ISSN 2007-9737

Gesture Recognition System For Isolated Word Sign Language Based On key-Point Trajectory Matrix

Sana Fakhfakh, Yousra Ben Jemaa

El Manar University, Signal and System Research Unit, ENIT, Tunisia

sana.fakhfakh@enis.tn, yousra.benjemaa@enis.rnu.tn

Abstract. This paper suggests a new system to help the deaf and the hearing-impaired community improve their connection with the hearing world and communicate freely. The most important thing in this system is how to help the users be free and finally have a more natural way of communication. For this reason, we present a new process based on two levels: a static-level aiming to extract the most head/hands key points and a dynamic-level with the objective of accumulating the key-point trajectory matrix. Also our proposed approach takes into account the signer-independence constraint. A SIGNUM database is applied in the classification stage and our system performances have improved with a 94.3% recognition rate. Furthermore, a reduction in time processing is obtained when the removing of redundant frame step is applied. The obtained results prove the superiority of our system compared to the state-of- the-art methods in terms of recognition rate and execution time.

Keywords. Sign language recognition, isolated word, signer-independence, particular filter, key-point trajectory matrix.

1 Introduction

A sign language is the natural communication tool between different hearing impaired people. In recent years, a sign language recognition system has become an active area of research. Many systems are proposed in order to facilitate communication between the hearing and the deaf communities [12, 32, 33, 20, 36, 14, 17].

The main goal of these systems is to offer an automatic system able to recognize the different signs used by the deaf community. These systems support isolated or continuous signs as presented in Fig. 1 [10, 43, 34]. In isolated systems, just one sign gesture is presented. In continuous signs, a complete clause sign is presented. In this paper, we treat only isolated words. The researchers' attention was generally focused on the powerful extraction of non-manual and manual features [2, 16]. Manual features are generally related to hand gestures which are composed of different features like hand motion, hand location, hand orientation and hand shape. Non-manual features like mouth movement, facial expressions, or body posture can also give an additional meaning to the sign. However, most signs are presented only with manual features. Hence our focus is solely on manual signing.

The goal of the sign recognition task can be achieved via two varieties of data acquisition: sensor-based [26, 32, 33, 27] or vision-based systems [42, 41, 13]. Sensor-based methods usually use different specialized equipment. Vision-based methods are based only on standard cameras and rely on image-processing techniques to interpret gestures. The vision-based approach is more natural and easier to use than the sensor-based approach.

In IWR(Isolated Word Recognition) vision-based approaches, after segmentation, extracting the suitable features is the important step to have an automatic recognition system. Its objective is to extract the most important information related to each word, which is difficult due to many conditions related to background, clothing, camera position, data acquisition. In this context, many features are

proposed focusing on hand characteristics (shape appearance, orientation and motion).

Two approaches are generally used to extract these features: no-tracking-based and tracking-based approaches [16]. With the no-tracking approaches, gestures are detected in a global manner [39] and the hand plays an important cue extracted from different images in sequence. Although these approaches were proposed in order to have a simple system and get rid of tracking approaches [41], they can't solve the signer variability problem [6] and the complexity of the database [13]. In view of the fact that the gesture is presented, generally, with quick movements of the hands and occluded head/hands parts, many methods are presented as a solution for this non-trivial mission using tracking techniques to predict how trackers(head/hands) move in the future frames. Also, using methods based on estimation theory to correct and update system state becomes a necessity to avoid data complexity and signer variability problem [4].

On the other hand, the tracking-based approach offers the possibility of extracting different regions of interest from an image. In this context, the detected positions of the object can be used as features, too. In the sign language recognition task, the tracked object(head/hand) is used for the calculation of positional and visual features. More advantages can be offered by the tracking-based approaches also, in particular, a robust feature extraction task in real-time condition [16]. Consequently, it will be wiser here to use the tracking based approaches.

A sign language recognition system can be classified as a signer-independent(SI) or signer-dependent system(SD). To have a good recognition rate, some works rely on the same signers in the training and testing phase with the SD system. But if a new signer is used in the training stage with the SI system, we have an adapting system able to accept all new signers. In our work, we have opted for the SI condition. So, a SIGNUM database for SI Continuous Sign Language Recognition published by RWTH Aachen University in Germany [36] will be used in the evaluation step.To sum up, in this paper we will deal only with isolated signs.

This can consist of a large vocabulary of signs. Also, we will concentrate only on manual features when introducing the vision-based approaches and applying the tracking-based techniques in order to have a more natural and robust system as a final step. In our work, we have also opted for the SI condition and decided to take into account the interpersonal variability when the user executes gesture. Figure 1 presents, with different circles, the overall context of our proposed sign language system and its main axis of research. The choice of this axis is well argumented in the next section based on the different existing works on IWR systems.

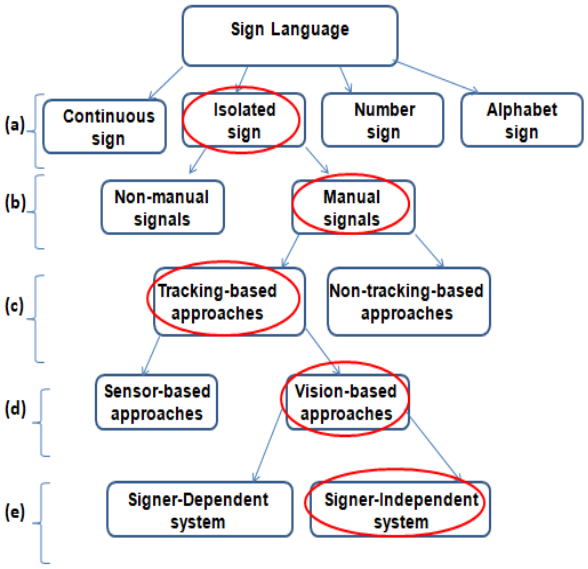

Fig. 1. Study context: (a) sign categories, (b) sign characteristics, (c) sign detection techniques, (d) sign data acquisition approaches and (e) sign classification conditions

This paper is organized as follows: Section 2 illustrates the related works. Section 3 presents the proposed IWR system based on tracking methods. Two levels are presented in order to extract the feature vector: the static-level which is presented only in the first frame of the video stream and the dynamic-level which takes all the rest of the frame in the video stream. Section 4 is concerned with the system's evaluation.

First, we are going to address the evaluation step with a public dataset, the SIGNUM corpora. Second, some ameliorations will be proposed in order to increase the obtained recognition rate and reduce time processing. Then we are going to present a comparative study with existing works. In this context the same classification conditions are applied. Finally, Section 5 is a conclusion to the whole paper.

2 Related Works

This section presents an overview of the existing IWR systems. The main problems of IWR come from the complexity of hand movement and shape. Counting on the manner how manual features are extracted for the classification step, IWR existing systems are based on different restrictions. In fact, in order to have a precise head/hands position detection and an easy hand motion and shape extraction stage, many works are based on Sensor-based approaches. In reference [26], the user was obliged to be placed before a single camera, wearing two gloves with different colors. In addition, different background and clothing conditions are suggested in order to facilitate the extraction of the proposed characteristics which are; head/hands centroids, the angle of the first principal component and the regions for the two hands. This approach gave 98% recognition rates with the Hidden Markov Model(HMM) classifier relative to 50 words presented with 500 signs and 95 % recognition rates with an enriched database of 300 words [27].

The approach presented in reference [32] requires that the user wears gloves to simplify the process of head/hand segmentation while applying the Zonal code. A frame difference is also used to characterize the gesture movement. This system achieved an 87% recognition rate with the classical classifier K nearest neighbors(KNN) and a database composed of 23 signs (words / phrases) captured with three different deaf people. Also this approach reached 97% [33] recognition rate while combining a KNN and a Bayesian network in the classification stage. With the same database, two levels of characteristics extraction [33] were proposed.

In the first level, motion information was detected using the temporal domain (frame difference). In the second level, two approaches were applied: Frequency Transformation (Zonal code) and Radian Transformation on each image difference collected in the first level. Thanks to this amelioration, 39% of the errors have been corrected using the Linear Fisher Discrimination.

Obviously, most illustrated works based on sensor-based approaches are expensive and limit the user's movement. In this situation, the user cannot manifest naturally and cannot be naturally attached to the workplace conditions. Further, with vision-based approaches, just one camera is employed to capture the gesture without limiting the user movement. Hence, hand segmentation and tracking present the most challenging and important steps toward IWR. A detailed discussion of the field of sign language recognition systems is presented in [16]. Thus, two research foci are proposed to extract manual features: no-tracking-based and tracking-based approaches.

No-tracking methods propose to extract features directly from the image. In fact, the authors in [42] propose to use skin intensity thresholding related to the original frames and extract its derivative feature vectors(first derivative, positive first derivative, negative first derivative, absolute first derivative, second derivative) to recognize 10 words in the American Sign Language. A 7% error rate is achieved when applying the leaving one out method and the HMM for the training and the classification steps. In this context, the variability of the signers presents a problem, so introducing an invariant feature vector respecting scale and position becomes a necessity to make feature vectors more robust.

The authors in [41] account for visual variabilities to recognize 50 ASL words without the tracking method and propose two different combinations of the local and the global transformations. The proposed features are generated by multiplication of the model: they are generated by multiplying the tangent distance and the zero-order local deformation model. In this context, the error rate is reduced from 22.2% to 17.2%.

But the question, now, is: what with a large dataset? No-tracking methods have good results with

a limited dataset. In this context, visual-tracking methods present the important task toward an automatic IWR system in order to account for a large gesture dictionary. However, tracking approaches guarantee an accurate estimation and offer an unlimited workspace [13] compared to no-tracking approaches. But also, generally with the tracking-based system, many constraints are imposed [16]:

— Movement speed: the problem is generally related to hand movement which is very fast.

— Shape variability: the hand is very flexible with a deformable shape and a changing posture in different situations.

— Occlusion condition: occlusion between hand/hand and hand/face increases the complexity of the skin-segmentation approaches.

In the literature, many of the problems presented above are solved by applying restriction conditions. In order to avoid occultation problems, some works propose to use only the dominant hand and concentrate on its important characteristics using a limited vocabulary. For example, in [43], in the first stage, a probabilistic model is applied to detect skin color region (gray level). Also, in order to characterize each gesture, this work introduces a principal component analysis(PCA) to measure the configuration and the overall direction of the hand. Features of the local image are represented with the flattening position (kurtosis position) and Motion chain code to design the hand's trajectory.

The database used, RWTH-BOSTON-50 [43], is released by the National Center of Boston University. Only the gray level of the frontal image is used in this work and a 10.91% versus 13.63% error rate with HMM classifier is obtained. Another work [17] applied a snake algorithm combined with a motion cue in the occultation condition (face/hand) to calculate the contour of the moving object. In this work, the head/hand are tracked by their geometric characteristics (position/shape). This method is not robust with complex and moving backgrounds. Also, it is not available with the short-sleeved clothing condition.

The work presented in [40] proposes a system without signer or word length restrictions by applying an HSV(Hue, Saturation, Value) space and a median filter. Edge detection has been applied also using a Canny filter. In this context, the Connected Component Analysis(CCA) has been introduced before the edge detection: the largest region is considered the head and the others are the hands. In this stage, 8 features have been introduced, namely the head's position, the coordinate centers of each hand, the hands' orientation relative to the head, the size and orientation of each hand and its global positions. An 82.2% recognition rate is obtained with an HMM classifier and a database composed of 20 gestures repeated 45 times. But what happens with a real time system, with a large dataset and with the SI condition?

Other researches propose some filters [35, 30] as a solution such as kalman and particular filters. Some works [28], in order to extract isolated words from continuous word problems, rely on a particle filter in the head/hands tracking stage. The tracking stage results have as output some ellipse characteristics (location, velocity, width, height and orientation) related to the face and each hand. The HMM classifier improved its performance and the obtained ratio between the searched window and the error is less than 30%.

Work [9] used a framework related to the particle filter combined with a PCA method based on a skin color region segmentation to develop a hand tracking system. In this context, an RWTH-Boston-104 [9] database was used and it improved the robustness of PCA for hand modeling compared to ground truth data and it also has a few minutes to execute videos of 100 frames. Generally, most of the proposed systems are based on shape and motion information(see Table 1). This, is due to the human nature when executing a dynamic gesture. Since the human being is based naturally on different movements when presenting a message dynamically, in addition we use the hand's shape to describe more precisely the meaning of each gesture.

According to Table 1, all existing approaches use and impose many constraints in order to overcome the encountered problems.

Table 1. Extracted feature and imposed condition in an isolated word recognition system

Works	Constraints	Hand shape feature	Hand motion feature
[26, 27]	Wearing gloves	Region	Centroids
	Background condition	Angle	
	Clothing condition		
[32]	Wearing gloves		Motion trajectory
[33]	Wearing gloves	Zonal code	Frame differencing
		Radian transform	
[43]	Only dominant hand	Kurtosis position	Motion Chain Code
		PCA	
[17]	Short-sleeved clothing	Shape position	Snake algorithm

Taking all this into account, in this research we propose a new IWR system which uses a tracking-based approach. We combined motion and shape information but with a new strategy to extract more discriminative features. The proposed approach has two-level features extraction:a static-level that detects, in the first frame of the video stream, only the region of interest related to head/hands and then extracts its key points, and a dynamic-level that accumulates, in the rest of the frame in the video stream, the key-point trajectory matrix(KPTM) related to each proposed head/hands key-point trajectory. The main contribution of this work is that it has a natural IWR system and relies on a vision-based approach without device limitation and without imposed constraints such as:

— Fixing environment and clothing restrictions.

— Using only one dominant hand or two hands.

— Using a reduced database.

Since SI is a fundamental precondition for the future IWR system, it is also an important subject of this paper.

3 Proposed System

Head/hand motion is one of the important cues for interpreting each word gesture's meaning. Also, the visual information related generally to shape, adds a lot of meaning when gestures are performed. This paper takes into account the two key pieces of information (motion/shape) in order to benefit from visual data (shape), evolution, and changeability in time gesture progression.

The main idea proposed in this context can be summarized in two global levels. The static-level extracts the head/hands region in order to select only the most important points which enable us to describe each head/hands position and shape in time. The dynamic-level accumulates trajectories of all the detected key points in a matrix to describe each word gesture. Finally, each word gesture is presented as a moving key-point group in time.

We propose to use tracking techniques able to generate each proposed point's trajectory. Finally, a word matrix trajectory (KPTM) composed of the concatenation of each detected key-point trajectory is generated. In the recognition step, a gesture-trajectory matrix is tried by using a support vector machine(SVM) [29] in order to judge which word it belongs to.

A key-point extraction stage becomes a necessity. Also, the choice of these points is important and needs a good detection of the regions of interest (head, left hand and right hand). Figure 2 illustrates our proposed system which is composed of three principal stages:

Stage 1: Static-level:

— Pre-processing stage: to extract the regions of interest(head, left hand and right hand).

— Key-points extraction stage: to extract the interest points which present each detected region(head, left hand and right hand).

Stage 2: Dynamic-level:

— Trajectory extraction stage: to have finally the word matrix's characteristics(KPTM).

Stage 3: Classification:

— Using an SVM classifier with SI conditions.

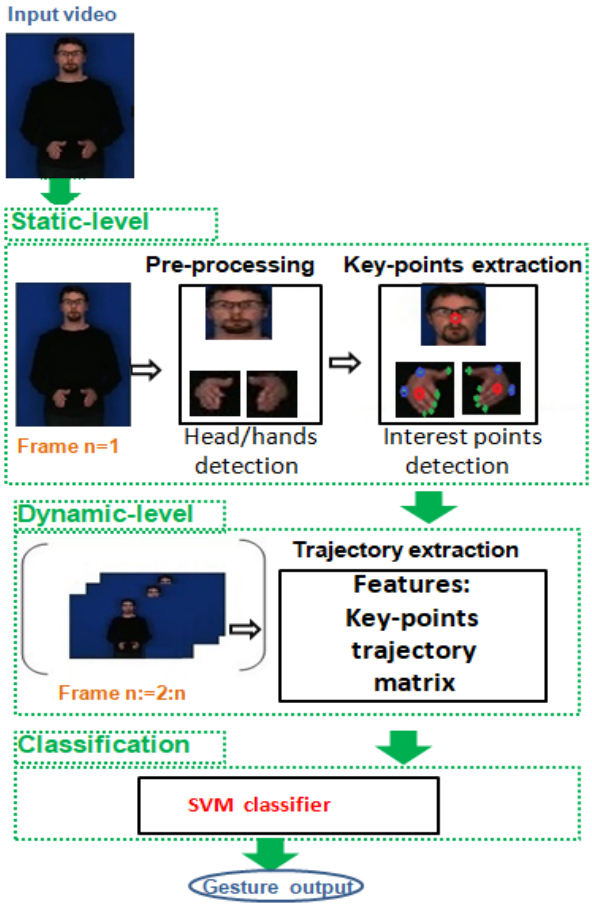

Fig. 2. Proposed IWR system

3.1 Static-Level

3.1.1 Pre-Processing

Head/ hand detection is the first prerequisite task in our IWR system. Its goal is the extraction of the region of the interest. We propose several operations illustrated in Fig. 3 and Alg. 1 based on three principal steps.

Skin Color Detection:

The goal of the skin color region detection is the limitation of the area of interest to reduce the time and the complexity of the detection algorithm and also to guarantee the presence of pertinent information related to a head/two hands regions in the detected bloc. In this step, we propose to apply a YCbCr colors space to improve the performance of the skin-color segmentation under lighting conditions(see Fig. 3a) and to search for color information in the chrominance value Cb in [77, 127] and Cr in [133, 173] [19].

Morphology Operation:

The detection of the skin color region can cause the existence of noisy information because some small regions can be present related to the skin colors' information which exists in the background picture. In order to eliminate it, we propose to apply the closing and the opening techniques (see Fig. 3b) and we keep solely the three biggest regions generally related to head/hand objects.

Head And Hands Detection:

In the occultation problem, the biggest region can be related to the left/right hand or the Head/hand region. So we propose to apply the Viola and Jones method [7, 11] to detect the face region and the rest is considered as the hands' regions. This, decreases detection errors and provides a good head/hands localization (see Fig. 3c).

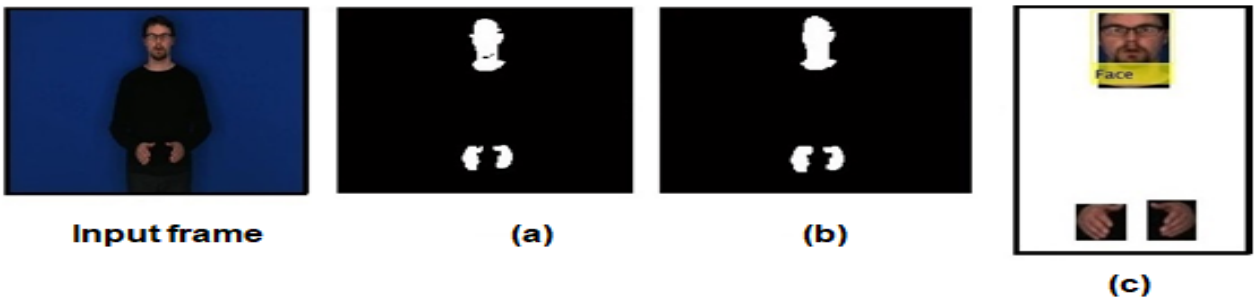

Fig. 3. Head/hand detection process, in the first frame: (a) Skin colors detection, (b) Morphological operation and (c) Head/hands detection operation

Algorithm 1 Pre-processing stage

Inupt:colors image
First frame
1. Skin color's detection: YCbCr color space.
2. Morphology operation application.
3. Face detection: Viola and Jones technique.
Output: ROI detected region.

3.1.2 Key-Point Extraction

In this section, we will try to exploit two principal claims: a general-shape reference (related to head/hands objects), and an intrinsic-shape reference (related only to hands objects).

First of all, we extract the head/hand center of gravity because it can refer to a head/hand object position in general [24]. The region designated as face object is referenced only by its gravity center but we extract more key points related to left and right hand in order to present the hand's shape faithfully.

To guarantee an intrinsic characterization in a robust way, we extract different points which are naturally used to represent hand poses because gesture can be defined as sequence of hand poses [22]. So, for each hand, we extract each finger position because it can give details about the hand pose versus time.

Also, we extract the two-points position relative to the extremity of the wrist line. They will be used in wrist detection step [12] as a solution to extract only the hand region when the user is wearing short-sleeved clothes. This can ensure our system robustness against clothing conditions.

Consequently, all in all, we obtain seven end points for each hand, which makes 17 key points (see Fig. 4):

— (Hg): Head gravity center.

— (LHg, RHg): Left and right hand gravity centers.

— $(FL1, FL2, FL3, FL4, FL5)$: Each finger points left hand position.

— $(FR1, FR2, FR3, FR4, FR5)$: Each finger points right hand position.

— $(WL1, WL2)$:Two-points related to the extremity of the wrist line; left hand.

— $(WR1, WR2)$: Two-points related to the extremity of the wrist line; right hand.

Head/Hand Center Of Gravity

This stage presents the detection of generic information relative to the head and hand positioning. We have chosen to calculate the center of gravity of each object (head/hand) detected(Hg, LHg, RHg) based on Eq. 1 where $Xi = (x, y)$ is a point in the detected region and n is the number of points:

$$\underset{c}{X} = \frac{\sum_{i=1}^{n} Xi}{n}. \tag{1}$$

Figure 5 illustrates the gravity center points detection.

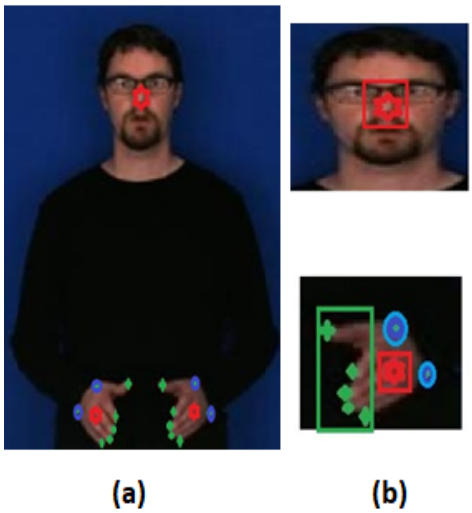

Fig. 4. The proposed 17 key points: (a) All 17 key points extracted and (b) Zoomed head and left hand: The squares represents the head and left-hand centers of gravity, the rectangle includes the 5 finger tips positions and the circles represent the position relative to the extremity of the wrist line

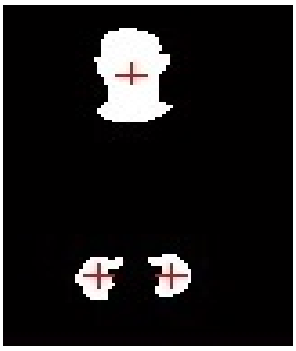

Fig. 5. Frame of ”BOOK” sign language gesture: Gravity center detection

End Points

The goal of this section is to extract the seven end points related to the 5 finger tips and wrist line positions. However, in our situation we propose to extract these points in the first frame relative to the beginning of the gesture when it has an open palm. So the constraint of finger occultation is totally absent (see Fig. 6).

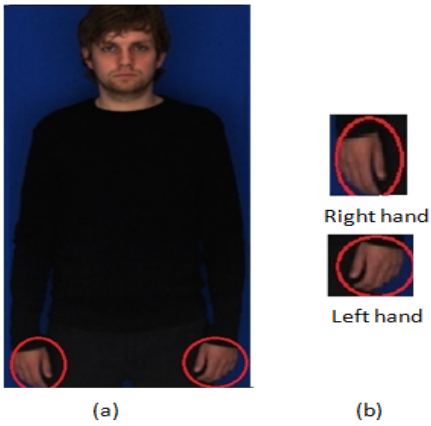

Fig. 6. The extraction of the first frame related to ”BOOK” sign having an open palm: SIGNUM corpora: (a) Original image, (b) Zoomed hands

We propose to use a convex hull. Some works, like the American Sign Language [18], use the convex Hull characteristic in the palm-tracking problem, but in our approach we use it to find the set of points enveloped in the hand.

Also, we can assume that the fingertip is a force related to the end points of the convex defect(see Fig. 7).

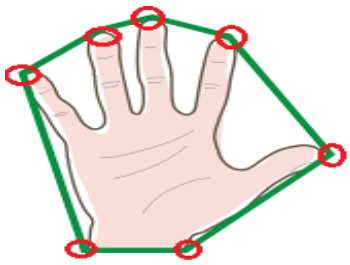

Fig. 7. Proposed points: the polygone represents the hand’s convex hull and the seven circles represent the seven end points of the convex defect

In each of these convex hulls we compute the convexity defect. At each pixel i in a convex hull c, we compute the curvature of defect using ’vector dot product’. Vc(l) represents the ending point of defect and Vc(i) represents the starting point of defect.

These points, including the points related to fingertip and wrist position [15],are represented with seven circles in Fig. 7. The seven stars presented in Fig. 8 illustrate our end-point-detection results related to the left and the right signer's hand.

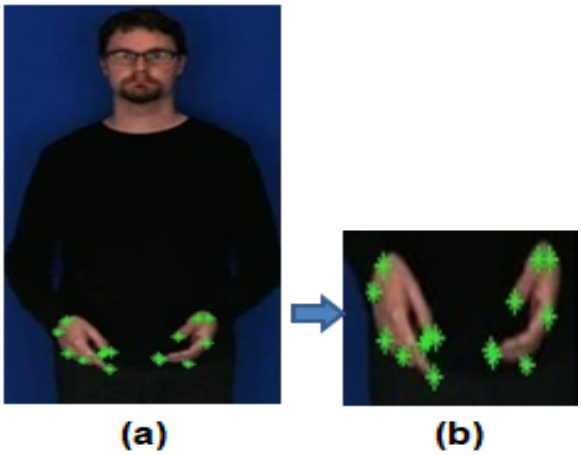

Fig. 8. Extracted end points: "Butter" sign language gesture: (a) Original image and (b) Zoomed hands

3.2 Dynamic-Level

3.2.1 Trajectory Extraction: Tracking With Particle Filter

Tracking can be presented as an estimated trajectory problem of an object during its movement. It is based on the motion prediction using the Kalman filter [35] or Particle filters [30]. These filters have a principal idea for the estimation of the current position based on the previous one. Unlike the Kalman filter, particle filters ensure:

— A stability tracking with non-linear movement robustness with the multiple-object tracking problems.

— Robustness in occulted situations.

For these reasons [31], we opted to apply a multi-point tracking system based on the Particle filter(PF). Each point is considered as an object and a Particle filter(PF)is initialized for each point. For each initialized Particle filter(PF) in each iteration stage, we estimate the position of the tracked point. A trajectory vector is obtained by accumulating the estimated position in each video frame. Finally, a key-point trajectory matrix (KPTM) related to each word based on the concatenation of the detected trajectory is obtained:
$KPTM = (Hg, LHg, FL1, FL2, FL3, FL4, FL5, WL1, WL2, RHg, FR1, FR2, FR3, FR4, FR5, WR1, WR2)$.

Figures 9, 10 and 11 present examples of extracted trajectories.

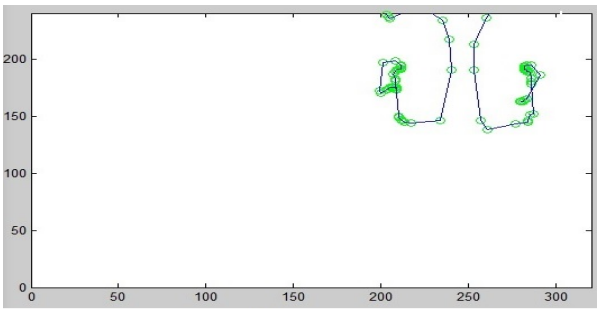

Fig. 9. Extracted trajectory related to "BOOK" sign: Left and right hand gravity center trajectory

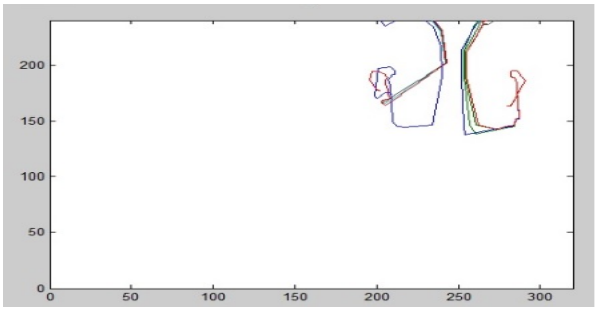

Fig. 10. Extracted trajectory related to "BOOK" sign: Left and right hand points trajectory

3.2.2 Trajectories' Normalization

In order to have a system which is robust to translation and scale invariance, the trajectories must be normalized. We apply the similar normalization technique used in reference [3] for each extracted trajectory. The normalized trajectory coordinates are calculated with the following equations.

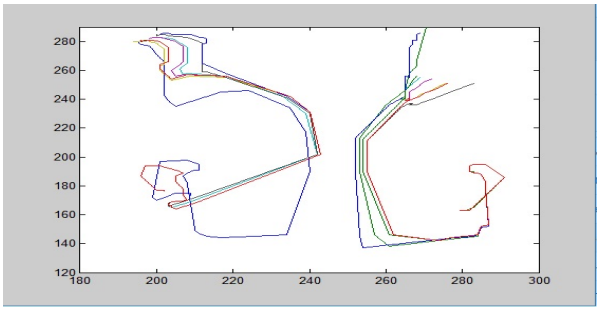

Fig. 11. A zoomed 16 points trajectory related to "BOOK" sign: Left and right hand points trajectory

— For translation normalization:

We define two mid-points of the range in the x and y coordinates respectively:

$$\underset{m}{x} = (\underset{max}{x} + \underset{min}{x})/2, \tag{2}$$

$$\underset{m}{y} = (\underset{max}{y} + \underset{min}{y})/2. \tag{3}$$

— For scale normalization:

We define two amount points of spread in the x and y coordinates respectively:

$$\underset{x}{d} = (\underset{max}{x} / \underset{min}{x})/2, \tag{4}$$

$$\underset{y}{d} = (\underset{max}{y} / \underset{min}{y})/2. \tag{5}$$

The scaling factor is selected to be the maximum of the spread in the x and y coordinates since scaling with different factors disturbs the shape:

$$d = max(\underset{x}{d}; \underset{y}{d}). \tag{6}$$

The normalized trajectory coordinates are:

$$((\underset{1}{x'}; \underset{1}{y'})...(\underset{t}{x'}; \underset{t}{y'})....(\underset{n}{x'}; \underset{n}{y'}))), \tag{7}$$

tacking into consideration that n is the sequence length and $0 \leqslant \underset{t}{x'}, \underset{t}{y'} \geqslant 1$.

All coordinates are calculated as follows:

$$\underset{t}{x'} = 0.5 + 0.5(\underset{t}{x} + \underset{m}{x})/d, \tag{8}$$

$$\underset{t}{y'} = 0.5 + 0.5(\underset{t}{y} + \underset{m}{y})/d. \tag{9}$$

The proposed steps in our isolated word gesture recognition system are summed up in Algorithm 2.

Algorithm 2 IWR system: Key Point Trajectory Matrix extraction

input system: Video stream.

A. *frame=1*

Input: color image

1.Pre-processing.

2.Key points' extraction.

Output: Vector V(x,y) of key extracted points: 17 points

3. Vector V normalization.

4.KPTM initialized matrix by V vector.

5. Initialize 17 PF.

For each PF and in each frame

Normalize and update KPTM position matrix with each updated position by each PF

Output system: updated KPTM position matrix.

4 System Evaluation

Two scenarios have been proposed to evaluate our proposed recognition system. First of all, we evaluate the different proposed steps as presented in Section 3. Second, we propose an amelioration step to reduce the time processing and increase the recognition rate.

4.1 Results of the Proposed System

4.1.1 Dataset: SIGNUM Corpora

We propose to use a public dataset with a large vocabulary. In order to guarantee the SI condition, we suggest using the SIGNUM database containing the German Sign Language (DGS) [37, 38]. The SIGNUM database has been created under laboratory conditions with a uniform background as well as dark clothes for the signer. The corpus is based on 450 signs in the German Sign Language and comprises 780 sentences. Each sign was performed once by 25 native signers of different sexes and ages.

Figure 12 shows some images from this database.

ISSN 2007-9737

Fig. 12. Three image examples of the SIGNUM database

4.1.2 Test Protocol

In the test protocol, the corpus is divided, as in work [37], into three sub corpora simulating a vocabulary of 150, 300, and 450 signs respectively. Also the evaluation is based on SI. Table 2 summarizes the test protocol.

Table 2. The signum test protocol: 5 signers

Word	150	300	450
Vocabulary size	750	1500	2250
Frames	22500	45000	67500
Training(70%)	525	1050	1575
Test(30%)	225	450	675

In order to have a real-time IWR system, we propose to use an SVM classifier.

4.1.3 Results

The experimental results presented in Table 3 confirm the robustness of our proposed system. It reaches a 93.7% recognition rate with manual features only applied in a large database. In addition to that, the SI condition is considered and it improves the performance of the proposed trajectory key-point matrix in the Isolated word Sign Language recognition system.

Table 3. Signum recognition rate with signer-independent condition: 5 signers

Word	150	300	450
SIGNUM database	93.7%	90.65%	87.3%

4.2 Amelioration Of The Proposed System: Eliminating Redundant Frames

The main idea in this section is the Removal of Redundant Frames(RRF) in order to have a reduction in the processing time. Generally, a video stream captured by webcam has a big dimension and contains many similar frames with no new significant information. In this situation, it can be considered as noise and it reduces the system's competence. So, it becomes necessary to eliminate this noise and to create a new stream composed of a key frame, only. This dimensional reduction also offers a reduction in time complexity. Work [1] improved the performance of applying this step in the feature-extraction time stage. So, we have opted for applying the Redundant Frame Removal algorithm proposed in [1] after the preprocessing stage.

Algorithm 3 IWR system: Eliminating redundant frames in each video stream [1]

1. Set the frame counter a = 1.
2. For each frame k = 2..N

a) Compute difference image between frames S(k) and S(a).
b) Convert difference image to binary image using global image threshold
c) Count the number of pixels in each connected component of the binary image.
Let Np be the number of pixels in largest component.
d) If Np $>$ A (threshold), set a = k else skip frame k.
e) Set k=k+1.

The details of this algorithm are illustrated in Algorithm 3 and Fig. 13 which presents the frame obtained after the elimination step, where 17 frames are deleted. Also, the two obtained frames presented in Fig.13(b) prove that only the frames having rotation and scale variations are extracted.

With the elimination of redundant frames, the processing time is reduced. For example, the execution time related to the BEURRE gesture is reduced by 48.45% (see Table 4). This reduction is related to the decrease in the number of created

ISSN 2007-9737

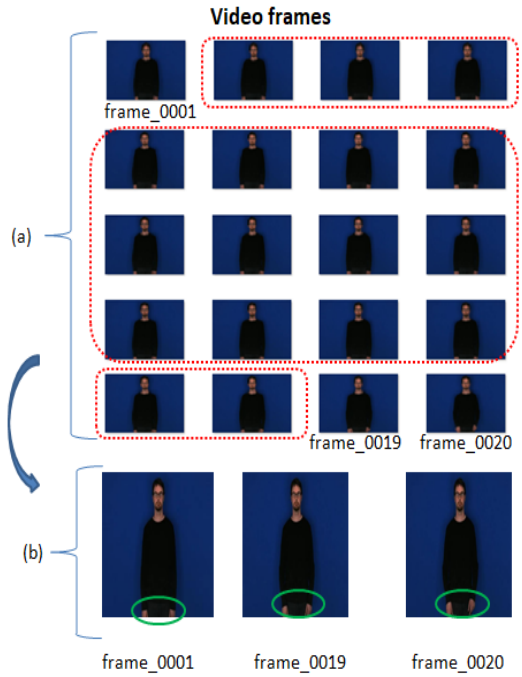

Fig. 13. Frame elimination step: "BOOK" German sign language gesture: (a) Without applying RRF step, (b) With applying RRF step

particles in the tracking stage and the construction of KPTM.

Table 4. Beurre gesture executing time: time reduced with the removal of redundant frames step

Beurre gesture	Executing time
Time with RRF	2.072 (s)
Time without RRF	4.020 (s)

Also, an amelioration in recognition rate is obtained; 95.01% with 150 signers when applying the 'removing redundant frames' step. Table 5 illustrates the recognition rate with and without the RRF step.

Consequently, it is better to add a key frame extraction stage, to the database, to improve performances. In our context, presenting gestures only by pertinent frames improves the data presentation which will be processed for recognition. This idea of working with pertinent data and presenting only key frames is applied in reference [25, 23, 21, 8]. In fact, when we use an efficient data presentation strategy, a relevant feature vector can be obtained, which can, on the one hand, offer a reduction of the training times and on the other hand, a reduction of the over-fitting condition during the training process of the SVM classifier. Hence the improvement of the recognition rate.

Table 5. Recognition rate with and without the removal of redundant frames step

Word	150	300	450
SVM with RRF	95.01%	92.9%	91.7%
SVM without RRF	93.7%	90.65%	87.3%

Algorithm 4 summarizes all the proposed steps in our Sign Language Word Recognition System.

Algorithm 4 IWR system: Key-Point Trajectory Matrix extraction

input system: Video stream.

A. *frame=1*

Input: color image
1.Pre-processing.
2.Key points' extraction.
Output: Vector V(x,y) of key extracted points: 17 points
3.Remove redundant frames.
4. Vector V normalization.
5.KPTM initialized matrix by V vector.
6. Initialize 17 PF.
For each PF and in each frame
Normalize and update KPTM position matrix with each updated position by each PF

Output system: updated KPTM position matrix.

4.3 Comparison with Existing Work

In this section, we compare our proposed system with the existing ones. In works [37] and [38] a dealing strategy based on manual and non-manual features is presented. An image-processing stage and a feature-extraction stage are introduced.

The image-processing stage is based on image enhancement, background subtraction and face localization. Hand tracking and overlap resolution are employed in the manual feature-extraction stage. In the recognition stage, only the manual features are tested.

Figure 14 presents the manual hand features extracted in [37] and [38]. Table 6 illustrates the extracted hand features in [37] and [38] and in our approach.

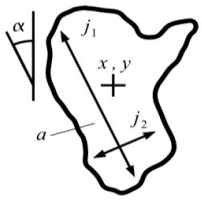

Fig. 14. Geometric features computed for each hand [37]

Table 6. Feature vector used in reference [37] and in our work at each time t

	Our feature vector	Extracted feature vector in reference [37]
Vector for each hand	KPTM	X
Final vector length	17	22

where:

$KPTM = [Hg, LHg, FL1, FL2, FL3, FL4, FL5, WL1, WL2, RHg, FR1, FR2, FR3, FR4, FR5, WR1, WR2]$.

$$X = [\underbrace{x,' x', y,' y', a,' a', o1, o2, r, c, e}_{lefthand},$$
$$\underbrace{x,' x', y,' y', a,' a', o1, o2, r, c, e}_{righthand}]$$

Where the template matching is used to accurately calculate the center coordinates x, y which is estimated from the width $\underset{F}{w}$ and the face position.

To ensure feature normalization step, the area a is normalized by $\underset{F}{w^2}$ and all coordinates by $\underset{F}{w}$.

α: Orientation of main axis. As $\alpha \in [-90°, 90°]$, it is split into $o1 = sin2\alpha$ and $o2 = cos\alpha$ to ensure the interval border stability.

r: Ratio of inertia along/perpendicular to main axis:

$$r = \frac{j_1}{j_2}, \tag{10}$$

c: Compactness:

$$c = \frac{4\pi a}{circumference^2}, \tag{11}$$

e: Eccentricity:

$$e = \frac{(U_{2,0} - U_{0,2})^2 + 4U_{1,1}}{a}, \tag{12}$$

where $U_{p,q}$: Central moments and the derivatives x', y', a' complete the 22-dimensional feature vector.

To have a faithful comparison, we use the same classification condition as work [37] and apply a HMM classifier. Also, we use the same database, SIGNUM corpora, with the same test protocol. The corpus is also divided into three sub corpora composed of 150, 300 and 450 isolated word gestures, respectively. The obtained classification rate, 94,3% with SI condition, presented in Table 7, confirms our proposed matrix feature robustness to present a word gesture and its capability to surpass existing works [37].

Table 7. Comparison study in terms of recognition rate

Word	150	300	450
Approach [37]	74.9%	71.2%	68.5%
Our approach with:	94.3%	92.7%	90.1%
+ HMM and RRF			
+ SI condition			

Our system was able to reach 90.1% of accuracy rate, but reference [37] attained only 68.5% of accuracy rate of 450 SIGNUM signs. These results displayed in Table 7, highlight the robustness of our proposed approach, especially the feature vector related to the head/hand key points and show the importance of presenting gestures as trajectories related to the most important points which describe

each pose in each time. They also prove the importance of eliminating redundant frames in the tracking phase.

5 Conclusion and Future Work

In this paper, we have presented a new approach for isolated word Sign Language gesture recognition tested on a vision-based approach. On the other hand, we have suggested presenting each isolated word as a trajectory related to the most moving key hand-points information. Our system is based on two principal levels: a static level and a dynamic level.

Also, the proposed step, being the removing redundant frames has demonstrated a good influence on our proposed system. All the experimental results have a satisfactory recognition rate and also prove their performance compared to existing studies in public datasets: SIGNUM corpora. The question, now, is: has the confusion between the similar gestures been totally solved.

In order to better improve the results, we envision working on the introducing of a new technique which represents each gesture's trajectory with an intrinsic formulation and to examine each trajectory in Kendall's shape space as in work [5]. Also, it is worth thinking more deeply about a real-time system that can increase our system performances and try to test it with continuous gestures.

References

1. **Agrawal, S. C., Jalal, A. S., & Bhatnagar, C. (2014).** Redundancy removal for isolated gesture in indian sign language and recognition using multi-class support vector machine. *Int. J. Comput. Vision Robot.*, Vol. 4, No. 1/2, pp. 23–38.
2. **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. *International Journal of Applied Pattern Recognition (IJAPR)*, Vol. 3, No. 2.
3. **Aran, O. & Akarun, L. (2006).** Recognizing two handed gestures with generative, discriminative and ensemble methods via fisher kernels, pp. 159–166.
4. **Balaji, S. R. & Karthikeyan, S. (2017).** A survey on moving object tracking using image processing. *2017 11th International Conference on Intelligent Systems and Control (ISCO)*, pp. 469–474.
5. **Ben Tanfous, A., Drira, H., & Ben Amor, B. (2018).** Coding Kendall's Shape Trajectories for 3D Action Recognition. *IEEE Computer Vision and Pattern Recognition*.
6. **Corradini, A. (2001).** Dynamic time warping for off-line recognition of a small gesture vocabulary. *Procs. of ICCV: Wkshp: Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems, Vancouver, BC, pp. 82–90. IEEE Comput. Soc., Los Alamitos*.
7. **Da'san, M., Alqudah, A., & Debeir, O. (2015).** Face detection using viola and jones method and neural networks. *2015 International Conference on Information and Communication Technology Research (ICTRC)*, pp. 40–43.
8. **Dong, Z., Zhang, G., Jia, J., & Bao, H. (2014).** Efficient keyframe-based real-time camera tracking. *Computer Vision and Image Understanding*, Vol. 118, pp. 97 – 110.
9. **Du, W. & Piater, J. (2012).** Hand modeling and tracking for video-based sign language recognition by robust principal component analysis, pp. 273–285.
10. **Elons, A. S., Abull-ela, M., & Tolba, M. (2013).** A proposed pcnn features quality optimization technique for pose-invariant 3d arabic sign language recognition. *Applied Soft Computing*, Vol. 13, No. 4, pp. 1646 – 1660.
11. **Ephraim, T., Himmelman, T., & Siddiqi, K. (2009).** Real-time viola-jones face detection in a web browser. *2009 Canadian Conference on Computer and Robot Vision*, pp. 321–328.
12. **Fakhfakh, S. & Jemaa, Y. B. (2017).** Hand and wrist localization approach for features extraction in arabic sign language recognition. *2017 IEEE/ACS 14th International Conference on Computer Systems and Applications (AICCSA)*, pp. 774–780.
13. **Filippeschi, A., Schmitz, N., Miezal, M., Bleser, G., Ruffaldi, E., & Stricker, D. (2017).** Survey of motion tracking methods based on inertial sensors: A focus on upper limb human motion. *Sensors*, Vol. 17, No. 6.
14. **González, G. S., Sánchez, J. C., Díaz, M. M. B., & Pérez, A. A. (2018).** Recognition and classification of sign language for spanish. *Computación y Sistemas*, Vol. 22.

15. **Gurung, D., Jiang, C., Deray, J., & Sidibé, D. (2013).** Hand Gestures Recognition and Tracking. Working paper or preprint.
16. **Helen Cooper, R. B., Brian Holt (2011).** Sign language recognition. *Visual Analysis of Humans: Looking at People*, pp. 539 – 562.
17. **Holden, E.-J., Lee, G., & Owens, R. (2005).** Australian sign language recognition. *Machine Vision and Applications*, Vol. 16, No. 5, pp. 312.
18. **Hussain, I., Talukdar, A. K., & Sarma, K. K. (2014).** Hand gesture recognition system with real-time palm tracking. *2014 Annual IEEE India Conference (INDICON)*, pp. 1–6.
19. **Jemaa, Y. B. & Khanfir, S. (2009).** Automatic local gabor features extraction for face recognition. *CoRR*, Vol. abs/0907.4984.
20. **Jmaa, A. B., Mahdi, W., Jemaa, Y. B., & Hamadou, A. B. (2016).** A new approach for hand gestures recognition based on depth map captured by rgb-d camera. *Computación y Sistemas*, Vol. 20.
21. **Kim, J., Yoon, K.-J., & Kweon, I. S. (2015).** Bayesian filtering for keyframe-based visual slam. *The International Journal of Robotics Research*, Vol. 34, No. 4-5, pp. 517–531.
22. **Krupka, E., Karmon, K., Bloom, N., Freedman, D., Gurvich, I., Hurvitz, A., Leichter, I., Smolin, Y., Tzairi, Y., Vinnikov, A., & Bar-Hillel, A. (2017).** Toward realistic hands gesture interface: Keeping it simple for developers and machines.
23. **Li, X. & Xu, T. (2011).** Face video key-frame extraction algorithm based on color histogram. Vol. 51.
24. **Lu, W.-L. & Little, J. J. (2006).** Simultaneous tracking and action recognition using the pca-hog descriptor. *The 3rd Canadian Conference on Computer and Robot Vision (CRV'06)*, pp. 6–6.
25. **Miao, Q., Li, Y., Ouyang, W., Ma, Z., Xu, X., Shi, W., & Cao, X. (2017).** Multimodal gesture recognition based on the resc3d network. *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, pp. 3047–3055.
26. **Mohandes, M. & Deriche, M. (2005).** Image based arabic sign language recognition. *Proceedings of the Eighth International Symposium on Signal Processing and Its Applications, 2005.*, volume 1, pp. 86–89.
27. **Mohandes, M., Deriche, M., Johar, U., & Ilyas, S. (2012).** A signer-independent arabic sign language recognition system using face detection, geometric features, and a hidden markov model. *Computers Electrical Engineering*, Vol. 38, No. 2, pp. 422 – 433.
28. **Pinar Santemiz, M. S. L. A., Oya Aran (2018).** Extraction of isolated signs from sign language videos via multiple sequence alignment.
29. **Platt, J. C. (1999).** Advances in kernel methods. chapter Fast Training of Support Vector Machines Using Sequential Minimal Optimization. MIT Press, Cambridge, MA, USA, pp. 185–208.
30. **Roussos, A., Theodorakis, S., Pitsikalis, V., & Maragos, P. (2012).** Hand tracking and affine shape-appearance handshape sub-units in continuous sign language recognition. *Trends and Topics in Computer Vision*, pp. 258–272.
31. **Shan, C., Tan, T., & Wei, Y. (2007).** Real-time hand tracking using a mean shift embedded particle filter. *Pattern Recognition*, Vol. 40, No. 7, pp. 1958 – 1970.
32. **Shanableh, T. & Assaleh, K. (2007).** Arabic sign language recognition in user-independent mode. *2007 International Conference on Intelligent and Advanced Systems*, pp. 597–600.
33. **Shanableh, T. & Assaleh, K. (2007).** Two tier feature extractions for recognition of isolated arabic sign language using fisher's linear discriminants. *ICASSP (2)*, IEEE, pp. 501–504.
34. **Shanableh, T. & Assaleh, K. (2007).** Video-based feature extraction techniques for isolated arabic sign language recognition. *2007 9th International Symposium on Signal Processing and Its Applications*, pp. 1–4.
35. **Stenger, B. (2006).** Template-based hand pose recognition using multiple cues. *Computer Vision – ACCV 2006*, pp. 551–560.
36. **von Agris, U., Knorr, M., & Kraiss, K. (2008).** The significance of facial features for automatic sign language recognition. *2008 8th IEEE International Conference on Automatic Face Gesture Recognition*, pp. 1–6.
37. **von Agris, U. & Kraiss, K.-F. (2007).** Towards a video corpus for signer-independent continuous sign language recognition. *GW 2007 The 7th International Workshop on Gesture in Human-Computer Interaction and Simulation*, pp. 10–11. Poster.
38. **von Agris, U., Zieren, J., Canzler, U., Bauer, B., & Kraiss, K.-F. (2008).** Recent developments in visual sign language recognition. *Universal Access in the Information Society*, Vol. 6, No. 4, pp. 323–362.

39. **Wong, S.-F. & Cipolla, R. (2005).** Real-time adaptive hand motion recognition using a sparse bayesian classifier, pp. 170–179.
40. **Youssif, A. A., Aboutabl, A. E., & Ali, H. H. (2011).** Arabic sign language (arsl) recognition system using hmm. *International Journal of Advanced Computer Science and Applications*, Vol. 2, No. 11.
41. **Zahedi, M., Keysers, D., Deselaers, T., & Ney, H. (2005).** Combination of tangent distance and an image distortion model for appearance-based sign language recognition, pp. 401–408.
42. **Zahedi, M., Keysers, D., & Ney, H. (2005).** Appearance-based recognition of words in american sign language. *Pattern Recognition and Image Analysis*, pp. 511–519.
43. **Zaki, M. M. & Shaheen, S. I. (2011).** Sign language recognition using a combination of new vision based features. *Pattern Recognition Letters*, Vol. 32, No. 4, pp. 572 – 577.

Article received on 08/08/2018; accepted on 05/10/2018.
Corresponding author is Sana Fakhfakh.