize the head skin blob. When more than one skin blob appears, the head is distinguished as the skin blob with the smallest Euclidean distance from the former position of the head. Let $H_p = (x_p, y_p)$ is the center of the previous head bounding rectangle while $B_i = (x_i, y_i)$ is the center of the current skin blobs where $i = \{1, 2, 3\}$. The Euclidean distance (ξ_{HB}) between the H_p and B_i is given by:

$$(\xi_{HB_i}) = \sqrt{(x_p - x_i)^2 + (y_p - y_i)^2} \tag{1}$$

The skin blob with the minimum ξ_{HB} is the current head (H_c). As shown in Fig. 2a, the head is marked with solid rectangle. In Fig. 2b, the previous head bounding rectangle is marked with a dotted rectangle, while the new skin blobs are marked with solid rectangles. Euclidean distances between the center of the previous location of the head and the center of the current skin blobs are calculated. The blob with the minimum Euclidean distance is recognized and marked as the new head with a solid rectangle as shown in Fig. 2c.

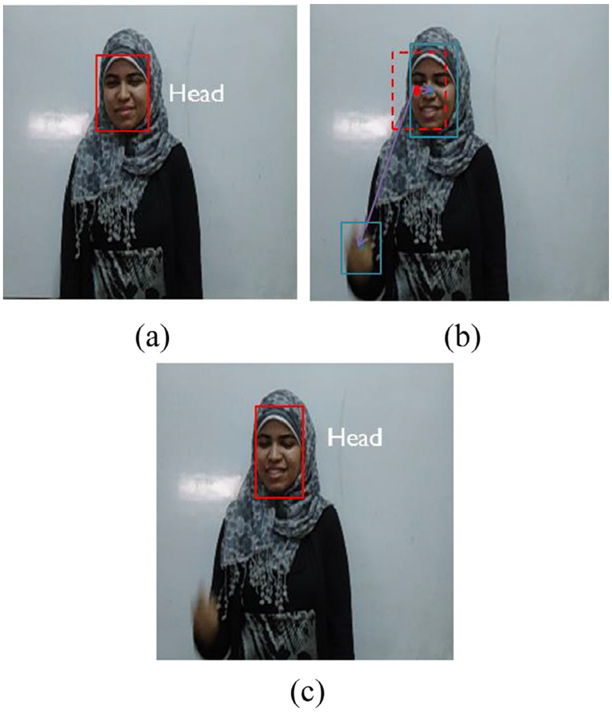


Fig. 2. Head tracking.

3.2.2. Hands tracking

The center of the bounding rectangle of the head can be used as a reference point to define the hands. Let B be a skin blob. To identify the skin blob as right hand (RH) or left hand (LH), the difference $(\triangle x)$ between the x-coordinates of the centers of the current head (H_c) and the skin blob (B) must be calculated as follows:

$$\triangle x = x_{H_c} - x_B \tag{2}$$

Then, the skin blob is identified according to the following conditions:

$$B = \begin{cases} RH, & \triangle x > 0 \\ LH, & Otherwise \end{cases} \tag{3}$$

Identifying of the right and the left-hand skin blobs is shown in Fig. 3.

After localizing the first appearance of the head and the hands, Euclidean distance is used to keep track of them, as shown in Fig. 4. This will work well till an occlusion takes place.

*Occlusion resolving*Occlusion is the overlapping of one or more of the tracked objects where one object may cover some or whole

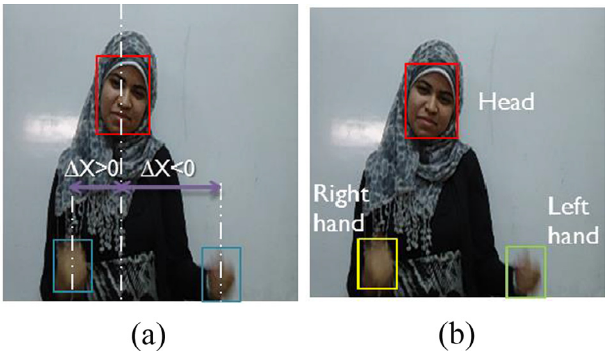


Fig. 3. Identifying the first appearance of the right and the left hands.

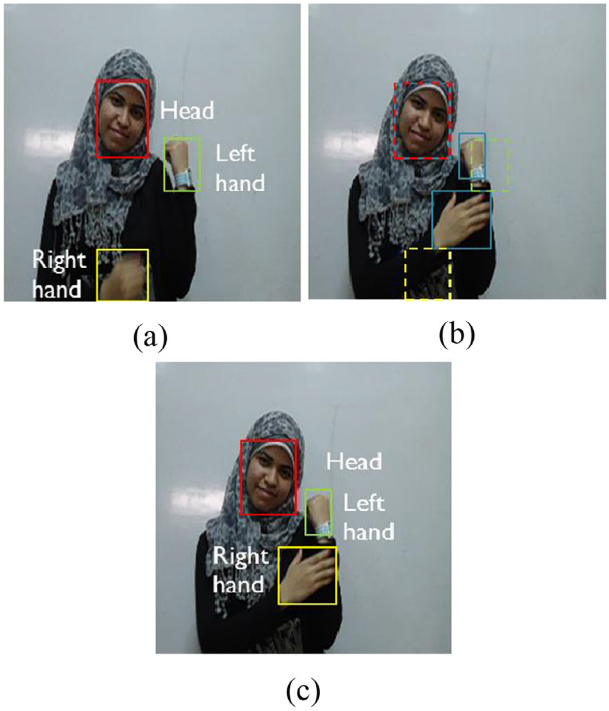


Fig. 4. Hands tracking.

of the other object. Hands segmentation in the occlusion situation is a challenging task. The hand shape change was estimated to be small where capturing of the signs was acquired on high-speed frame recording.

In this paper, an occlusion resolving technique is proposed. This technique splits the problem into two sub-problems. The first is the occlusion of the head with one or both hands which is the general case, while the second one is the occlusion between the two hands only.

Occlusion with head The head plays an important role where it is considered as a reference point and also as an indicator for occlusion. If the area of the head increases by nearly the third, this is an occlusion state. If the Euclidean distance calculated for the head and one or both of the hands by using Eq. 1 labeled the same skin blob as the head and one or both of the hands then this is an occlusion situation.

Any bounding rectangle of an object has four corners: right upper (RU), right lower (RL), left upper (LU) and left lower (LL) as illustrated in Fig. 5. The proposed algorithm uses the corners to identify the occluded objects location.

Occlusion between head and right hand can be resolved by calculating the difference between the y-coordinates (Δy) of the RU corners of the current skin blob (B) bounding rectangle and the previous right hand (PRH) bounding rectangle. (Δy) can be calculated as follows:

$$\triangle y = y_{RU_B} - y_{RU_{PRH}} \tag{4}$$

Then,

$$if\,\triangle y = \begin{cases} \geqslant 0 \;\rightarrow\; \text{move } RU_{PRH} \text{ to } RU_B \text{ and move } LL_{H_P} \text{ to } LL_B \\ < 0 \;\rightarrow\; \text{move } RL_{PRH} \text{ to } RL_B \text{ and move } LU_{H_P} \text{ to } LU_B \end{cases} \tag{5}$$

On the other hand, for the occlusion of the head and the left hand, the difference between the y-coordinates of the RU corners of the current skin blob (B) bounding rectangle and the previous left hand (PLH) bounding rectangle are calculated (Δy) as follows:

$$\triangle y = y_{RU_B} - y_{RU_{PLH}} \tag{6}$$

Then,

$$if\,\triangle y = \begin{cases} \geqslant 0 \;\rightarrow\; \text{move } LU_{PRH} \text{ to } LU_B \text{ and move } RL_{H_P} \text{ to } RL_B \\ < 0 \;\rightarrow\; \text{move } LL_{PRH} \text{ to } LL_B \text{ and move } RU_{H_P} \text{ to } RU_B \end{cases} \tag{7}$$

The case of occlusion between the head and the left hand and how to resolve it is shown in detail in Fig. 6. In Fig. 6a, the head and the left hand are marked with solid rectangle. Then, Euclidean distance between the current skin blobs and the previous head and left hand is calculated, as shown in Fig. 6b. These calculations indicate that the head and the left hand share the same skin blob, as illustrated in Fig. 6c. Δy is calculated and the arrow shown in Fig. 6d indicates the translation of the head and the left hand from the previous locations to the new locations. Finally, the occlusion is resolved and the location of the new head and hand is in Fig. 6e. Finally, occlusion between the head and both hands is solved by calculating the Euclidean distance between both hands and the head. If this distance is less than a predefined threshold, then remove occlusion as mentioned previously for both hands with keeping the head position as its previous location. On the other hand, if the distance is greater than the predefined threshold, then remove this hand and resolve the occlusion using the previous methods for the remaining hand and the head.

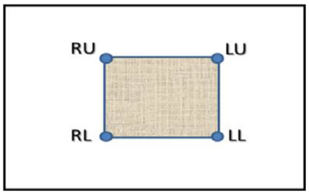


Fig. 5. Corners of the bounding rectangle of an object.

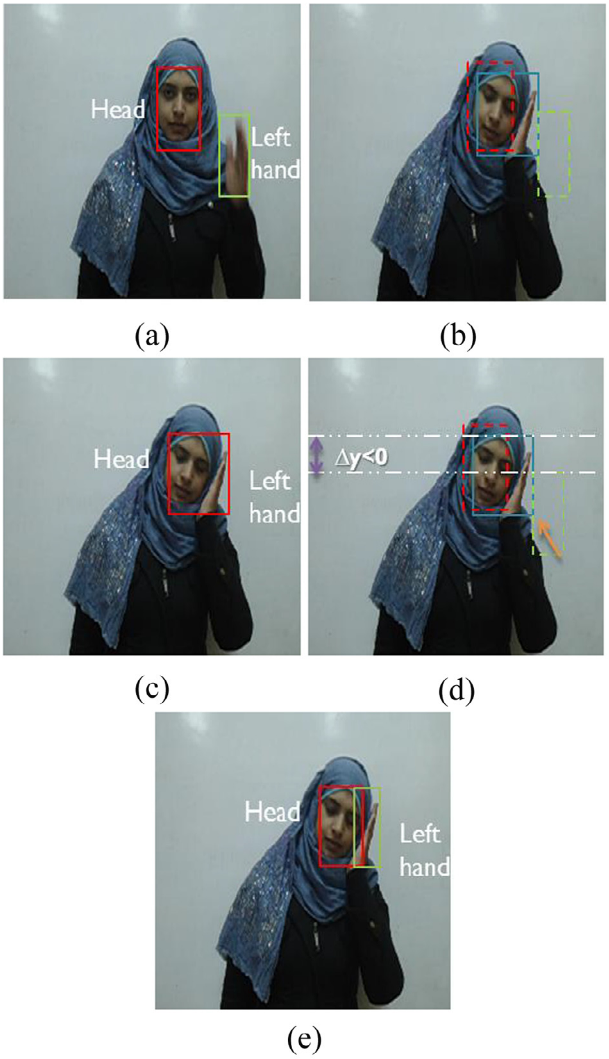


Fig. 6. Hands tracking.

Occlusion between hands The partial occlusion of the hand is indicated by its area increase by one half. If the occlusion takes place, then solve it as if it was an occlusion between the hands and the head.

Fig. 7a is the frame that contains the head and both hands before occlusion. In Fig. 7b the hands have moved and its region has increased by more than the half which indicates an occlusion situation. The head area doesn't increase by more than the third; therefore, the occlusion happened between hands only. As shown in Fig. 7c the head is identified and the other skin blob is recognized as the hands. Δy is calculated for hands as shown in Fig. 7d using Eq. 4 and Eq. 6. The arrow in Fig. 7d shows the movement of the previous bounding rectangles of both hands. The RU of the right hand is moved to the RU of the current skin blob while the LL of the left hand is moved to the LL of the current skin blob. The result of the tracking is shown in Fig. 7e.

For full occlusions between both hands, it is indicated when the previous location of both hand points to the same skin blob as the

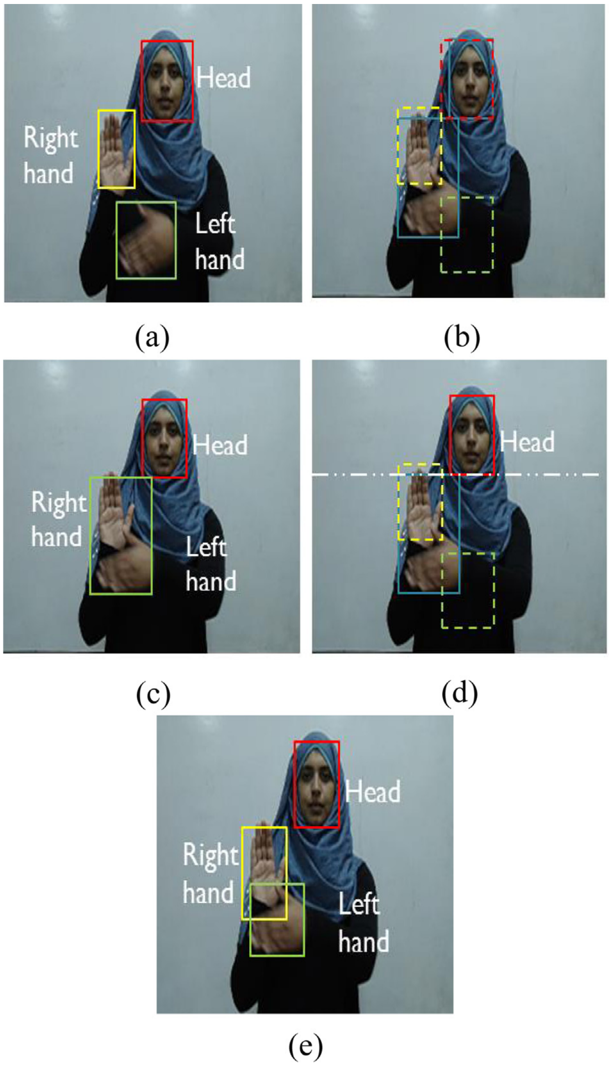


Fig. 7. Resolving partial occlusion between two hands.

nearest one. If the Euclidean distance between each of the previous hands locations and the new skin blob is greater than a threshold value, then there is a full occlusion between the two hands. This occlusion is solved by labeling the skin blob as the new position of both hands. If one of the hands has its Euclidean distance less than the threshold, then this hand is out of the scene and the skin blob is labeled as the other hand only.

3.3. Hand feature extraction

The next step is extracting hands features. Extracting good features leads to a significant increase in the performance of the SLRS. The features fall in two domains: temporal domain and spatial domain (Al-Rousan et al., 2009).

Temporal domain is sometimes referred as frequency domain where the frame is converted to one of its frequency domain transformation forms. While the spatial domain is based on direct manipulation of the pixels in the image that is divided into two categories: geometric feature and statistical features (El-Jaber et al., 2010). Geometric features describe the two-dimensional projection of the hand in the image while statistical features describe the statistical properties of the hand shape.

In this paper, Geometric features of the spatial domain are used. The chosen features include: coordinates of the center of gravity of the hands, the velocity of the hand movement and the orientation of the main axis of the hand. The feature vector of any sign is represented as follows:

$$\text{Feature vector} = \{x_{RH}, y_{RH}, v_{RH}, \phi_{RH}, x_{LH}, y_{LH}, v_{LH}, \phi_{LH}\} \quad (8)$$

where x , y , v and ϕ are the coordinates of the gravity of the hands, the velocity of the movement of the hands and the orientation of the main axis of the hands, respectively.

3.4. Recognition

The most used recognition techniques are: Hidden Markov Model (HMM), support vector machine (SVM), artificial neural networks (ANN), adaptive Neuro-fuzzy inference system (ANFIS) and Euclidean distance. In this study, the dataset is not overly big, so HMM, ANN, SVM and ANFIS are not used as there are no decent data for training. Euclidean distance is used for classification as it acts to compare directly the feature vectors.

Let the feature vector of the original sign is $v_o = \{x_1, x_2, x_3, \ldots\}$ and the feature vector of the testing sign is $v_t = \{y_1, y_2, y_3, \ldots\}$ then, the Euclidean distance (ED) for the feature vector is computed utilizing the following equation:

$$ED = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + (x_3 - y_3)^2 + \ldots} \quad (9)$$

4. Results and evaluation

Three scenarios have been followed to evaluate the proposed system. The first is to understand the effect of changing skin color and illumination on correctly segmenting the hands. The proposed dynamic mid-way threshold histogram skin detector was evaluated in our previous work (Ibrahim et al., 2012). This evaluation proposed that this detector turns down the False positive rate (FPR) by nearly half while holding on the False negative rate (FNR) approximately the same. It also diminishes the number of pixels to deal with to be about 52% of the entire face which dramatically decreases the detection time. Ultimately, it is recommended for real-time applications and is applicable for different races due to its adaptive dynamic nature. On the other hand, using the YCbCr color space decrease the effect of illumination. Nevertheless, the light must be uniform and the skin tone of the face and the hands must be the same.

The second scenario is to investigate the performance of head and hand tracking algorithms. To achieve this, a data with ground-truth annotations is required as well as an evaluation measure.

For an image sequence $X_1^T = X_1, \ldots, X_T$ and corresponding annotated object positions $u_1^T = u_1, \ldots, u_T$ the tracking error rate (TER) of tracked positions $\hat{u}_1^T$ is defined as the relative number of frames where the Euclidean distance between the tracker and the annotated position is larger than or equal to a tolerance (τ) (Dreuw et al., 2006):

$$TER = \frac{1}{T}\sum \delta_\tau(u_t, \hat{u}_t) \text{ with } \delta_\tau(u_t, v_t) = \begin{cases} 0, \|u - v\| < \tau \\ 1, \text{Otherwise} \end{cases} \quad (10)$$

A RWTH-BOSTON-104 tracking benchmark database (Dreuw et al., 2010) for video-based sign language recognition is applied to assess the proposed tracking technique. The ground truth of the head and the hands' position are annotated to evaluate the tracking technique. The sequences have a lot of dynamic variations in movement and have most of the occlusion cases to be tested.

Different methods are applied to RWTH-BOSTON-104 database (SIGNSPEAK, 2012) like dynamic programming tracking (DPT), principle component analysis (PCA), Viola and Jones (VJ), active

Table 2
TER for the proposed tracking technique and other state-of-the-art tracking methods from SIGNSPEAK (2012).

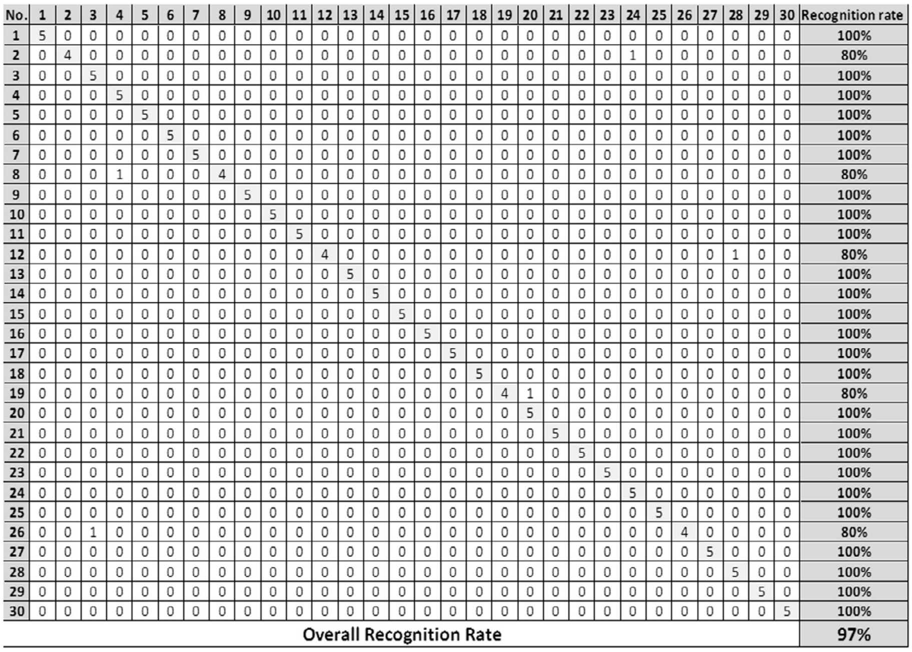
	TER%			
Tracking methods	$\tau = 5$	$\tau = 10$	$\tau = 15$	$\tau = 20$
DPT + PCA	26.77	17.32	12.7	10.86
DPT + VJ	10.06	0.4	0.02	0
VJD	9.75	1.23	1.09	1.07
VJT	10.04	0.81	0.73	0.68
FJAAM	10.17	6.85	6.82	6.81
FJAAM	10.92	7.92	7.88	7.76
POICAAM	3.54	0.12	0.08	0.08
Proposed	0.97	0.70	0.22	0.08

appearance model (AAM), Project-Out Inverse Compositional AAM (POICAAM) and Fixed Jacobian active appearance model (FJAAM). Most of these methods are model-based signer-dependent tracking approaches and have been evaluated on all 15732 ground-truth annotated frames of the RWTH-BOSTON-104 dataset. The results of these different algorithms are compared to the results of the proposed technique for evaluation. The proposed technique has the lowest TER at $\tau = 5$, as shown in Table 2. By increasing the τ, the TER decrease, but the proposed technique has small changes unlike other algorithms. This demonstrates the robustness of the proposed occlusion resolving technique as it can accurately specify the position of the hands and the head with an improvement of 25.7% compared to the result from POICAAM at $\tau = 5$. The proposed technique is not a model-based technique compared to other methods which guarantee the less computation needed and the signer independency of the method. Integrating the proposed technique with other methods may improve the accuracy, especially, for the tolerance $10 \leqslant \tau \leqslant 20$. The third scenario is evaluating the whole system by considering the percentage of the total number of Arabic sign words that were correctly recognized. Euclidean distance is used to classify the video of each gesture. The signer of the tested gesture is asked to face the camera with no orientation and freely perform the sign (remember that signer must begin and end with a silence state). The environment is controlled as one signer at a time is performing with stationary environment around him. The system attains a recognition rate of 97% in signer independent mode. Form the confusion matrix illustrated in Fig. 8, it was indicated that gestures (2, 8, 12, 19 and 26) confused with the gestures (24, 4, 28, 20 and 3), respectively. These gestures have great similarity in either partial or full hand movement. Despite that, only one gesture is recognized wrongly. This indicates the ability of the proposed ArSLRS to differentiate between gestures with high similarity.

Finally, it is clear that the proposed system has no demand for more than one camera or complicated calculations to adjust the two cameras to achieve high recognition rates as in El-Jaber et al. (2010). The proposed system did not have to group similar gesture to increase its recognition rate as in AL-Rousan et al. (2007) where gesture that has common parts decrease dramatically its rate. The developed system does not construct a 3D model of hand posture that need two cameras and a sensitive calculation to weight the two views from both cameras. The system proves its robust occlusion resolving technique that outperform other methods.

5. Conclusions

This paper presents an automatic visual SLRS that translates isolated Arabic word signs into text. The proposed system is signer-independent system that utilizes a single camera and the signer does not employ any type of gloves or markers. The system has four primary stages: hand segmentation, hand tracking, hand feature extraction and classification. Hand segmentation is performed utilizing a dynamic skin detector based on the color of the face. Then, the segmented skin blobs are used to identify and track hands with the aid of the head. The system proved its robust performance against all states of occlusion as 83% of the words in the dataset has different occlusion states. Geometric features are employed to construct the feature vector. Finally, Euclidean distance classifier is applied to classification stage. A dataset of 30 isolated words that are utilized in the daily school life of the hearing-impaired children was developed. The experimental results indicate that the proposed system has a recognition rate of 97% with the low misclassification rate for similar gestures. In addition, the system proved its robustness against different cases of occlusion with a minimum TER of 0.08%.



No.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	Recognition rate
1	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100%
2	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	80%
3	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100%
4	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100%
5	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100%
6	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100%
7	0	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100%
8	0	0	0	1	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	80%
9	0	0	0	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100%
10	0	0	0	0	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100%
11	0	0	0	0	0	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100%
12	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	80%
13	0	0	0	0	0	0	0	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100%
14	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100%
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100%
16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100%
17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	100%
18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	100%
19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	1	0	0	0	0	0	0	0	0	0	0	80%
20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0	100%
21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0	100%
22	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	0	0	0	0	0	0	0	100%
23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	0	0	0	0	0	0	100%
24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	0	0	0	0	0	100%
25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	0	0	0	0	100%
26	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	80%
27	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	0	0	100%
28	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	0	100%
29	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	100%
30	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	100%
Overall Recognition Rate																															97%

Fig. 8. Confusion matrix of the proposed system.

References

Agrawal, S.C., Jalal, A.S., Tripathi, R.K., 2016. A survey on manual and non-manual sign language recognition for isolated and continuous sign. Int. J. Appl. Pattern Recognit. 3 (2), 99–134.

Ahmed, A.A., Aly, S., 2014. Appearance-based arabic sign language recognition using hidden markov models. In: Engineering and Technology (ICET), 2014 International Conference on, IEEE, 2014, pp. 1–6.

AL-Rousan, M., Al-Jarrah, O., Nayef, N., 2007. Neural networks based recognition system for isolated arabic sign language. In: Information Technology, 2007. ICIT 2007. The 3rd International Conference on, AL-Zaytoonah University.

Al-Rousan, M., Assaleh, K., Tala'a, A., 2009. Video-based signer-independent arabic sign language recognition using hidden markov models. Appl. Soft Comput. 9 (3), 990–999.

Al-Rousan, M., Al-Jarrah, O., Al-Hammouri, M., 2010. Recognition of dynamic gestures in arabic sign language using two stages hierarchical scheme. Int. J. Knowl.-Based Intell. Eng. Syst. 14 (3), 139–152.

Asaari, M.S.M., Suandi, S.A., 2010. Hand gesture tracking system using adaptive kalman filter. In: Intelligent systems design and applications (ISDA), 2010 10th international conference on, IEEE, 2010, pp. 166–171.

Assaleh, K., Shanableh, T., Fanaswala, M., Amin, F., Bajaj, H., et al., 2010. Continuous arabic sign language recognition in user dependent mode. JILSA 2 (1), 19–27.

Baskaran, J., Subban, R., 2014. Compressive object tracking–a review and analysis. In: Computational Intelligence and Computing Research (ICCIC), 2014 IEEE International Conference on, IEEE, 2014, pp. 1–7.

Bilal, S., Akmeliawati, R., Salami, M.J.E., Shafie, A.A., 2015. Dynamic approach for real-time skin detection. J. Real-Time Image Proc. 10 (2), 371–385.

Center, M., 2017. Deafness and hearing loss Tech. rep.. World Health Organization.

Cooper, H., Holt, B., Bowden, R., 2011. Sign language recognition. In: Visual Analysis of Humans. Springer, pp. 539–562.

Dreuw, P., Deselaers, T., Rybach, D., Keysers, D., Ney, H., 2006. Tracking using dynamic programming for appearance-based sign language recognition. In: Automatic Face and Gesture Recognition, 2006. FGR 2006. 7th International Conference on, IEEE, 2006, pp. 293–298.

Dreuw, P., Forster, J., Ney, H., 2010. Tracking benchmark databases for video-based sign language recognition. In: ECCV Workshops (1), 2010, pp. 286–297.

El-Bendary, N., Zawbaa, H.M., Daoud, M.S., Hassanien, A.E., Nakamatsu, K., 2010. Arslat: Arabic sign language alphabets translator. In: Computer Information Systems and Industrial Management Applications (CISIM), 2010 International Conference on, IEEE, 2010, pp. 590–595.

El-Jaber, M., Assaleh, K., Shanableh, T., 2010. Enhanced user-dependent recognition of arabic sign language via disparity images. In: Mechatronics and its Applications (ISMA), 2010 7th International Symposium on, IEEE, 2010, pp. 1–4.

Elons, A.S., Abull-Ela, M., Tolba, M.F., 2013. A proposed PCNN features quality optimization technique for pose-invariant 3d arabic sign language recognition. Appl. Soft Comput. 13 (4), 1646–1660.

Gianni, F., Collet, C., Dalle, P., 2007. Robust tracking for processing of videos of communication's gestures. In: International Gesture Workshop, Springer, 2007, pp. 93–101.

Holden, E.-J., Lee, G., Owens, R., 2005. Australian sign language recognition. Mach. Vis. Appl. 16 (5), 312.

Hoshino, K., 2006. Dexterous robot hand control with data glove by human imitation. IEICE Trans. Inf. Syst. 89 (6), 1820–1825.

Ibrahim, N.B., Selim, M.M., Zayed, H.H., 2012. A dynamic skin detector based on face skin tone color. In: Informatics and Systems (INFOS), 2012 8th International Conference on, IEEE, 2012, pp. MM–1.

Kawulok, M., 2008. Dynamic skin detection in color images for sign language recognition. Image Signal Process., 112–119

Lee, C.-Y., Lin, S.-J., Lee, C.-W., Yang, C.-S., 2012. An efficient continuous tracking system in real-time surveillance application. J. Network Comput. Appl. 35 (3), 1067–1073.

Li, Y.-B., Shen, X.-L., Bei, S.-S., 2011. Real-time tracking method for moving target based on an improved camshift algorithm. In: Mechatronic Science, Electric Engineering and Computer (MEC), 2011 International Conference on, IEEE, 2011, pp. 978–981.

Mohandes, M.A., 2013. Recognition of two-handed arabic signs using the cyberglove. Arabian J. Sci. Eng., 1–9

Mohandes, M., A-Buraiky, S., Halawani, T., Al-Baiyat, S., 2004. Automation of the arabic sign language recognition. In: Information and Communication Technologies: From Theory to Applications. In: Proceedings. 2004 International Conference on, IEEE, 2004, pp. 479–480.

Mohandes, M., Deriche, M., Liu, J., 2014. Image-based and sensor-based approaches to arabic sign language recognition. IEEE Trans. Hum.-Mach. Syst. 44 (4), 551–557.

Rautaray, S.S., Agrawal, A., 2015. Vision based hand gesture recognition for human computer interaction: a survey. Artif. Intell. Rev. 43 (1), 1–54.

Sadek, M.I., Mikhael, M.N., Mansour, H.A., 2017. A new approach for designing a smart glove for arabic sign language recognition system based on the statistical analysis of the sign language. In: Radio Science Conference (NRSC), 2017 34th National, IEEE, 2017, pp. 380–388.

Shukor, A.Z., Miskon, M.F., Jamaluddin, M.H., bin Ali, F., Asyraf, M.F., bin Bahar, M.B., et al., 2015. A new data glove approach for malaysian sign language detection. Proc. Comput. Sci. 76, 60–67.

SIGNSPEAK, 2012. Scientific understanding and vision-based technological development for continuous sign language recognition and translation, Tech. rep., European Commission.

Viola, P., Jones, M.J., 2004. Robust real-time face detection. Int. J. Comput. Vision 57 (2), 137–154.

Wang, R.Y., Popović, J., 2009. Real-time hand-tracking with a color glove. ACM Transactions on Graphics (TOG), vol. 28. ACM, p. 63.

Yang, H., Shao, L., Zheng, F., Wang, L., Song, Z., 2011. Recent advances and trends in visual tracking: A review. Neurocomputing 74 (18), 3823–3831.

Yilmaz, A., Javed, O., Shah, M., 2006. Object tracking: a survey. ACM Computing Surveys (CSUR) 38 (4), 13.

Zaki, M.M., Shaheen, S.I., 2011. Sign language recognition using a combination of new vision based features. Pattern Recogn. Lett. 32 (4), 572–577.