IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2023.3332250

Two-Stage Deep Learning Solution for Continuous Arabic Sign Language Recognition Using Word Count Prediction and Motion Images

Tamer Shanableh, Senior Member, IEEE
Tamer Shanableh is with the American University of Sharjah, Sharjah, UAE
Corresponding author: Tamer Shanableh (e-mail: tshanableh@aus.edu).

ABSTRACT Recognition of continuous sign language is challenging as the number of words is a sentence and their boundaries are unknown during the recognition stage. This work proposes a two-stage solution in which the number of words in a sign language sentence is predicted in the first stage. The sentence is then temporally segmented accordingly and each segment is represented in a single image using a novel solution that entails summation of frame differences using motion estimation and compensation. This results in a single image representation per sign language word referred to as a motion image. CNN transfer learning is used to convert each of these motion images into a feature vector which is used for either model generation or sign language recognition. As such, two deep learning models are generated; one for predicting the number of words per sentence and the other for recognizing the meaning of the sign language sentences. The proposed solution of predicting the number of words per sentence and thereafter segmenting the sentence into equal segments worked well. This is because each motion image can contain traces of previous or successive words. This byproduct of the proposed solution is advantageous as it puts words into context, thus justifying the excellent sign language recognition rates reported. It is shown that bidirectional LSTM layers result in the most accurate models for both stages. In the experimental results section we use an existing dataset that contains 40 sentences generated from 80 sign language words. The experiments revealed that the proposed solution resulted in a word and sentence recognition rates of 97.3% and 92.6% respectively. The percentage increase over the best results reported in the literature for the same dataset are 1.8% and 9.1% for both word and sentences recognitions respectively.

INDEX TERMS Sign language, feature extraction, video processing, deep learning

I. INTRODUCTION

Sign language recognition facilitates the communications between the deaf and hearing communities. In general, the two-way communication entails converting sign language into text or speech and the other way around. The latter is a simpler task, as speech recognition is a mature technology and transforming the transcribed text into sign language is a deterministic task.

Sign language recognition is performed using various technologies including vision-based [1], glove-based [2] and sensor-based [3]. The most challenging and the least accurate of all is the vision-based approach. This is because in vision-based approaches, there are many variances including distance from camera, tilt and movement of the camera and non-stationary backgrounds. On the other hand, gloved-based and sensor-based approaches are more accurate but are more restrictive and require setup and batteries.

Additionally, sign language recognition can be applied to either sign language alphabet [4], [5] and [6], words (a.k.a. gestures) [7] and continuous sentences [8]. Clearly, the simplest of all is the alphabet recognition as it deals with static images only, this is followed by word recognition and then sentence recognition, which is the most challenging of all. This is because the number of words in a sentence and their boundaries are unknown during the recognition stage. Generally, alphabet and word recognitions are nearly solved problems and the challenge remains in sentence recognition. The challenge is sign language

recognition is even more elevated when it entails signer-independent recognition where the system is trained on a number of signers and tested on other signers [9].

In this work, we focus on signer-dependent Arabic sign language recognition of sentences using a camera and without the use of gloves or sensors.

In this work, we focus on the recognition of Arabic sign language sentences using a camera in user-dependent mode. We propose a two-stage solution in which deep learning is used to predict the number of words in a sentence in the first stage. This is followed by a second stage in which a sentence is segmented into words and recognized using a novel solution that relies on motion images and recurrent neural nets using biLSTM layers.

The novelty of the proposed solution pertains to providing a statistical insight into the sentence-based dataset used. The paper also introduces a two-stage solution that predicts the number of sign language words prior to classification. The paper also introduces the use of motion compensation in the formation of motion images used in both the prediction of the sign language words and the sign language recognition.

The rest of this paper is organized as follows; Section 2 presents the literature review of related work. Section 3 describes the dataset used and provides statistical insights regarding the number of frames and words used. Section 4 introduces the proposed feature extraction and classification solution. Section 5 introduces the proposed system of predicting the number of sign language words in a sentence. Section 6 describes the deep learning architectures used for model generation and recognition. Section 7 presents the experimental results and section 8 concludes the paper.

II. Related work

Recently, the work reported in [10] created the largest Saudi Sign Language database, which belongs to the Arabic sign language with 293 signs and 33 signers. The signs pertain to various domains including healthcare, numbers, days, family, and so forth. This rich dataset is word-based and does not contain sentences. Moreover, in [11] a new dataset of 80 common Arabic sign language words are recorded from 40 signers each repeated five times. In [13], 32 Arabic sign language sign and alphabets are collected from 40 signers with different age groups. The images have different dimensions and different variations resembling real-life acquisition. In [13] a multi-modality Arabic dataset that integrates facial expressions is recorded which consists of 50 words performed by four signers. Again, all these recent Arabic sign language datasets focus on isolated words not sentences.

Very few papers worked on recognizing Arabic sign language sentences, this includes [14] in which connected sequence of gestures are recognized using graph-matching techniques as a component of a real-time Arabic sign language recognition system. In [15] an Arabic sign language dataset was introduced with alphabet, words and sentences; however, the number of sentences are limited to five. More recently, in [16] 650 annotated Qatari sign language sentences are recorded and the authors intend to make them publicly available.

In this work, we make use of our sentence-based Arabic sign language dataset, which contains 40 sentences composed from 8 sign language words [17] and [18]. The dataset is made available to the research community through a dataset release form.

In [17] the authors proposed a time-sensitive KNN solution for recognizing continuous Arabic sign language sentences composed of 40 sentences and 80 words. A follow-up work was reported in [18] in which HMMs were used on the same datasets using vision-based and sensor-based solutions. Both solutions present signer-dependent approaches to sign language recognition.

In [19], a framework is proposed for signer-independent sign language recognition where hand shape features are extracted using a convolutional self-organizing map. The sequence of extracted feature vectors are then recognized using deep Bi-directional Long Short-Term Memory (BiLSTM) recurrent neural network.

In an attempt to perform robust gesture recognition, the authors in [20] proposed a rotation, translation and scale-invariant sign language recognition systems that uses CNN. Likewise, a sign language solution that caters for different angles and distances is reported in [21] where a convolution and transformer-based multi-branch network is used. Lastly, the work in [22] proposed an optimal segmentation solution for identifying hand gestures, which makes the recognition system more robust as well.

III. The Dataset

In this work, we use an existing Arabic sign language dataset, which was reported in [17] and [18]. The dataset contains 40 sentences with 17~19 repetitions per sentence. The sentences are composed of 80 sign language words that required the use of one or two hands. The data acquisition is performed using a single camera, sensors and/or data gloves are not used. The description of the dataset is listed in Table 1.

TABLE I
DESCRIPTION OF SIGN LANGUAGE DATASET USED

Sentence-based Arabic Sign Language Dataset	
Language	Arabic
Number of sentences	40
Number of words	80
Number of hands used	2
Repetitions per sentence	17~19
Number of signers	1
Vision-based acquisition	Yes
Sensor-based acquisition	No
Requires colored gloves or data gloves	No

The full list of the sign language sentences are shown in Table 2.

TABLE II
FULL LIST OF ARABIC SIGN LANGUAGE SENTENCES

Arabic Meaning / Arabic Transcription	English Meaning
نام ابي في الامس Ams Ab Naam	My dad slept yesterday
ذهبت الى نادي كرة القدم Ana Zahab Nadi Kurahkadam	I went to the soccer club
كم عمر اخيك؟ Istifhaam Umar Akh	How old is your brother?
اليوم ولدت امي بنتا Yawm Umm Wildat Bint	My mom had a baby girl today
اخي لا يزال رضيعا Akh Ana Radeeh	My brother is still breast feeding
في الامس نمت عند الساعة العاشرة Ana Ams Naam Saaah Asharah Masaah	Yesterday I went to sleep at 10:00 o'clock
ذهبت الى العمل في الصباح بسيارتي Subaah Zahab Amal Sayyaarah Ana	I am going to work in the morning in my car
ذهبت الى بيت جدي Jad Bayt Ana	I went to my grandfather's house
انا احب سباق السيارات Ana Uhub Sibaaka Sayyaarah	I like car racing
انا لا آكل قبل النوم Ana Akel Laa Kabl Naam	I do not eat close to bedtime
اشتريت كرة ثمينة Ishtarah Kurah Thameen	I bought an expensive ball
اشترى ابني كرة رخيصة Ishtarah Ibn Ana Kurah Rakhees	My kid bought an inexpensive ball
اكلت طعاما لذيذا في المطعم Akel Lazeez Matam	I eat delicious food at the restaurant
قرأت اختي كتابا Karaah Akh Bint Kitaab	My sister read a book
امي ذاهبة الى السوق هذا الصباح Umm Zahab Souk Subaah	My mom is going to the market in the morning
شاهدت انفجارشديد لبيت بالتلفاز Raaah Tilifizyoon Innfijaar Shadeed Bayt	I saw a big explosion on TV
يوم السبت عندي مباراة كرة قدم الساعة العاشرة Sabt Ana Mubaaraat Kurahkadam Saaah Asharah Subaah	On Saturday I have a soccer match at 10:00 o'clock
انا احب شرب الحليب في المساء Ana Uhub Shurb Haleeb Masaah	I like drinking milk in the evening
اين يعمل صديقك؟ Ana Uhub Akel Laham Akhthar Dajaaj	Where does your friend work?
هل اخوك في البيت؟ Istifhaam Akh Bayt	Is your brother home?
رأيت بنتا جميلة Ana Raaah Bint Jameel	I saw a beautiful girl
صديقي طويل Ana Sadeek Taweel	My friend is tall
ساشتري سيارة جديدة بعد شهر Baad Shahr Ishtarah Sayyaarah Jadeed	I will buy a new car in a month
هو توضأ ليصلي الصبح Huwa Tuwaddah Salla Subaah	He made Wadu for morning prayer
بيت عمي كبير Bayt Akh Ab Kabeer	My uncle's house is big
سيتزوج اخي بعد شهر Baad Shahr Akh Ana Zawaaj	In one month my brother will get married
سيطلق اخي بعد شهرين Baad Shahr Ithnayn Akh Ana Talaaq	In two months my brother will get divorced
اين يعمل صديقك؟ Istifhaam Amal Sadeek Huwa	Where does your friend work?
اكلت جبنة مع عصير Akel Jibnah Mah Aseer	I ate cheese and drank juice
اخي يلعب كرة سلة Ana Akh Kurahsallah	My brother plays basketball
في النادي ملعب كرة قدم Fi Nadi Malab Kurahkadam	There is a soccer field in the club
غدا سيكون هناك سباق دراجات Fi Gadaa Sibaaka Darrajah	There will be a bike race tomorrow
يوم الاحد القادم سيرتفع سعر الحليب Ahad Kadim Maal Haleeb Thameen	Next Sunday the price of milk will go up
عندي أخوين Ana Akh Ithnayn	I have two brothers
ذهبت الى صلاة الجمعة عند الساعة العاشرة Saaah Asharah Zahab Jumah Salla	I went to Friday prayer at 10:00 o'clock
أكلت زيتونا صباح الامس Ams Subaah Akel Zaytoon	Yesterday morning I ate olives
ما اسم ابيه؟ Istifhaam Ism Ab Huwa	What is his father's name?
وجدت كرة جديدة في الملعب Wajad Malab Kurah Jadeed	I found a new ball in the field
كان جدي مريضا في الامس Ams Jad Ana Mareed	Yesterday my grandfather was sick
انا احب شرب الماء Ana Uhub Shurb Maa	I like drinking water

Specific to this work, we report statistics on the dataset including the number of times each word appeared in the unique 40 sentences. This is reported in Figure 1 below where it is shown that 43 words occur in only one sentence, 17 words appeared in 2 sentences, 10 words appeared in 3 sentences, 6 words appeared in 4 sentences, 2 words appeared in 5 sentences, 1 word appeared in 9 sentences and lastly, 1 word appeared in 19 sentences.

These are important numbers as in any classification system, the number of samples per class are impotent to report. Recall that each sentence is repeated around 19 times. So, if a word appears in three sentences then the total number of repetitions of that particular word is 3x19 times.

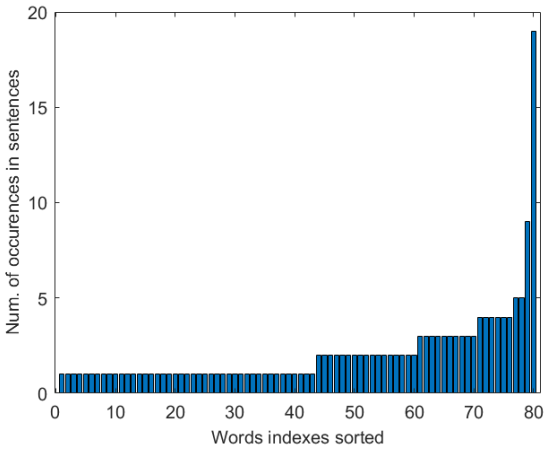


Figure 1. Number of word occurrences in sentences

Likewise, we carried out statistics to find the average and standard deviation of the number of video frames required for each sign language word. The statistics are reported in Figure 2.

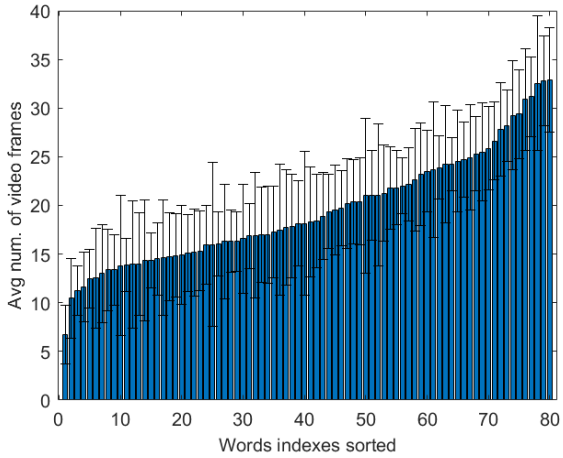


Figure 2. Average number of video frames per sign language word

It is shown in the figure that the at the average number of video frames per sign language word varies from 6 to 30 frames with a mean value of 18 frames. The variance in the number of frames across words is expected as some sign language words are shorter than others. On the other hand, the variance in the number of frames within one word is due to the speed at which the signer acts the sign.

IV. Proposed solution

In this work, to recognize a sign language sentence, a two-stage solution is proposed. In the first stage, we train and use a sequence-to-label classifier to predict the number of words in a sentence. In the second stage, the predicted number of words is used to temporally segment the sentence and use a sequence-to-sequence classifier to recognize the sign language sentence. The system overview is further illustrated in Figure 3. In this section, we start by making the assumption that the number of sign language words are known and explain in details the proposed feature extraction and classification solutions. In Section V, the solution is taken one-step further by introducing a solution that predicts the number of sign language words in a sentence using a sequence-to-label classifier.

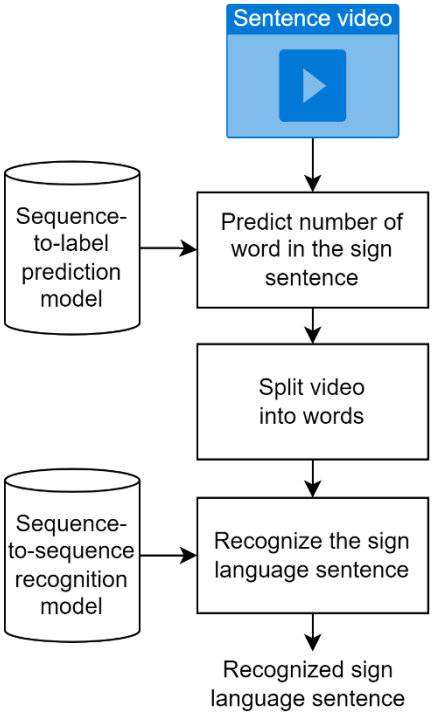


Figure 3. Proposed sequence-to-sequence training for continuous sign language recognition

With more details, to recognize a sign language sentence, it is first segmented into words and each word is represented into what we refer to as a motion image. Each motion image is then represented as one feature vector using a pertained CNN Inception-v3 network. Thus, the sentence becomes a sequence of feature vectors that can be fed into a sequence-to-sequence deep learning network that uses LSTM or biLSTM layers. Other pertained CNN networks can be used for feature extraction as well, however Inception-v3 is used as it generated the best results for our work.

Clearly, the sign language recognition step is preceded by a model generation step that uses labeled training data to form a feature matrix and it corresponding labels that represent the sign language words.

To create the proposed motion images we use a novel solution in which motion estimation is carried out between consecutive video frames belonging to a sign language word. The resultant motion vectors are used to subtract each frame from it preceding motion-compensated frame. All differences are then summed into one motion image. Motion estimation is used to generate the motion vectors using either block-based motion estimation or optical flow. In this work motion estimation is performed using optical flow and motion compensation is a video compression technique in which a frame content is displaced by a negative value of the motion vector pointing to it. As such,

when a frame is subtracted from it preceding motion-compensated frame, each pixel is subtracted from a pixel in the preceding frame located at the best match location found by the motion estimation process. Mathematically, the motion image is represented as:

$$Motion\ image = \sum_{i=1}^{N} sub^i \quad (1)$$

Where N is the number of video frames in a given sign language word and sub^i represents the result of subtracting two video frames at indices i and $i+1$ using motion compensation, more specifically:

$$sub_{x,y}^i = frame^i(x,y) - frame^{i-1}\left(x - Vx_{x,y}^{i-1}, y - Vy_{x,y}^{i-1}\right) \quad (2)$$

Where Vx^{i-1} *and* Vy^{i-1} are the x and y motion vector components of $(i\text{-}1)^{th}$ image and w and h are image width and height respectively. The index x runs from 1 to the width of the frame and the index y runs from 1 to the height of the frame.

An example of summed image differences with motion estimation and compensation are shown in the top 4 images of Figure 7 below. The images shown belongs to the sentence "I am going to the soccer club" which is represented in 4 Arabic sign language words.

According to Equations 1 and 2, which are used to compute motion images, consecutive images are subtracted by means of motion compensation using the generated motion vectors. As such, only the motion part is retained in the resultant motion images, which is the important part needed for gesture recognition. Additionally, the number of motion images is set according to the number of words in a sentence thus generating the same number of motion images regardless of the number of video frames in a sentence.

Once a sign language sentence is segmented into words and each word is converted into a motion image, a recurrent sequence-to-sequence neural network is trained as illustrated in Figure 4.

As illustrated in the figure, each motion image is converted into a feature vector using CNN transfer learning. These feature vectors are time-dependent and are excellent candidates for a recurrent neural network. In this work, we experiment with deep learning architectures that contain LSTM and biLSTM layers. The model is trained in a sequence-to-sequence manner as the input is a sequence of time-dependent feature vectors and the desired output is a sequence of sign language words.

Up to this point, there is one aspect that needs to be addressed in the proposed system which is related to segmenting the sign language sentences into words. This can be done during model generation as the labels or the ground truth of individual video frames is available, however, during the testing stage, the ground truth is not available and thus a solution is needed to segment sentences into words. In the next section, we proposed a prequel stage to the proposed system in which the number of sign language words is predicted and the sentence are segmented accordingly prior to both model generation and testing.

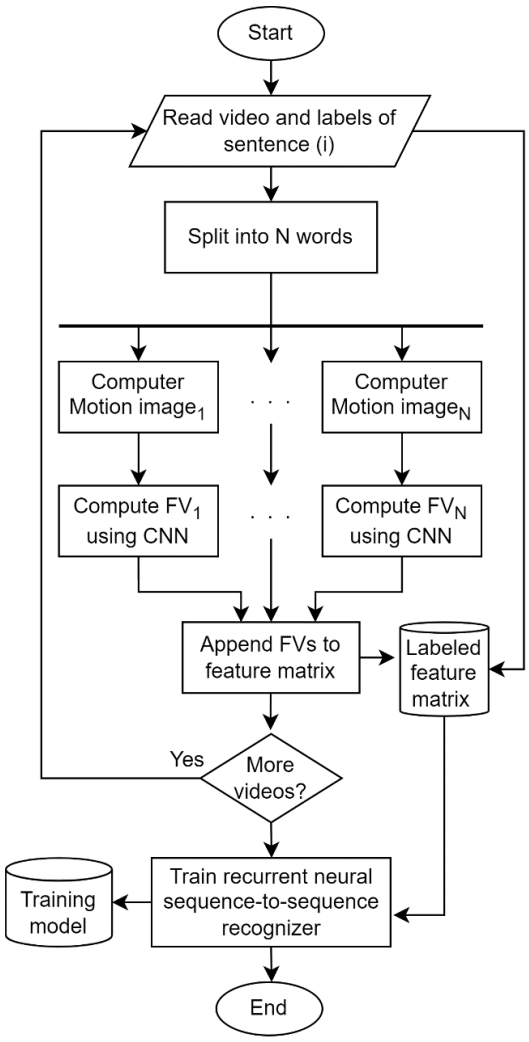


Figure 4. Proposed sequence-to-sequence training for continuous sign language recognition

It is worth noting that the proposed solutions are developed for sentence-based Arabic sign language recognition, and therefore, no claims are made about the suitability of these solutions for other languages without further experimentations and fine-tuning.

Additionally, no claims are made about the suitability of the proposed solution for real-life deployment of sign language recognition. Verifying such a challenging claim entails collecting sign language sentences with covariates including indoors, outdoors, lighting conditions, distance from the camera, resolution of the camera, tilt of video capturing, signing speed, signing accuracy, non-stationary backgrounds and so forth.

V. Predicting the number of words

In the proposed solution of Section 4 above, we made the assumption that the number of sign language words are known in each sentence. However, in real life, this is not acceptable as the number of words is unknown in test sentences that are unlabeled. In general, the word boundaries cannot be

automatically detected in sign language sentences, as the hands motion is continuous.

Therefore, in this work, we propose a classification system that predicts the number of words per sentence. This will be a sequence-to-label classifier that precedes the recognition of sign language sentences proposed in Section 4 above.

In this proposed solution, the input video frames are paired in an overlapping manner, and the image difference with motion compensation is computed similar to the solution proposed in Section 4 above. The reason for computing motion images on pairs of frames here is that the number of words per sentence is unknown. Inception-v3 is used to compute the FVs of each frame pair constituting the input to the training network. The training labels or targets here are simply the number of words per sentence, hence this is a sequence-to-label model generation.

Once the number of words per sentence is predicted, the sign language sentence is split into frame segments of equal sizes, with a total number of segments equal to the number of predicted words. In the experimental results section, we show that this solution works well despite the variation in the true number of frames per word as shown in statistics presented in Figure 2 above.

The motion images computed from image pairs can also be combined to form three or six equal splits of the sign language sentences. In this work, we found that combining the image pairs into six splits generates the best results for predicting the number of words. In Figure 6, the confusion matrix of prediction the number of words is shown. The deep learning architecture used and the train parameters are presented in the next section.

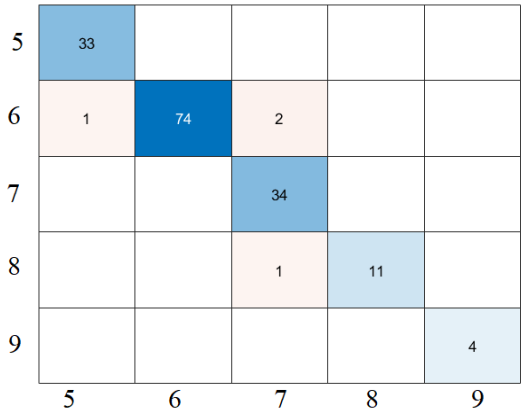
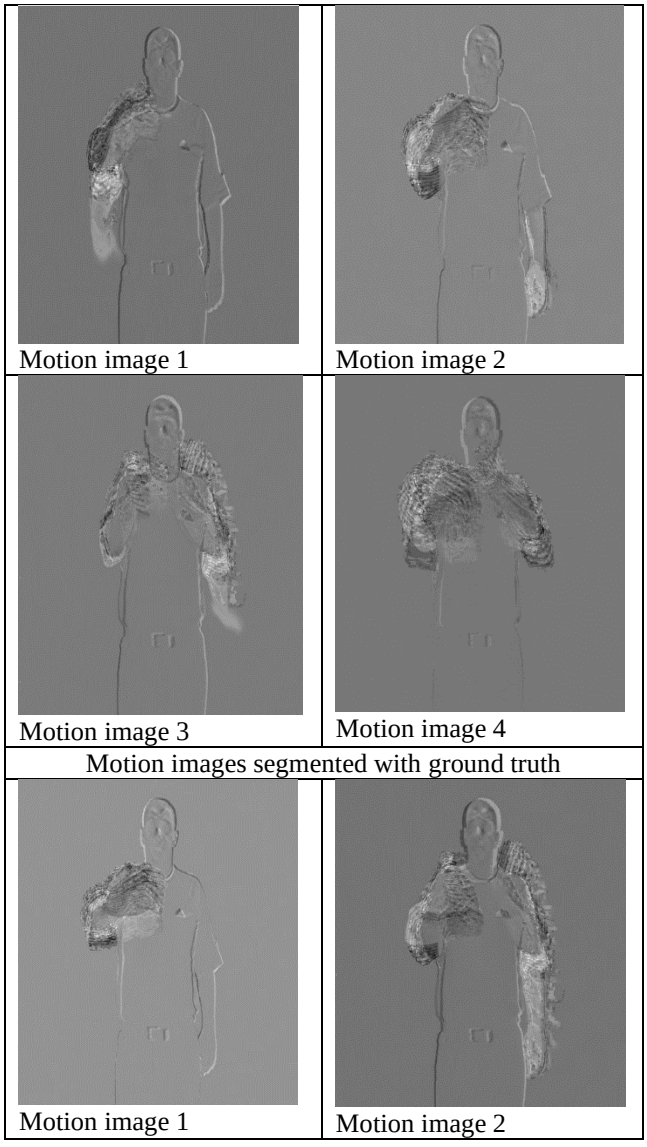


Figure 6. Confusion matrix of predicting number of words in a sentence. The numbers 5 to 9 are the number of words in sentences

The accuracy is 97.5%, however, further investigation revealed that the prediction's inaccuracy necessitates one of two sign language recognition errors, word deletion and word insertion. These errors are represented as values below and above the diagonal in the confusion matrix respectively. The word deletion cannot be detected and it is considered an error in sign language recognition. However, overestimating the number of words in a sentence results in sign word replication that can be detected as a post recognition process. Consequently, the error rate is reduced from 2.5% to 1.25% as shown in the values below the diagonal of the confusion matrix. The sign language word and sentence recognition rates are provided in the experimental results section.

Using the proposed solution of Section 4, In Figure 7, we present an example sign language sentence with summed image differences using motion estimation and compensation (i.e. motion images). The motion images shown belongs to the "I am going to the soccer club" sentence that is represented using four Arabic sign language words. The top 4 motion images contains words segmented with manually labeled data, while the second row contains motion images with automatically segments words using the proposed solution of this section.



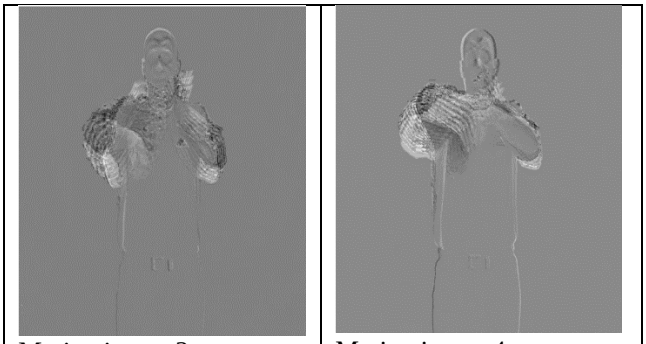


Figure 7. Motion images: Summation of image differences after motion estimation and compensation.

Using the true labels for segmenting the sentences results in one motion image per word as shown in top four motion images of Figure 7. Whereas, using the proposed solution, each motion image can contain traces of a previous or a successive word as shown in the bottom 4 motion images. This byproduct of the proposed solution is a great advantage as it puts words into context, thus justifying the excellent classification accuracies of sign language recognition as shown in the experimental results section.

To summarize, the proposed solution of sentence segmentation employs the proposed feature extraction technique that uses motion estimation and motion compensation to create motion images. These images are used to create a sequence-to-label classifier that predicts the number of sign language words in a sentence. Consequently, the sentence is split into equal-sized segments and each segment is classified as a sign language word. It is noticed that these segments can contain traces of a previous or a successive words as the number of video frames varies from one word to the other. This is an advantage as it puts sign language words into context. The experimental results show that this proposed solution works well when integrated with sentence-based recognition. Although, the dataset used contains only 80 words and 40 sentences, the success of this solution cannot be a coincidence as the number of video frames per word in this dataset varies from 7 to 33 as shown in Figure 2 above. Nonetheless, it is not claimed that the proposed solution generalizes to larger datasets until such datasets are compiled and experimented with.

Now that the two stages of the proposed system are introduced, we provide the overall system block diagram in Figure 8.

As illustrated in the figure, the first stage makes use of a sequence-to-label classifier that predicts the number of sign language words in a sentence and the second stage segments the sentence accordingly and makes use of a sequence-to-sequence classifier to recognize the sign language words.

Lastly, the architecture and parameters used in the proposed sequence-to-label and sequence-to-sequences classifiers to predict the number of sign language words and recognize sentences respectively, include a LSTM layer with 2048 nodes, followed by a dropout layer of 50%, followed by a fully connected, software and classification layers. The minimum batch size is 32 and the maximum epochs are 100. The number of iterations per epoch is set to the number of feature vectors divided by the minimum batch size. The Adam optimizer is used with an initial learn rate of 1e-4.

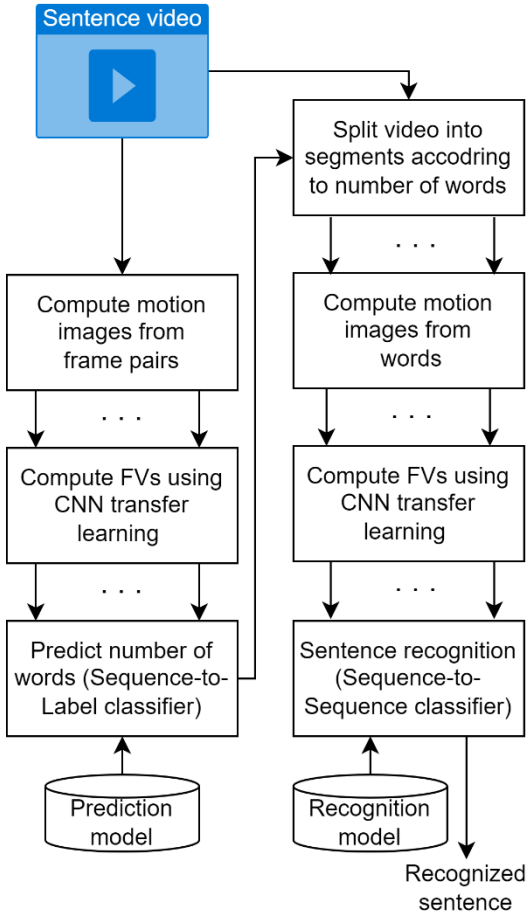


Figure 8. Block diagram of the proposed two-stage recognition system

VI. Deep Learning Architectures

As introduced in the previous sequences, the input video is converted into motion images, which are then converted into a sequence of FVs. Hence, in this work, we use LSTM or biLSTM layers in our recurrent network. We experiment with one LSTM layer, one biLSTM layer, two biLSTM layer and three biLSTM layers as well. An example deep learning architecture using two biLSTM layers is presented in Figure 9. The parameters used include 50% for the dropout layers, the min batch size was set for 64 and the max epochs to 100. The Adam optimizer is used with an initial learn rate of 1e-4.

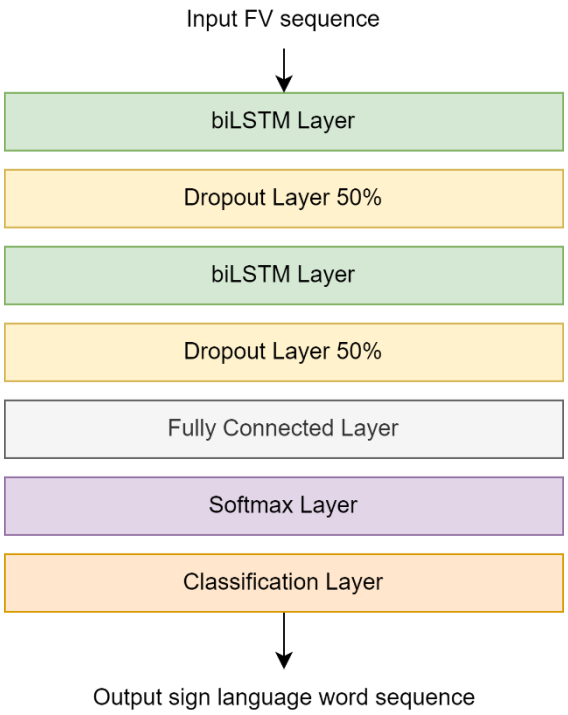


Figure 9. Example deep learning architecture using two biLSTM layers

VII. Experimental Results

This section presents the experimental results of the proposed solution using a number of deep learning configurations. Additionally, the proposed solution of automatic prediction of number of sign language words is compared against the case of full manual labeling. We also compare our results against existing literature that used the same dataset.

The metrics used in quantifying the experimental results contain the standard word recognition rates and sentence recognition rates. The former is calculated as:

$$Word\ recognition\ rate = 1 - \frac{D+S+I}{N} \qquad (3)$$

Where D, S and I are the number of word deletions, substitutions and insertions respectively, and N is the total number of words. On the other hand, the sentence recognition rate metric is more strict as a recognized sentence is considered correct only if it is exactly as that of the ground truth without any insertion, deletion or substitution. Therefore, the sentence recognition rates are lower than the word recognition rates, which might justifies why it is not commonly used in the literature.

Figure 10 presents the word recognition rates using the proposed two-stage solution in which the number of sign language words are predicted first followed by sign language recognition. The results are compared against the case of using the manually generated class labels for detecting word boundaries. Both approaches use the proposed solution of Section 4, however they differ in terms of segmenting the sentence into sign language words as clarified in Section 5 above.

The results include a number of observations. First, biLSTM resulted in higher recognition accuracies than LSTM. This is expected as the former considers the context of sign language words in both the forward and backward directions. Second, with two and three layers of biLSTM, we found that the proposed solution of Section 5 resulted in close recognition accuracies compared to the solution that uses manual labeling for segmenting sentences. The highest recognition rate for the proposed two-stage solution is this case is 97.3%.

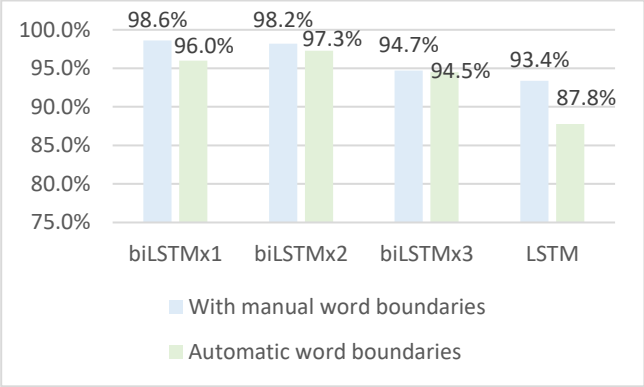


Figure 10. Word recognition rates of proposed two-stage solution versus recognition rates resulting from manual sentence segmentation.

In Figure 11, we repeat the same experiment but we use sentence recognition rates instead of word recognition rates. It is found that with two biLSTM layers, the proposed solution of Section 5 results in slightly higher recognition accuracies than the solution with manual labeling for segmenting sentences. The highest recognition rate for the proposed two-stage solution in this case is 92.6%. Again, this accuracy is clearly lower than the word recognition rates reported in Figure 10 as the metric of sentence recognition rates counts a sentence as correctly recognized only if it is identical to the ground truth.

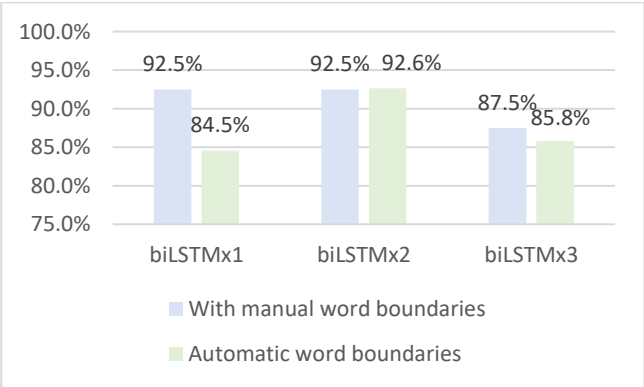


Figure 11. Sentence recognition rates of proposed two-stage solution versus recognition rates resulting from manual sentence segmentation.

As mentioned in Section 5, once the number of words per sentence is predicted, the sentence is segmented into frame segments of equal sizes. This results in a total number of

segments equal to the number of predicted words. This solution resulted in excellent classification accuracy as shown in figures 9 and 10 despite the variation in the true number of frames per word as shown in statistics presented in Figure 2 above. The number of frames in the equal size segments and the true number of frames per word are different, hence, the segments presented to the classifier in the proposed solution can contain a whole word and/or parts of the preceding and successive words as illustrated in Figure 7 above. Surprisingly, all of these scenarios are handled well by the biLSTM layers in the sequence-to-sequence classifier. Using the proposed solution, each motion image can contain traces of previous or successive words. This byproduct of the proposed solution is advantageous as it puts words into context, thus justifying the excellent sign language recognition rates as presented in Figures 10 and 11. This is an important conclusion as otherwise, all video frames belonging to a sentence need to be labeled manually which is an exhausting and a non-scalable task. In the proposed solution, on the other hand, it is enough to know the meaning of the sign language words for model generation and testing without the needs for labeling each video frame.

Additionally, we compare our work against the results reported in [17] and [18] as both used the same dataset. In [17], as mentioned in the introduction, a modified KNN algorithm is proposed which is suitable for image sequences and in [18] two different implementations of Hidden Markov Models are used, G2k [23] and RASR [24].

Figure 12 presents the comparison against existing work in terms of word recognition rates. It is shown that the proposed solution has higher recognition accuracy with a percentage increase of 6.2%, 3.5% and 1.8% in comparison to [17] and [18] using G2k and RASR respectively. The percentage increase in accuracy is calculated as (accuracy of proposed work – accuracy of reference)/ (accuracy of reference) * 100.

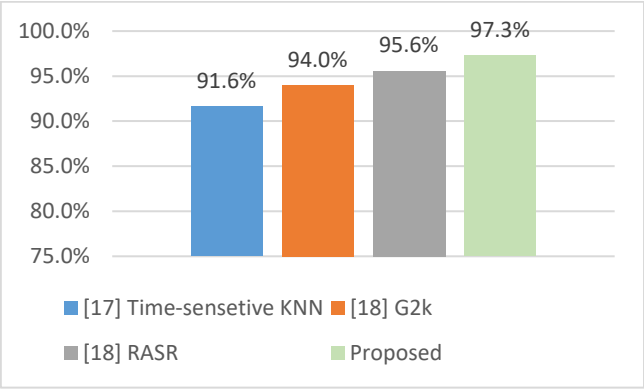


Figure 12. Comparison with existing work in terms of word recognition rates.

Likewise, Figure 13 presents the comparison against existing work in terms of sentence recognition rates. It is shown that the proposed solution has higher recognition accuracy with a percentage increase of 9.1%, 21.6% and 14.5% in comparison to [17] and [18] using G2k and RASR respectively. The percentage increase in accuracy is calculated as mentioned above.

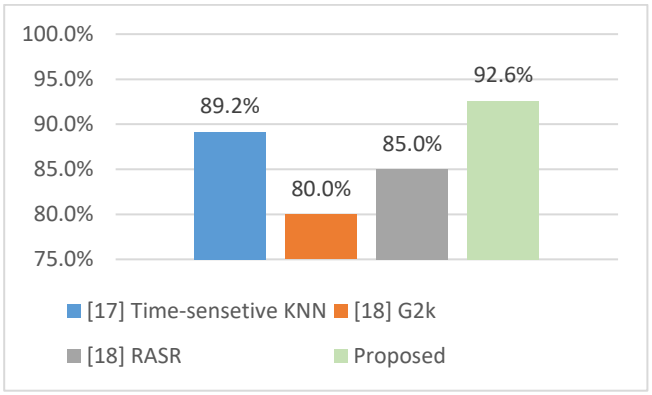


Figure 13. Comparison with existing work in terms of sentence recognition rates.

Clearly, in real-life applications, the reported recognition rates will be lower as this work used an existing dataset collected in a lab environment. Extending this work to real-life applications requires collecting a dataset with a complete Arabic sign language vocabulary. It also requires capturing data with many variants that resemble real life including non-stationary backgrounds, the existence of more than one person in the scene, distance of the signer form the camera, tilt of the camera, illumination of the scene and so forth. Additionally, the work can be extended to signer-independent recognition mode in which data needs to be collected from hundreds of signers.

Moreover, although this work focuses on recognizing sentence-based Arabic sign language, nonetheless, for comparison, we provide a summary of word recognition rates of other sign languages in Table III.

TABLE III.
Summary of Word Recognition Rates of sentence-based sign language recognition for other languages

Reference	Language	Dataset	W. Recognition Rate
HLSTM-att [26]	Chinese	CSL[25]	89.8%
Align-jOpt [27]	Chinese	CSL[25]	93.9%
DPD [28]	Chinese	CSL[25]	95.3%
Multi-Stream CNN-LSTM-HMMs [30]	German	RWTH-2014T [29]	73.5%
Cross-modal alignment [31]	German	RWTH-2014T [29]	75.7%
key frame extraction [32]	Indian	10 sentences	91.3%

The reported work in Table III is all video-based. The Chinese dataset reported in [15] contains 100 sentences and the German dataset reported in [19] contains around 600 sentence, hence the rather low word recognition rates. However, the dataset contains videos of isolated gestures, which makes the training process easier. The word recognition rates reported for the CSL dataset are consistent with what is reported in this work, in fact

the recognition rates of the reported work is lower as the number of sentence is higher.

Lastly, for completeness, we present the train and test times of the proposed solution, although we do not claim that this work is suitable for real-time sign language recognition. The results are generated using MATLAB R2021b and the time is measured using its *cputime()* function. The machine used to generate the results runs Windows 10 with a 10th gen Intel Core i9 processor, 16GB RAM and NVIDIA Quadro T2000 GPU.

The proposed feature extraction of this work includes motion estimation and compensation hence it can result in high computational time. In this work, the average time required to extract features from one sign language word is 1.4 seconds. This is with the fact that we are using the full temporal resolution of 25 frames per second. Thus, for a sentence with 6 words, the average time required for feature extraction is around 8.4 seconds. This computational time can be easily reduced by spatio-temporal subsampling of the input videos. However, this can have an impact on the overall recognition accuracy that needs to be investigated in future work. On the other hand, the time required for sign language recognition using deep learning is much lower than the feature extraction time as reported in Table 4.

TABLE IV.
TRAIN AND TEST TIME IN SECONDS PER SIGN LANGUAGE SENTENCE.

Net Architecture	Train (sec/sentence)	Test (sec/sentence)
LSTM	0.7284	0.0036
biLSTM	0.9555	0.0041
biLSTMx2	1.3833	0.0044
biLSTMx3	1.7174	0.0044

The results in Table 4 present the model generation and recognition times in seconds per sign language sentence. It is shown that the train and test times increase when biLSTM is used and it also shown that the train time is a function of the total number of biLSTM layers, which is expected. It is also shown that the recognition time is very fast and is performed in a fraction of a second per sign language sentence. Reducing the computational time and making the proposed solution suitable for real-time recognition of sign language remains the topic of a future work.

VIII. Conclusion

This worked focused on vision-based recognition of continuous Arabic sign language in user-dependent mode. One challenge with such an approach is knowing the number of words in a sentence as video frames are continuous and there are no pauses between words. A two-stage solution was proposed in which the number of words is predicted first, followed by the second stage in which individual words are recognized. The solution is based on computing motion images based on motion estimation and motion compensation. Each motion image was converted into a motion vector using CNN transfer learning prior to model generation and testing.

The novelty of the proposed solution pertains to providing a statistical insight into the sentence-based dataset used. The paper also proposed the use of motion compensation in the formation of motion images used in both the prediction of the sign language words and the sign language recognition. Existing techniques on gesture-based sign language recognition can benefit from the proposed use of motion compensation in the formation of motion images and existing techniques on sentence-based sign language recognition can benefit from the proposed solution in predicting the number of words in a sentence.

Experimental results revealed that the proposed solution works well when applied to a dataset composed of 40 sentences. The word and sentence recognition rates were 97.3% and 92.6% respectively.

The proposed solution of predicting the number of words per sentence and splitting the sentence into equal segments accordingly worked well. Consequently, each motion image potentially contained traces of previous or successive words. This byproduct of the proposed solution is advantageous as it puts words into context, thus justifying the excellent sign language recognition rates. This is an important conclusion as otherwise, video frames belonging to a sentence need to be labeled manually which is an exhausting and a non-scalable task. In the proposed solution, on the other hand, it is enough to know the meaning of the sign language words for model generation and testing without the needs for labeling each video frame. Lastly, the computational time of the proposed feature extraction, model generation and testing were presented. It was mentioned that reducing the computational time and making the proposed solution suitable for real-time recognition of sign language remains the topic of a future work. Additional future work also include collecting and labeling a larger sentence-based dataset with more than one signer.

ACKNOWLEDGMENT

The work in this paper was supported, in part, by the Open Access Program from the American University of Sharjah, award number OAPCEN-1410-E00215. This paper represents the opinions of the authors and does not mean to represent the position or opinions of the American University of Sharjah.

REFERENCES

[1] T. Shanableh, K. Assaleh and M. AL–Rousan, “Spatio–Temporal feature extraction techniques for isolated Arabic sign language recognition,” IEEE Transactions on Systems, Man and Cybernetics Part B, 37(3), June, 2007

[2] M. Alzubaidi, M. Otoom and A. Rwaq, “A Novel Assistive Glove to Convert Arabic Sign Language into Speech” ACM Transactions on Asian and Low-Resource Language Information Processing. 22.10.1145/3545113, 2022

[3] M. Mohandes, S. Aliyu and M. Deriche, "Arabic sign language recognition using the leap motion controller," 2014 IEEE 23rd International Symposium on Industrial

Electronics (ISIE), Istanbul, Turkey, pp. 960-965, doi: 10.1109/ISIE.2014.6864742, 2014

[4] M. A. Bencherif, M. Algabri, M. Amine and M. Faisal, "Arabic Sign Language Recognition System Using 2D Hands and Body Skeleton Data," IEEE Access, vol. 9, pp. 59612-59627, doi: 10.1109/ACCESS.2021.3069714, 2021

[5] Z. Alsaadi, E. Alshamani, M. Alrehaili, A. Alrashdi, S. Albelwi, A. Elfaki, "A Real Time Arabic Sign Language Alphabets (ArSLA) Recognition Model Using Deep Learning Architecture," Computers. 11(5):78. https://doi.org/10.3390/computers11050078, 2022

[6] S. M. Miah, M. A. M. Hasan and J. Shin, "Dynamic Hand Gesture Recognition Using Multi-Branch Attention Based Graph and General Deep Learning Model," in IEEE Access, vol. 11, pp. 4703-4716, 2023.

[7] W. Abdul, M. Alsulaiman, S. Umar Amin, M. Faisal, G. Muhammad, F. Albogamy, M. Bencherif, H. Ghaleb, "Intelligent real-time Arabic sign language classification using attention-based inception and BiLSTM," Computers and Electrical Engineering, Volume 95,107395,ISSN 0045-7906,https://doi.org/10.1016/j.compeleceng.2021.107395, 2021

[8] K. Assaleh, T. Shanableh, M. Fanaswala, F. Amin, H. Bajaj," Continuous Arabic Sign Language Recognition in User Dependent Mode," Journal of Intelligent Learning Systems and Applications, 2(1), DOI: 10.4236/jilsa.2010.21003, 2010

[9] T. Shanableh and K. Assaleh, "User-independent recognition of Arabic sign language for facilitating communication with the deaf community," Digital Signal Processing, Elsevier, 21(4), July, 2011

[10] M. Alsulaiman, M. Faisal, M. Mekhtiche, M. Bencherif, T. Alrayes, G. Muhammad, H. Mathkour, W. Abdul, Y. Alohali, M. Alqahtani, H. Al-Habib, H. Alhalafi, M. Algabri, M. Alhamadi, H. Altaheri and T. Alfaqih, "Facilitating the Communication with Deaf People: Building a Largest Saudi Sign Language Dataset," Journal of King Saud University - Computer and Information Sciences, 101642, https://doi.org/10.1016/j.jksuci.2023.101642, 2023

[11] M. A. Bencherif et al., "Arabic Sign Language Recognition System Using 2D Hands and Body Skeleton Data," in IEEE Access, vol. 9, pp. 59612-59627, doi: 10.1109/ACCESS.2021.3069714, 2021

[12] G. Latif, N. Mohammad, J. Alghazo, R. AlKhalaf and R. AlKhalaf, "ArASL: Arabic Alphabets Sign Language Dataset," Data in Brief, 23: 103777, 2019

[13] H. Luqman, E-SM. El-Alfy, "Towards Hybrid Multimodal Manual and Non-Manual Arabic Sign Language Recognition: mArSL Database and Pilot Study," Electronics, 10(14):1739. https://doi.org/10.3390/electronics10141739, 2021

[14] M. Tolba, A. Samir and M. Abul-Ela, "A proposed graph matching technique for Arabic sign language continuous sentences recognition," 8th International Conference on Informatics and Systems (INFOS), IEEE, MM-14-MM-20, 2012

[15] S.M. Shohieb, H.K. Elminir and A.M. Riad, "Signs world atlas; a benchmark Arabic sign language database," Journal of King Saud University-Computer and Information Sciences, 27, pp. 68-76, 2015.

[16] A. Othman and O. El Ghoul, "Intra-linguistic and extra-linguistic annotation tool for the "Jumla Dataset" in Qatari sign language," 8th International Conference on ICT & Accessibility (ICTA), Tunis, Tunisia, pp. 01-06. doi: 10.1109/ICTA54582.2021.9809778, 2021

[17] N. Tobaiz, T. Shanableh and K. Assaleh, "Glove-Based Continuous Arabic Sign Language Recognition in User-Dependent Mode," IEEE Transactions on Human-Machine System, doi: 10.1109/THMS.2015.2406692, 45(4), August, 2015

[18] M. Hassan, K. Assaleh and T. Shanableh, "Multiple Proposals for Continuous Arabic Sign Language Recognition," Sensing and Imaging, Springer, 20: 4. https://doi.org/10.1007/s11220-019-0225-3, 2019

[19] S. Aly and W. Aly, "DeepArSLR: A Novel Signer-Independent Deep Learning Framework for Isolated Arabic Sign Language Gestures Recognition," in IEEE Access, vol. 8, pp. 83199-83212, doi: 10.1109/ACCESS.2020.2990699, 2022

[20] A. S. Musa Miah, J. Shin, M. A. M. Hasan, M. A. Rahim and Y. Okuyama, "Rotation, translation and scale invariant sign word recognition using deep learning," Computer Systems Science and Engineering, vol. 44, no.3, pp. 2521–2536, 2023

[21] S. Jungpil, A. S. Musa Miah, Md. Hasan, K. Hirooka, K. Suzuki, H.-S. Lee, and S.-W. Jang. 2023. "Korean Sign Language Recognition Using Transformer-Based Deep Neural Network" *Applied Sciences* 13, no. 5: 3029.

[22] M. A. Rahim, A. S. M. Miah, A. Sayeed and J. Shin, "Hand Gesture Recognition Based on Optimal Segmentation in Human-Computer Interaction," *2020 3rd IEEE International Conference on Knowledge Innovation and Invention (ICKII)*, Kaohsiung, Taiwan, 2020, pp. 163-166, doi: 10.1109/ICKII50300.2020.9318870

[23] T. Westeyn, H. Brashear, A. Atrash and T. Starner, "Georgia tech gesture toolkit: Supporting experiments in gesture recognition. In 5th international conference on multimodal interfaces, pp. 85–92. New York, 2003

[24] S. Wiesler, A. Richard, P. Golik, R. Schlüter and H. Ney, "RASR/NN: The RWTH neural network toolkit for speech recognition," 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Florence, Italy, pp. 3281-3285, doi: 10.1109/ICASSP.2014.6854207, 2014

[25] J. Huang, W. Zhou, Q. Zhang, H. Li and W. Li, "Video-based sign language recognition without temporal segmentation", Proc. 32nd AAAI Conf. Artif. Intell., pp. 1-8, 2018.

[26] D. Guo, W. Zhou, H. Li and M. Wang, "Hierarchical LSTM for sign language translation", Proc. 32nd AAAI Conf. Artif. Intell., pp. 1-8, 2018.
[27] J. Pu, W. Zhou and H. Li, "Iterative alignment network for continuous sign language recognition", Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 4165-4174, Jun. 2019.
[28] H. Zhou, W. Zhou and H. Li, "Dynamic pseudo label decoding for continuous sign language recognition", Proc. IEEE Int. Conf. Multimedia Expo (ICME), pp. 1282-1287, Jul. 2019.
[29] N. C. Camgoz, S. Hadfield, O. Koller, H. Ney and R. Bowden, "Neural sign language translation", Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., pp. 7784-7793, Jun. 2018.
[30] O. Koller, C. Camgoz, H. Ney and R. Bowden, "Weakly supervised learning with multi-stream CNN-LSTM-HMMs to discover sequential parallelism in sign language videos", IEEE Trans. Pattern Anal. Mach. Intell., Apr. 2019.
[31] I. Papastratis, K. Dimitropoulos, D. Konstantinidis and P. Daras, "Continuous Sign Language Recognition Through Cross-Modal Alignment of Video and Text Embeddings in a Joint-Latent Space," in IEEE Access, vol. 8, pp. 91170-91180, 2020
[32] K. Tripathi and N. B. G. Nandi, “Continuous Indian sign language gesture recognition and sentence formation,” Procedia Computer Science, vol. 54, pp. 523–531, 2015.

Tamer Shanableh a senior member of the IEEE and a professional engineer. Was born in Scotland, UK. He studied at the University of Essex where he received his Ph.D. in electronic systems engineering in 2002 and his MSc in software engineering in 1998.
He then worked as a senior research officer at the University of Essex for three years, during which, he collaborated with BTexact on inventing video transcoders. He then joined Motorola UK Research Labs and contributed to establishing a new profile within the ISO/IEC MPEG-4 known as the Error Resilient Simple Scalable Profile. He joined the American University of Sharjah in 2002 and is currently a professor of computer science. During the summer breaks, Dr. Shanableh worked as a visiting professor at Motorola Labs in five different years. He spent his sabbatical leave as a visiting academic at the Multimedia and Computer Vision and Lab at Queen Mary, University of London, U.K. Dr. Shanableh has six patents and authored more than 80 publications including 10 IEEE transaction papers. His research interests include digital video coding and processing and pattern recognition.