

ArabSign: A Multi-modality Dataset and Benchmark for Continuous Arabic Sign Language Recognition

Hamzah Luqman

Information and Computer Science Department, King Fahd University of Petroleum and Minerals
SDAIA-KFUPM Joint Research Center for Artificial Intelligence, Dhahran 31261, Saudi Arabia.
Email: hluqman@kfupm.edu.sa

Abstract—Sign language recognition has attracted the interest of researchers in recent years. While numerous approaches have been proposed for European and Asian sign languages recognition, very limited attempts have been made to develop similar systems for the Arabic sign language (ArSL). This can be attributed partly to the lack of a dataset at the sentence level. In this paper, we aim to make a significant contribution by proposing ArabSign, a continuous ArSL dataset. The proposed dataset consists of 9,335 samples performed by 6 signers. The total time of the recorded sentences is around 10 hours and the average sentence’s length is 3.1 signs. ArabSign dataset was recorded using a Kinect V2 camera that provides three types of information (color, depth, and skeleton joint points) recorded simultaneously for each sentence. In addition, we provide the annotation of the dataset according to ArSL and Arabic language structures that can help in studying the linguistic characteristics of ArSL. To benchmark this dataset, we propose an encoder-decoder model for Continuous ArSL recognition. The model has been evaluated on the proposed dataset, and the obtained results show that the encoder-decoder model outperformed the attention mechanism with an average word error rate (WER) of 0.50 compared with 0.62 with the attention mechanism. The data and code are available at <https://github.com/Hamzah-Luqman/ArabSign>

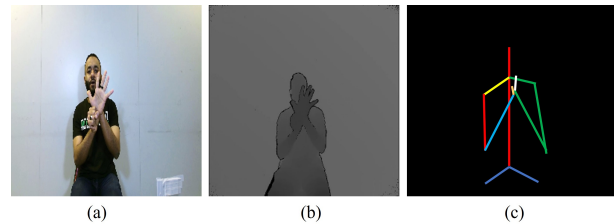


Fig. 1: An illustrative example from the ArabSign dataset for the three modalities provided for each sentence sample: (a) color, (b) depth, and (c) skeleton joint points.

Sign Language (BSL) and American Sign Language (ASL). Other popular sign languages are Chinese (CSL), German (GSL), Indian (ISL), and Arabic (ArSL) sign languages. ArSL is one of the main languages used in Arab countries. It is currently the main language used in translating television programs such as news and interviews. This language has a dictionary consisting of 3200 sign words published in two parts [5], [6].

Sign language is a non-verbal language that uses multi-modality data to express thoughts [15]. Manual and non-manual gestures are the two modalities used in sign language for communication. These gestures are combined during signing in a way that complements each other. Manual gestures are the dominant element used in sign languages. These gestures employ body movements through the hands and head. The majority of sign language signs depend on manual gestures. The non-manual modality consists mainly of facial expressions that are simultaneously performed with manual gestures. Non-manual gestures are used to show emotions and feelings in sign language in addition to linguistic properties such as grammatical structure, adjectival or adverbial content, and lexical distinction.

Translating sign language into spoken language is accomplished through sign language recognition (SLR) and translation [28]. Automatic SLR involves using pattern recognition and computer vision to identify sign gestures and convert them into their equivalent words in the natural language [40]. Sign language translation involves using natural language processing and linguistics to translate the recognized sign language

I. INTRODUCTION

Hearing loss is a serious problem facing the world today, and it is getting worse. It is estimated that nearly 2.5 billion people are projected to have some degree of hearing loss by 2050, and at least 700 million will require hearing rehabilitation [1]. Modern lifestyles and unsafe listening practices mean that over 1 billion young adults are at risk of permanent hearing loss.

Sign language is the main communication language of hearing impaired people. This language is a complete and rich language with grammar and structure that differ from spoken languages. Sign language has its own lexicon that is usually smaller than spoken languages’ vocabulary.

Sign language is not a universal language and it does not depend on spoken languages [4]. Sign languages are “not mutually intelligible with each other” although there are some similarities in some signs. There are many sign languages that differ in their gestures, lexicon, and grammar. Most of the sign languages are related to the country more than the spoken language of that country. There are some countries that speak one language but have different sign languages, such as British

sentences into spoken languages to meet their structure and grammar. Extensive research has been conducted on SLR compared with translation since translation depends on the output of the SLR at the sentence level.

Based on the type of the recognized signs, SLR systems can be categorized into isolated and continuous SLR systems. Isolated sign recognition systems target isolated sign words while continuous sign language recognition (CSLR) systems target more than one sign performed continually. Most of the techniques that have been proposed for SLR during the last three decades have targeted isolated signs [15]. CSLR is still in its infancy compared with isolated SLR, where the growth of CSLR studies is close to linear compared with the exponential growth of isolated SLR studies [25]. One of the challenges associated with CSLR is the lack of movement epenthesis clues between sentence signs and the lack of temporal information that can help in signs segmentation. In addition, the high variance between signs performed by different signers made the learning of segmentation clues very difficult. Another challenge is the lack of datasets, which can be considered the main challenge that makes most of the researchers target isolated signs.

To our knowledge, no available vision-based annotated sentences of ArSL that can be used for ArSL CSLR and translation. The datasets that have been proposed for Arabic CSLR are collected using glove sensors. However, sensor-based SLR requires signers to keep wearing the electronic sensor gloves during signing. This makes these sensors unsuitable for real-time applications. In addition, the sensors used for sign acquisition can not capture the non-manual features of the sign language. This motivated us to propose a continuous ArSL dataset that can be used for CSLR and translation. The main contributions of this research are as follows:

- Propose a continuous ArSL dataset (ArabSign). The proposed dataset was collected using a multi-modality Kinect V2 camera. The dataset is available in three modalities: color, depth, and joint points shown in Figure 1. The proposed dataset consists of 9,335 samples representing 50 ArSL sentences. Each sentence was performed by 6 signers, and each sentence was repeated several times by each signer.
- Provide the annotation of the performed sentences according to the structure of ArSL and Arabic language. This makes the dataset useful for studying the grammar and structure of ArSL and developing machine translation systems between ArSL and natural languages.
- Propose an encoder-decoder model for benchmarking the proposed ArabSign dataset. The model has been trained on features extracted from the color frames of the sentences using different pre-trained models. In addition, the proposed model has been compared with an attention mechanism.

This paper is organized as follows: a literature review of the available continuous sign language datasets is presented in Section II. A detailed description of the proposed ArabSign

dataset is presented in Section III. Section IV describes the experimental work that has been conducted to benchmark the proposed dataset, and the conclusions are presented in Section V.

II. LITERATURE REVIEW

The work on SLR can be dated back to the middle of the 1990s [25]. The SLR systems at the sign level are the most common SRL systems compared with the sentence level due to the availability of datasets at the sign level and the similarity between this problem and gesture recognition problems [15]. In contrast, few approaches have been proposed for CSLR due to the challenges associated with recognizing sign languages' sentences. One of these challenges is the lack of annotated datasets.

Few datasets have been proposed for continuous SL compared with isolated sign datasets. The majority of these datasets target ASL and DGS. There are some datasets that are used by researchers for their work. However, these datasets are either limited in size or unavailable for researchers. The most commonly used continuous sign language datasets were proposed by a group at RWTH Aachen University. This group proposed four datasets for continuous ASL and DGS, namely RWTH-BOSTON-104 [13], RWTH-BOSTON-400 [12], RWTH-PHOENIX-Weather [17], and RWTH-PHOENIX-Weather-2014 [18]. RWTH-BOSTON-104 [13] was recorded at Boston University and it consists of 201 sentences of ASL performed by three signers. The vocabulary size of this dataset is 168 sign words.

RWTH-BOSTON-400 [12] is an extension of RWTH-BOSTON-104. It consists of 843 sentences with a vocabulary size of 406 sign words performed by four signers. RWTH-PHOENIX-Weather [17] includes weather forecasts collected from German television. This dataset is performed by seven signers and it consists of 1,980 sentences of DGS with a vocabulary size of 911 sign words. This dataset is extended in RWTH-PHOENIX-Weather-2014 [18] to 6,861 sentences performed by nine signers. Both datasets were recorded in a controlled environment where signers were wearing a dark T-shirt with grey background. How2Sign [14] is a multi-view ASL dataset consisting of around 35K samples performed by 11 signers for a duration of 79 hours.

SIGNUM [39] is a DGS dataset consisting of 780 sentences performed by 25 signers. The SignsWorld Atlas [35] is an ArSL dataset consisting of five sentences performed by four signers. TheRuSLan [24] is a Russian SL dataset consisting of 164 sentences performed by 13 signers. Huang et al. [22] proposed a CSL dataset consisting of 100 sentences with a vocabulary size of 178 sign words performed by 50 signers. Table I summarizes the available continuous sign language datasets. The missing information in the table is not reported in the respective reference.

Another challenge of CSLR is sign segmentation. This can be attributed to the lack of movement epenthesis clues between sentence's signs and the lack of temporal information that can

TABLE I: A summary of the publicly available continuous sign language datasets.

Dataset	Language	Sentences	Duration (h)	Vocabulary Size	Signers	Samples
RWTH-BOSTON-104 [13]	ASL	400	-	104	3	-
RWTH-BOSTON-400 [12]	ASL	843	-	400	4	-
MS-ASL [23]	ASL	1000	-	-	222	25,513
How2Sign [14]	ASL	-	79	16K	11	35,191
RWTH-PHOENIX-Weather [17]	DGS	1,980	-	911	7	1,980
RWTH-PHOENIX-Weather-2014 [18]	DGS	6,861	-	1,080	9	6,861
SIGNUM [39]	DGS	780	55	455	25	780
TheRuSLan [24]	Russian	164	-	-	13	-
Video-based CSL [22]	CSL	100	100	178	50	5,000
SignsWorld Atlas [35]	ArSL	5	-	-	4	-



Fig. 2: A sample of Al-Jazeera ArSL sentence videos that was followed by the signers of ArabSign dataset [2].

help in sign segmentation [15]. In addition, the high variance between signs performed by different signers made the learning of segmentation clues very difficult. These challenges motivated researchers to use recognition techniques that do not require pre-segmented sentences such as Hidden Markov models [8], [20], [29]–[31], [44] and deep learning techniques [9], [10], [22], [27], [32], [34], [38], [45], [46]. However, some researchers converted the CSLR problem into an isolated signs recognition problem by segmenting the sentences into signs and recognizing each sign separately [16], [19], [26], [36], [42]–[44].

III. ARABSIGN DATASET

The ArabSign dataset consists of ArSL video sentences with their corresponding annotations in Arabic and English languages. A total of 10 hours and 13 minutes of ArSL were recorded by 6 signers in different recording sessions. The sentences used in this dataset are based on the ArSL tutorial videos performed by ArSL translation experts and produced by Al-Jazeera media network [2]. We extracted fifty most commonly used sentences in ArSL from these videos. Then, we used these sentences as a reference for our signers, and each signer was asked to repeat those sentences. The reference sentences will be published with the dataset for interested researchers. A sample of a reference video is shown in Figure 2.

All the sentences of the proposed dataset were annotated with their glosses in both Arabic and English languages. The

annotation followed the structure of the ArSL sentence as performed by the signer. Figure 3 shows an example of an ArabSign sentence with its glosses and its equivalent in Arabic language. To make the dataset useful for machine translation tasks between ArSL and Arabic language, we also provide the annotation of each sentence according to the structure and grammar of Arabic language. This helps in studying the linguistic characteristics of ArSL.

A. Recording setup

The ArabSign dataset was performed by 6 signers. Three of these signers have experience with sign language and have been involved in sign language gesturing before. The other signers were trained on sign language during the project. The signers first watched the sentences video performed by an expert in sign language translation who is working in Al-Jazeera news media. The video was played at a slow speed of 0.75 to help the signers in learning the sentence’s signs correctly. Each sentence was transcribed to show the sign of each word to help the signers be familiar with the sentence glosses. The signers were asked to repeat each ArSL sentence several times before the recording session to minimize the variations between sentence samples.

The dataset’s sentences have been recorded in several sessions. Each signer was asked to wear casual clothes and use different clothes in each recording session to add more variability to the dataset. The dataset was recorded in an unconstrained environment. We used the room lights during the recordings. The sessions were recorded at different times of the day. In addition, all the sentences were recorded with a white background.

The dataset was recorded using Microsoft Kinect V2. This camera is a multi-modality camera that provides three types of information recorded simultaneously for each sentence. It provides color, depth, and skeleton information. The color video is recorded with a resolution of 1920×1080 at 30 fps. The depth information is available as a video stream with a resolution of 512×424 . Kinect V2 provides 25 joint points for each signer. The joint points were captured for each frame and are available in Matlab file format. This file contains the position of each joint point in the camera space, represented using X, Y, and Z coordinates. It also contains the orientation,

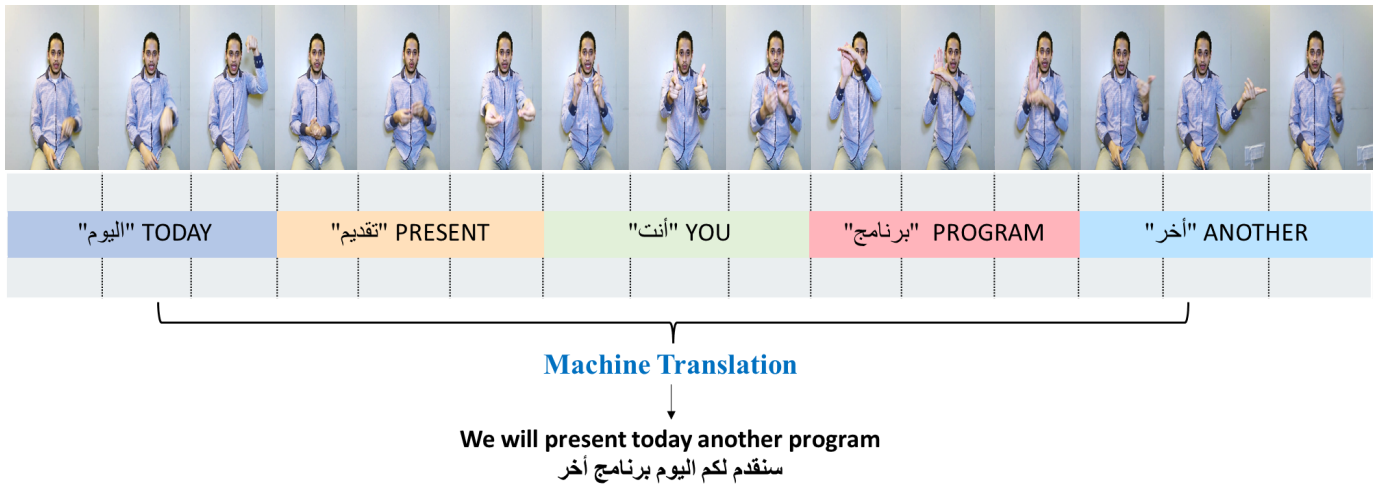


Fig. 3: ArabSign sentence with its glosses and the corresponding sentence resulting from machine translation systems.

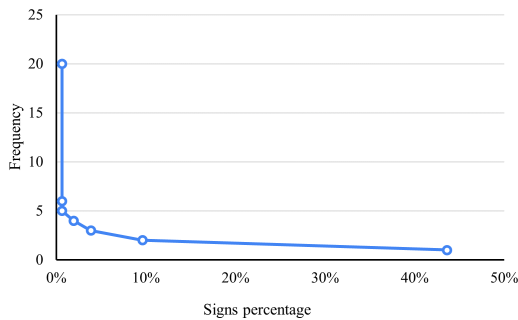


Fig. 4: Signs' frequency in the dataset.

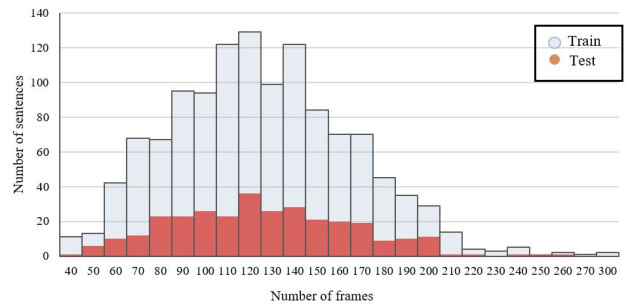


Fig. 5: Frames' frequency over sentences.

tracking state, left and right hands' states, and the cardinal position of each joint point in the color and depth videos.

B. Dataset statistics

The dataset consists of 9,335 samples representing 50 sentences of ArSL. The dataset's sentences were performed by 6 signers. Each sentence was repeated by each signer at least 30 times at different sessions. All signers are male with different skin colors. The signers' ages range between 21 and 30 years old. All signers are right-handed, and one of them was wearing eyeglasses.

The dataset's sentences consist of 155 signs, and the dataset's vocabulary consists of 95 signs. Figure 4 shows the frequency of the signs in the dataset. As shown in the figure, more than 40% of the dataset signs appeared less than 5 times. Having a large number of unique signs or signs that appear a few times makes the dataset appropriate for evaluating real-time recognition systems.

The total time of the recorded sentences is around 10 hours and 13 minutes. The duration of each sentence depends on its length in terms of the number of signs and the signer's signing speed. The average sentence length is 3.1 signs. The dataset was recorded at the normal speed of signing. Each sentence

TABLE II: Statistics of the proposed ArabSign dataset

RGB resolution	1920×1080	# of signers	6
Depth resolution	512×424	Vocab. size	95
Body joints	21	Average words/sample	3.1
Min. video duration (sec.)	1.3	Repetitions/sentence	≥30
Max. video duration (sec.)	10.4	FPS	30
Total hours	10.13	Total Samples	9,335

was signed continuously with no pauses between sentence's signs. This resulted in around 200,000 frames for all sentences performed by one signer, with an average of 130.3 frames per sentence, as shown in Figure 5. This figure shows the frames over sentence level clips for one signer. Table II shows the statistics of the proposed ArabSign dataset.

IV. EXPERIMENTAL EVALUATION

A. System Architecture

In this work, we propose a continuous ArSL recognition framework for benchmarking the ArabSign dataset. The architecture of the proposed framework is shown in Figure 6. The features are extracted from sentence videos using pre-trained models and fed into an encoder-decoder model for features learning and classification. To learn the corresponding natural language text of the recognized sentences, we employed word

embedding. This section describes the components of the proposed framework for CSLR.

1) *Features Extraction*: Sign language learning depends on two types of information, spatial and temporal. Spatial information in sign language represents the shape and orientation of the body parts used during signing, such as the hand, head, and mouth. The temporal information of sign gestures involves the motion of the signer’s body parts during signing.

Spatial information is important for sign language understanding. Learning these features efficiently contributes to improving the model accuracy. Although time-series learning techniques, such as Long short-term memory (LSTM) and Gated recurrent unit (GRU), are efficient for temporal information learning, they lack the ability to learn spatial information [41]. To address this issue, we employed CNN models for features extraction.

Two pre-trained CNN models are used in this work for extracting the spatial features from the sign frames. These models are MobileNet [21] and InceptionV3 [37]. All of these models have been trained originally on the ImageNet dataset, which consists of images grouped into 21,841 subcategories. Although these models have been trained on ImageNet, the performance of each model can vary due to the structure and specifications of the model. These differences make each model appropriate for different computer vision tasks.

Different number of features were extracted from each model. For each frame in the sign gesture, 1024 features were extracted using the MobileNet pre-trained model. We also extracted 2048 features from each frame using the InceptionV3 model. The variation in the number of extracted features is related to the architecture of each model. These features are used as inputs to the proposed models.

2) *Words Embedding*: The CSLR system accepts the sentence gestures as a sequence of frames and outputs their equivalent glosses in the form of an ArSL sentence. The sentence frames are fed into the encoder, and the ground truth of the sentence will be fed into the decoder during the training stage of the proposed models. The ground truth is an Arabic sentence consisting of a sequence of glosses representing the signs of the sentence, as shown in Figure 3.

Several techniques have been proposed in the literature for word representation, such as TF-IDF and N-gram. These techniques ignore the word context that can affect the performance of several natural language processing systems. Therefore, word embedding techniques have been proposed recently to address this issue [3]. These embeddings played an important role in boosting the performance of several machine learning models [11]. In this work, we used an embedding layer to represent each sentence’s word as a vector of size 300 and used these vectors as an input to the decoder.

3) *Encoder-Decoder model*: An encoder-decoder model has been proposed in this work. The structure of the proposed model is shown in Figure 6. The encoder component accepts the features extracted from each sign’s frame using the pre-trained models discussed in Section IV-A1. Each feature vector is fed into the encoder component, which consists of

three bidirectional GRU layers separated by a dropout layer. Each GRU layer consists of 1024 neurons. To mitigate the overfitting, we used two dropout layers with a probability of 0.4. These hyper-parameters have been selected empirically.

The decoder component of the model is responsible for generating the equivalent sentence of the sign sentence video. This model accepts two inputs during training. The first input is the output of the encoder component, and the second input is the sentence’s ground truth. The ground truth input is represented as a sequence of word glosses. These words are fed first into a word embedding layer and the output of this layer is used as an input to the decoder, as illustrated in Figure 6. The decoder model consists of three stacked GRU layers with 1024 neurons, selected empirically. The outputs of the first two GRU layers are fed into dropout layers with a probability of 0.4. The output of the model is a sequence of word glosses representing the recognized sentence.

We also proposed a variant of the encoder-decoder model with attention layer. We will refer to this variant by attention model. The attention mechanism is a deep learning technique that provides more focus on some components of the input. Attention was proposed first for neural machine translation [7]. Later, this mechanism, or its variants, was used in other applications, including speech processing, computer vision, etc.

The attention model consists of four components. The convolutional component which is used to extract features from each sign frame, as discussed in Section IV-A1. The feature vector of each frame is fed into the encoder component, which consists of three stacked bidirectional GRU layers separated by two dropout layers. The architecture of the encoder component is similar to the encoder-decoder model discussed before.

The output of the encoder component is fed into the decoder and attention fusion components. We used the attention fusion module as a layer to obtain the attention weights, which were multiplied by the encoder’s output to obtain the attended encoder output. We use this attended encoder output as a context tensor, which represents a weighted sum indicating which parts of the encoder’s output to pay attention to. The output of the decoder with attention is used to predict the corresponding word gloss.

B. Results and Discussion

Several experiments have been conducted to evaluate the proposed models and benchmark the proposed ArabSign dataset. We ran all the experiments using Pytorch 1.12 on a workstation with an Nvidia GeForce RTX 2080 TI GPU with 11 GB of GPU memory and 64 GB of RAM memory.

We evaluated the proposed models in the signer-dependent and signer-independent modes. In signer-dependent mode, the models are trained and tested on samples from the same signer(s). In contrast, the signer-independent mode involves testing the model on a signer that has not been seen during the model training.

We used BLEU and WER metrics to evaluate the proposed models. Bilingual Evaluation Understudy (BLEU) performs

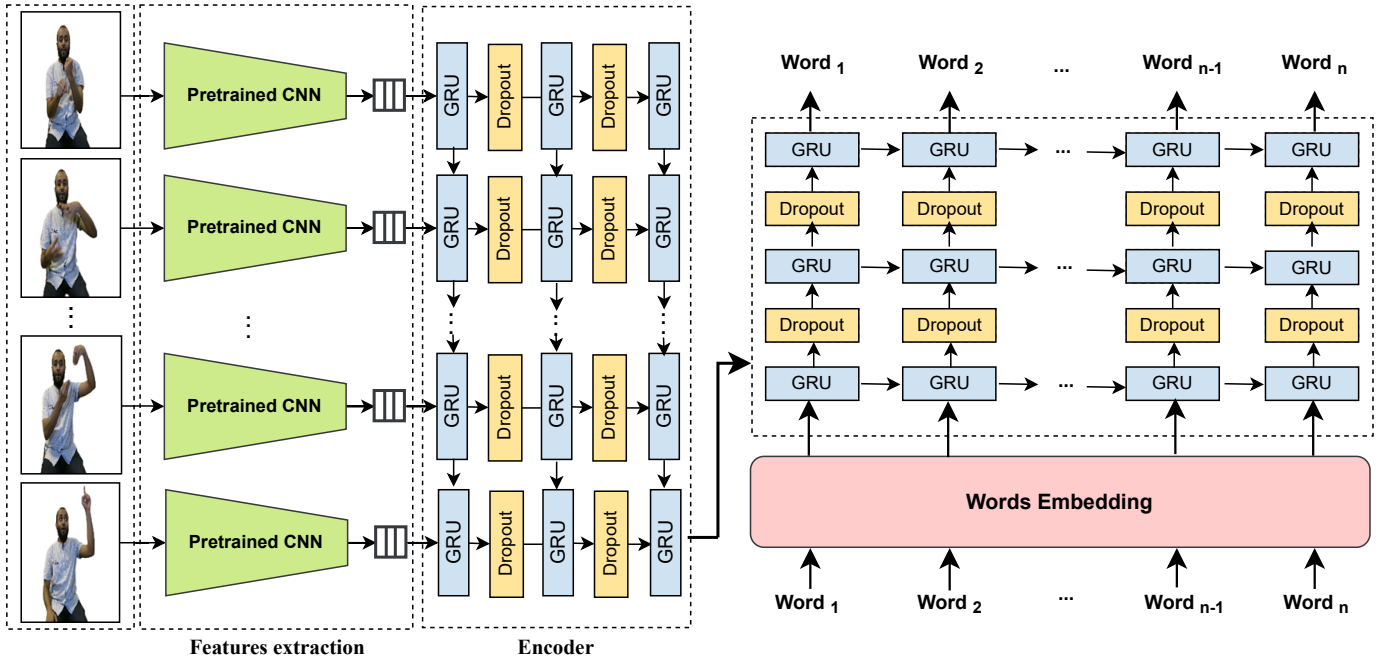


Fig. 6: The framework of the encoder-decoder model.

TABLE III: The performance of the proposed models in the signer-dependent mode.

	Encoder-decoder-Inception		Encoder-decoder-MobileNet		Attention-Inception		Attention-MobileNet	
	BLEU-4	WER	BLEU-4	WER	BLEU-4	WER	BLEU-4	WER
Signer 1	0.33	0.00	0.32	0.01	0.33	0.01	0.32	0.02
Signer 2	0.32	0.05	0.32	0.02	0.32	0.01	0.32	0.00
Signer 3	0.33	0.00	0.33	0.00	0.33	0.00	0.33	0.00
Signer 4	0.34	0.02	0.34	0.00	0.34	0.01	0.34	0.01
Signer 5	0.32	0.01	0.32	0.01	0.32	0.00	0.32	0.00
Signer 6	0.32	0.06	0.32	0.02	0.32	0.01	0.31	0.11
All	0.32	0.11	0.30	0.14	0.31	0.06	0.33	0.04
Average	0.33	0.04	0.32	0.03	0.32	0.01	0.32	0.03

TABLE IV: The performance of the proposed models in the signer-independent mode.

	Encoder-decoder-Inception		Encoder-decoder-MobileNet		Attention-Inception		Attention-MobileNet	
	BLEU-4	WER	BLEU-4	WER	BLEU-4	WER	BLEU-4	WER
Signer 1	0.20	0.44	0.27	0.27	0.18	0.54	0.24	0.34
Signer 2	0.24	0.46	0.19	0.51	0.21	0.61	0.16	0.65
Signer 3	0.19	0.50	0.19	0.6	0.15	0.64	0.16	0.76
Signer 4	0.13	0.67	0.14	0.58	0.09	0.76	0.14	0.52
Signer 5	0.25	0.36	0.22	0.37	0.18	0.53	0.12	0.67
Signer 6	0.12	0.61	0.06	0.65	0.09	0.66	0.04	0.76
Average	0.19	0.51	0.18	0.50	0.15	0.62	0.14	0.62

n-gram matching between the sentences resulting from the CSLR models and the reference sentences [33]. We used the BLEU metric with a 4-gram and a brevity penalty in all our experiments. The WER is derived from the Levenshtein distance and it works at the word level by comparing the CSLR output and the reference sentence word by word. This metric finds the differences between these sentences by computing the

number of insertions, deletions, and substitutions normalized by the total number of words in the sentence. Both BLEU and WER metrics score ranges between 0 and 1, where 1 indicates an exact match between the CSLR output and the reference sentences.

Table III shows the obtained BLEU-4 and WER results in the signer-dependent mode. We trained and tested the models

on samples of one signer. We refer to these settings in the table by Signer 1, Signer 2..etc. We also combined all signers' samples and split them into training and testing, and we refer to this setting as 'All'. As shown in the table, the proposed models performed well with all signers' data in the signer-dependent mode. The lowest WER was obtained using an attention model with an Inception pre-trained network. In addition, the BLEU score was almost similar across all models. It is also noticeable that all models were able to recognize the sentences of signer 3 with a WER of 0, whereas signer 6 was the most challenging signer for all models. This can be attributed to the small variations between the samples of signer 3 in contrast to signer 6, who is not an expert in sign language. This resulted in some variations between the samples of the same sign. These variations affected the capabilities of the models in sentence learning and recognition.

CSLR in the signer-independent mode is more challenging than recognition in the signer-dependent mode, as can be seen from the reported results in Table IV. As shown in the table, the obtained results using the encoder-decoder models outperformed the attention models with all signers. In addition, the lowest WER is obtained using the encoder-decoder model trained with the MobileNet pre-trained model. It is also noticeable that the highest WER is obtained with signer 6. These results align with the obtained results with signer 6 in the signer-dependent mode that reveal the variations between the signer's sentences.

V. CONCLUSIONS

CSLR is a hot computer vision problem with several approaches that have been proposed in the literature for CSLR. Most of these techniques depend on signs segmentation, whereas a few techniques recognize the sentence without the need for sign segmentation. To our knowledge, no technique has been proposed for continuous ArSL recognition. This can be attributed to the lack of an ArSL dataset at the sentence level. In this work, we proposed a continuous ArSL dataset. The dataset is composed of 9,335 samples, performed by 6 signers. The dataset has been acquired using Kinect V2 and all the samples are available in three forms: color, depth, and joint points. The dataset will be made publicly available to researchers.

Moreover, we have proposed encoder-decoder and attentions models for CSLR. The spatial features have been extracted from the sentence frames using two pre-trained models that are fed into the proposed models. We evaluated the models on the proposed dataset in the signer-dependent and independent modes. The obtained results show that the encoder-decoder model with features extracted using the MobileNet pre-trained network outperformed other models in the signer-independent mode.

ACKNOWLEDGMENT

The author would like to acknowledge the support received from Saudi Data and AI Authority (SDAIA) and King Fahd University of Petroleum and Minerals (KFUPM)

under SDAIA-KFUPM Joint Research Center for Artificial Intelligence Grant JRC-AI-RFP-05.

REFERENCES

- [1] Hearing loss statistics. URL: <https://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss>. Last visit: August. 13, 2022.
- [2] Sign language tutorial produced by al-jazeera media network. URL: shorturl.at/iN569. Last visit: August. 16, 2022.
- [3] Sentiment analysis using deep learning architectures: a review. *Artificial Intelligence Review*, 53(6):4335–4385, 2020.
- [4] N. Aloysius and M. Geetha. Understanding vision-based continuous sign language recognition. *Multimedia Tools and Applications*, 79:22177–22209, 2020.
- [5] Arab League Educational Cultural and Scientific Organization. *First part of the Unified Arabic Sign language Dictionary*. League Arab States Arab League Educ. Cult. Sci. Organ, 2000.
- [6] Arab League Educational Cultural and Scientific Organization. *Second part of the Unified Arabic Sign language Dictionary*. The League of Arab States & the Supreme Council for Family Affairs, Qatar, 2006.
- [7] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.
- [8] H. Brashear, T. Starner, P. Lukowicz, and H. Junker. Using multiple sensors for mobile sign language recognition. Georgia Institute of Technology, 2003.
- [9] N. C. Camgoz, S. Hadfield, O. Koller, and R. Bowden. Subnets: End-to-end hand shape and continuous sign language recognition. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, pages 3075–3084, 2017.
- [10] R. Cui, H. Liu, and C. Zhang. A deep neural framework for continuous sign language recognition by iterative training. *IEEE Transactions on Multimedia*, 21(7):1880–1891, 2019.
- [11] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2019.
- [12] P. Dreuw, C. Neidle, V. Athitsos, S. Sclaroff, and H. Ney. Benchmark databases for video-based automatic sign language recognition. In *LREC*, 2008.
- [13] P. Dreuw, D. Rybach, T. Deselaers, M. Zahedi, and H. Ney. Speech recognition techniques for a sign language recognition system. In *Eighth Annual Conference of the International Speech Communication Association*, 2007.
- [14] A. Duarte, S. Palaskar, L. Ventura, D. Ghadiyaram, K. DeHaan, F. Metzger, J. Torres, and X. Giro-i Nieto. How2sign: A large-scale multimodal dataset for continuous american sign language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2735–2744, June 2021.
- [15] E.-S. M. El-Alfy and H. Luqman. A comprehensive survey and taxonomy of sign language research. *Engineering Applications of Artificial Intelligence*, 114:105198, 2022.
- [16] I. Farag and H. Brock. Learning motion disfluencies for automatic sign language segmentation. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7360–7364. IEEE, 2019.
- [17] J. Forster, C. Schmidt, T. Hoyoux, O. Koller, U. Zelle, J. H. Piater, and H. Ney. Rwth-phoenix-weather: A large vocabulary sign language recognition and translation corpus. In *LREC*, volume 9, pages 3785–3789, 2012.
- [18] J. Forster, C. Schmidt, O. Koller, M. Bellgardt, and H. Ney. Extensions of the sign language recognition and translation corpus rwth-phoenix-weather. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1911–1916, 2014.
- [19] W. Gao, G. Fang, D. Zhao, and Y. Chen. Transition movement models for large vocabulary continuous sign language recognition. In *Proc. Sixth IEEE International Conference on Automatic Face and Gesture Recognition, 2004. Proceedings.*, pages 553–558, 2004.
- [20] M. Hassan, K. Assaleh, and T. Shanableh. Multiple proposals for continuous arabic sign language recognition. *Sensing and Imaging*, 20(1):1–23, 2019.
- [21] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

- [22] J. Huang, W. Zhou, Q. Zhang, H. Li, and W. Li. Video-based sign language recognition without temporal segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.
- [23] H. R. V. Joze and O. Koller. Ms-asl: A large-scale data set and benchmark for understanding american sign language. *arXiv preprint arXiv:1812.01053*, 2018.
- [24] I. Kagirow, D. Ivanko, D. Ryumin, A. Axyonov, and A. Karpov. Thruslan: Database of russian sign language. In *Proc. 12th Language Resources and Evaluation Conference*, pages 6079–6085, 2020.
- [25] O. Koller. Quantitative survey of the state of the art in sign language recognition. *arXiv preprint arXiv:2008.09918*, 2020.
- [26] W. Kong and S. Ranganath. Towards subject independent continuous sign language recognition: A segment and merge approach. *Pattern Recognition*, 47(3):1294–1308, 2014.
- [27] H. Luqman and E.-S. M. El-Alfy. Towards hybrid multimodal manual and non-manual arabic sign language recognition: Marsl database and pilot study. *Electronics*, 10(14):1739, 2021.
- [28] H. Luqman and S. A. Mahmoud. Automatic translation of arabic text-to-arabic sign language. *Universal Access in the Information Society*, 18(4):939–951, 2019.
- [29] R. M. McGuire, J. Hernandez-Rebollar, T. Starner, V. Henderson, H. Brashear, and D. S. Ross. Towards a one-way american sign language translator. In *Sixth IEEE International Conference on Automatic Face and Gesture Recognition, 2004. Proceedings.*, pages 620–625. IEEE, 2004.
- [30] M. Mohandes and M. Deriche. Arabic sign language recognition by decisions fusion using Dempster-Shafer theory of evidence. *2013 Computing, Communications and IT Applications Conference (Com-ComAp)*, pages 90–94, 2013.
- [31] H. Nagendraswamy and B. C. Kumara. Lbpv for recognition of sign language at sentence level: An approach based on symbolic representation. *Journal of Intelligent Systems*, 26(2):371–385, 2017.
- [32] I. Papastatis, 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. *IEEE Access*, 8:91170–91180, 2020.
- [33] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.
- [34] J. Pu, W. Zhou, and H. Li. Iterative alignment network for continuous sign language recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4165–4174, 2019.
- [35] S. M. Shohieb, H. K. Elminir, and A. Riad. Signsworld atlas; a benchmark arabic sign language database. *Journal of King Saud University-Computer and Information Sciences*, 27(1):68–76, 2015.
- [36] A. A. I. Sidig, H. Luqman, and S. A. Mahmoud. Arabic sign language recognition using vision and hand tracking features with hmm. *International Journal of Intelligent Systems Technologies and Applications*, 18(5):430–447, 2019.
- [37] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.
- [38] N. C. Tamer and M. Saraçlar. Keyword search for sign language. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8184–8188. IEEE, 2020.
- [39] U. Von Agris, M. Knorr, and K.-F. Kraiss. The significance of facial features for automatic sign language recognition. In *Proc. 8th IEEE International Conference on Automatic Face & Gesture Recognition*, pages 1–6, 2008.
- [40] A. Wadhawan and P. Kumar. Sign language recognition systems: A decade systematic literature review. *Archives of Computational Methods in Engineering*, 28(3):785–813, 2021.
- [41] L. Wang, Y. Xu, J. Cheng, H. Xia, J. Yin, and J. Wu. Human action recognition by learning spatio-temporal features with deep neural networks. *IEEE access*, 6:17913–17922, 2018.
- [42] Y. Ye, Y. Tian, M. Huenerfauth, and J. Liu. Recognizing american sign language gestures from within continuous videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 2064–2073, 2018.
- [43] J. Zhang, W. Zhou, and H. Li. A threshold-based hmm-dtw approach for continuous sign language recognition. In *Proceedings of International Conference on Internet Multimedia Computing and Service*, pages 237–240, 2014.
- [44] J. Zhang, W. Zhou, and H. Li. A new system for chinese sign language recognition. In *2015 IEEE China summit and international conference on signal and information processing (ChinaSIP)*, pages 534–538. IEEE, 2015.
- [45] H. Zhou, W. Zhou, and H. Li. Dynamic pseudo label decoding for continuous sign language recognition. In *Proc. IEEE International Conference on Multimedia and Expo (ICME)*, pages 1282–1287, 2019.
- [46] H. Zhou, W. Zhou, Y. Zhou, and H. Li. Spatial-temporal multi-cue network for continuous sign language recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13009–13016, 2020.