electronics

MDPI

Article

Towards Hybrid Multimodal Manual and Non-Manual Arabic Sign Language Recognition: mArSL Database and Pilot Study

Hamzah Luqman [1,2,*,†] and El-Sayed M. El-Alfy [1,2,†]

[1] Information and Computer Science Department, King Fahd University of Petroleum and Minerals, Dhahran 31261, Saudi Arabia; alfy@kfupm.edu.sa

[2] The Interdisciplinary Research Center for Intelligent Secure Systems (IRC-ISS), King Fahd University of Petroleum and Minerals, Dhahran 31261, Saudi Arabia

* Correspondence: hluqman@kfupm.edu.sa; Tel.: +966-13-860-1349

† These authors contributed equally to this work.

Abstract: Sign languages are the main visual communication medium between hard-hearing people and their societies. Similar to spoken languages, they are not universal and vary from region to region, but they are relatively under-resourced. Arabic sign language (ArSL) is one of these languages that has attracted increasing attention in the research community. However, most of the existing and available works on sign language recognition systems focus on manual gestures, ignoring other non-manual information needed for other language signals such as facial expressions. One of the main challenges of not considering these modalities is the lack of suitable datasets. In this paper, we propose a new multi-modality ArSL dataset that integrates various types of modalities. It consists of 6748 video samples of fifty signs performed by four signers and collected using Kinect V2 sensors. This dataset will be freely available for researchers to develop and benchmark their techniques for further advancement of the field. In addition, we evaluated the fusion of spatial and temporal features of different modalities, manual and non-manual, for sign language recognition using the state-of-the-art deep learning techniques. This fusion boosted the accuracy of the recognition system at the signer-independent mode by 3.6% compared with manual gestures.

Keywords: sign language database; sign language translation; gesture recognition; non-manual gestures; sign language recognition; facial expressions

check for updates

Citation: Luqman, H.; El-Alfy, E.-S.M. Towards Hybrid Multimodal Manual and Non-Manual Arabic Sign Language Recognition: mArSL Database and Pilot Study. *Electronics* **2021**, *10*, 1739. https://doi.org/10.3390/electronics10141739

Academic Editor: Kenji Suzuki

Received: 30 May 2021
Accepted: 13 July 2021
Published: 20 July 2021

1. Introduction

According to the World Health Organization (WHO), the deaf and hard-hearing community forms around 6.1% of world's population in 2018, which is close to 470 million people worldwide and is expected to be over 900 by 2050 [1]. Hearing impairment can be classified in several categories ranging from mild to profound. This community depends mainly on sign language to communicate with their society. This language is a complete natural language that has its own vocabulary and linguistic properties. It is not universal and there are many sign languages worldwide. There is no clear connection between spoken language and sign language and even countries that speak one language can have different sign languages such as American Sign Language (ASL) and British Sign Language (BSL) [2]. Moreover, some countries may have several sign languages in the same way as having several dialects of spoken language [3].

Unfamiliarity with sign language adds a barrier between deaf people and society. With the advances in computer vision and machine learning, different digital-aid systems have been developed to automatically recognize, synthesize and translate sign languages. Existing work on sign language can be classified in a variety of ways, e.g., based on the part of the body or type of features considered. Sign gesture is the basic component of sign language that can be classified, based on motion involvement, as static and dynamic. A static sign does not involve motion and largely depends on the shape and rotation of the signer's hands and fingers during signing [4]. Fingerspelling of alphabet letters and digits in most sign languages is expressed mostly using static signs. In contrast, dynamic signs involve motion of hands and other parts of the body during signing [5]. The majority of sign gestures can be categorized as dynamic where the motion plays a crucial role to convey the meaning. Sign gestures can be manual or non-manual or a combination of both. Gestures that involve hands and body motion can be considered as manual gestures. Non-manual gestures depend on other parts of the body such as facial expressions and head movement to express thoughts and clarify or emphasize meaning [6]. These gestures, manual and non-manual, are simultaneously utilized for the majority of signs.

Facial expressions are the dominant component of non-manual gestures in sign languages. They depend on mouth, eyes, eyebrows, lips, noses and cheeks to express feelings and emotions that can not be conveyed by manual gestures. In addition, facial expressions play an important role in expressing the linguistic proprieties of sign languages. They are used for grammatical structure, lexical distinction, and discourse functions such as negation and adverbial and adjectival contents [7]. An example of two signs of German Sign Language (GSL), "BROTHER" and "SISTER", that use the same hand gestures can be found in [8]. The difference between these signs depends on the facial expressions through lip pattern. Lip pattern is a commonly used parameter of non-manual gestures. Few lip patterns are dedicated to sign languages whilst the majority of lip patterns correspond to the pronunciation of the signed words in the spoken language. Deaf people are good lip readers and they read the lip patterns to gain a full understanding of the signs, especially from people who can hear.

Eyebrows and forehead are other examples primary components of the facial expressions in sign languages. They can be used alone or in combination with other facial expression components such as lip patterns. Figure 1a shows how eyebrows and forehead are employed with the face posture of "UGLY" sign of Arabic sign language. This face posture is also used with "BEAUTIFUL" sign, but with different facial expressions. Head motion is also an important component of the non-manual articulators of sign languages. Head posture can be used as an independent sign or integrated with manual gesture such as the "SLEEP" sign in Arabic sign language that consists of hand gestures and head motion, as shown in Figure 1b.

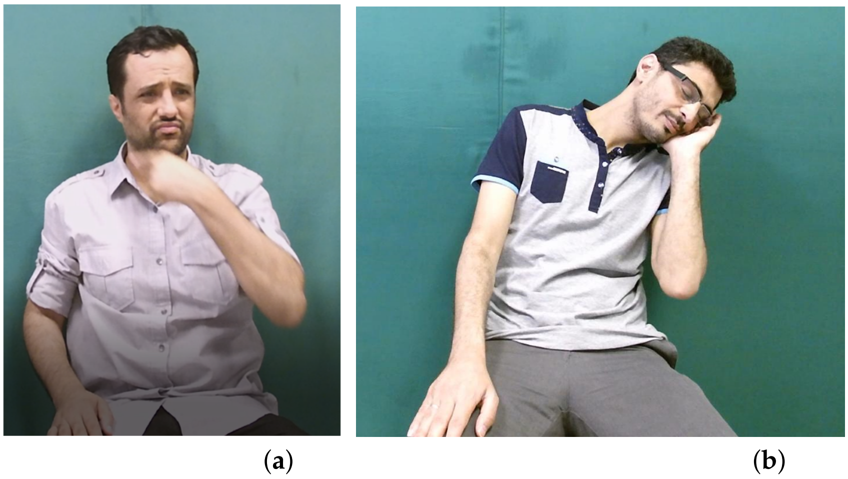

(**a**) (**b**)

Figure 1. Two signs showing the employment of facial expressions with manual gestures: (**a**) "UGLY"; (**b**) "SLEEP".

Arabic Sign Language (ArSL) is one of the languages that is used in Arab countries. This language is the unified language of several sign languages that exist in Arabic countries [9]. It was proposed in 1999 by the League of Arab States (LAS) and the Arab League Educational, Cultural and Scientific Organization (ALECSO) and a dictionary consisting of 3200 sign words was published in two parts in 2000 and 2006 [10,11]. This language is currently used mainly in the Arab gulf countries and is the main language used in the media channels such as Al-Jazeera. Research on automatic recognition of ArSL is still in its infancy and one of the main challenges associated with ArSL recognition systems is the lack of databases with sufficient numbers of relevant videos representing different articulators of the sign language [9]. In this work, we propose a multi-modality ArSL database with a focus on signs that employ manual and non-manual gestures. The proposed database consists of 6748 videos of 50 signs of ArSL performed by four signers. The signs of this database were recorded using Microsoft Kinect V2 sensors. In addition, we propose a hybrid model and quantitatively assess its effectiveness as a baseline for benchmarking future contributions on the dataset. The proposed model combines manual and non-manual gestures to enhance the recognition rates of a sign language recognition system. The prominent component of non-manual gestures, facial expressions, is integrated with the manual gestures and a high accuracy is obtained compared with only manual gestures. These experiments are conducted on different input representations of the sign gestures, such as RGB and depth data.

The remainder of this paper is organized as follows: Section 2 briefly reviews most of the related work dealing with non-manual features. Section 3 describes the motivations and details of the constructed database. Section 4 presents a pilot study using several state-of-the-art deep learning techniques for sign language recognition using examples of both manual and non-manual features. Finally, Section 5 concludes the paper and highlights the contributions of the paper.

2. Literature Review

Several techniques have been proposed in the last two decades for automatic recognition of sign language. The majority of these techniques targeted the dominant features of manual gestures. However, few approaches studied non-manual features (e.g., facial expressions) either alone or in combination with manual features. One of the main challenges of recognition systems is the lack of datasets, especially for ArSL. This section reviews available datasets of sign languages and surveys the most relevant recognition techniques of sign languages.

2.1. Sign Language Databases

The availability of databases is one of the challenges for advancing the research and development of sign language recognition and translation [12]. Although there is a large number of sign languages videos available online, these videos are not annotated, which makes them not useful for recognition and translation systems.

Sign language databases can be categorized into three main categories: fingerspelling, isolated signs, and continuous. Fingerspelling databases contain signs that depend mainly on finger shape and orientation. Most of the digits and alphabet letters of sign languages are static and use only fingers. Isolated sign words are equivalent to spoken words and they can be static or dynamic. Continuous sign language databases contain more than one sign word performed continuously. This section will focus on isolated signs databases since other databases are out of the scope of this paper.

Isolated sign language databases can be classified based on the acquisition device into sensor-based or vision-based databases. Sensor-based databases are collected using cumbersome sensors which can be worn on the signer's hand or wrist. The most commonly sensors used sensors for this purpose are electronic gloves. The need to wear these sensors during signing was one of the main issues with sensor-based recognition techniques and motivated researchers to use vision-based techniques. Vision-based databases are collected using single- or multi-camera acquisition devices. Single-camera devices provide a single piece of information about the signer, such as color video stream. A multi-camera device consists of more than one camera that provides different information about the signer, such as color and depth data. Multi-modal Kinect is one example of these devices that provide several types of information such as color, depth, and joint point information.

Several aspects are important to consider for evaluation of sign language databases, such as variability, size, and sign representation. The number of signers in the database is one of the factors that controls the database variability. This factor is important for evaluating the generalization of the recognition systems. Increasing the number of signers serves signer-independent recognition systems that are evaluated on signers different than signers involved in the system training. The number of samples per sign is another factor for sign language database evaluation. Having several samples per sign with some variations per sample is an important for training machine learning-based techniques that require large numbers of samples per sign. Sign representation data are also an important factor for evaluating databases. All the samples of vision-based sign language databases are available in RGB format. However, some databases [13–16] are recorded using multi-modality devices that provides other representations for the sign sample such as depth and joint points. Table 1 lists the surveyed sign language databases of non-Arabic sign languages at the sign word level. As shown in the table, the majority of the databases are for ASL. It is also noticeable that databases published before 2012 were only available in RGB format since multi-modality acquisition devices were released in 2011. In addition, datasets with large numbers of signs [16–18] do not have large numbers of samples per sign relative to their signs number compared with databases with low numbers of signs [13,19].

Table 1. Publicly available sign language datasets in other languages.

Year	Dataset	Language	Type	#Signs	#Signers	#Samples/Videos
2002	Purdue ASL [20]	American	RGB	184	14	3576
2005	RWTH-BOSTON50 [21]	American	RGB	50	3	483
2010	IIITA -ROBITA [22]	Indian	RGB	22	-	605
2012	ASLLVD [17]	American	RGB	3300+	6	9800
2012	MSR Gesture 3D [23]	American	RGB	12	10	336
2012	DGS Kinect 40 [13]	German	Multi *	40	15	3000
2012	GSL 982 [13]	Greek	Multi	982	1	4910
2013	PSL Kinect 30 [14]	Polish	Multi	30	1	300
2015	PSL ToF 84 [15]	Polish	Multi	84	1	1680
2015	DEVISIGN-L [16]	Chinese	Multi	2000	8	24,000
2016	LSA64 [24]	Argentine	RGB	64	10	3200
2016	Indian Kinect [25]	Indian	Multi	140	18	5041
2016	BosphorusSign Kinect [26]	Turkish	Multi	855	10	-
2019	BVCSL3D [19]	Indian	Multi	200	10	20,000
2020	BosphorusSign22k [27]	Turkish	Multi	744	6	22,542
2020	ASL-100-RGBD DATASET [28]	American	Multi	100	22	4150
2020	KSL DATASET [29]	Korean	RGB	77	20	1229
2020	WLASL2000 [18]	American	RGB	2000	119	21,083

* Multi: Multi-modality.

2.2. Sign Language Recognition Systems

The correlation between manual and non-manual gestures of sign language has been studied by Krnoul et al. [30]. This study was conducted on Czech sign language and the findings of this study showed that hand and head gestures are correlated mainly in signs with vertical movement of the head and hands. Caridakis et al. [31] discussed the grammatical and syntactic relation of manual and non-manual gestures to sign language. They also investigated the efficiency of including facial expressions in sign language recognition. Sabyrov et al. [32] used Logistic Regression for Kazakh-Russian sign language recognition. OpenPose was used to extract key points from manual gestures and facial expressions. The reported results show that combining manual key points with mouth key points improved the accuracy by 7%, whereas eyebrow key points improved the accuracy by only 0.5%. This conclusion was also reported by Elons et al. [33], who found that combining facial features with manual gestures improved the accuracy from 88% to 98%.

Paulraj et al. [34] extracted the area and discrete cosine transform (DCT) coefficients from the signer hand and head separately. Theses features were combined and classified using a simple neural network model to obtain an accuracy of 92.1% with 32 signs of Malaysian Sign Language. However, this technique depends on wearing colored gloves to facilitate hand segmentation, which makes this approach difficult to deploy in real time. DCT was also used by Rao and Kishore [35] for Indian sign language recognition. The 2D-DCT was used to extract features from the signer's head and hands, which were detected using a Sobel edge detector. This approach was evaluated on a dataset consisting of 18 signs and an accuracy of 90.58% was reported. DCT with HMM was used by Al-Rausan et al. [36] for ArSL recognition. A dataset consisting of 30 signs performed by 18 signers was used to evaluate this approach and accuracies of 96.74% and 94.2% were reported with signer-dependent and -independent modes, respectively. HMM was also used by Kelly et al. [37] to classify a set of statistical features extracted from the signer's hands and head.

An active appearance was used by Agris et al. [8] to detect the signer's mouth and eyes and a numerical description was computed from those components. For the signer's hands, a set of geometric features were extracted and concatenated with facial expressions features. This fusion of features improved the accuracy of GSL recognition by around 1.5%. Sarkar et al. [38] reported an improvement of around 4.0% using 39 signs of ASL through combining manual and non-manual gestures. A support vector machine (SVM) was used by Quesada et al. [39] for classifying manual and non-manual markers captured using a natural user interface device. This device captures hand shapes, body position, and facial expressions using 3D cameras. This approach achieved an accuracy of 91.25% using five face gestures and four handshapes of ASL. Kumar et al. [40] used two sensors to capture signers' hands and facial expressions. They used leap motion controller for manual gesture acquisition, whereas Kinect was used for capturing facial expressions. Then, HMM was used to recognize each component separately to be combined later using a Bayesian classification method. This combination boosted the recognition accuracy over a single modality by 1.04%. Camgoz et al. [41] employed a multi-channel transformer for continuous sign language recognition. Fusing a signer's hands and face improved the results to 19.21% compared with 16.76% using the signer's hands only. This approach was evaluated on the RWTH-PHOENIX-Weather-2014T [42] dataset, which consists of 1066 signs of GSL performed by nine signers.

3. mArSL Database

In this section, we present our proposed multi-modality database for Arabic sign language (ArSL). We will first explain the motivation for proposing this database and its properties as compared to other available ArSL. We will then describe the recording setup and sign capturing system and discuss the database components and organization.

3.1. Motivation

ArSL sign language is a low-resource language, yet there are 22 Arab countries with a total population of more than 430 million (https://worldpopulationreview.com/country-rankings/arab-countries (accessed on 30 May 2021)). Although the vocabulary of this language is limited compared with the spoken language, no database is available that accommodates all sign words. However, few datasets have been proposed recently with a main focus on the signs that depend only on manual gestures. These datasets ignored signs that combine manual and non-manual gestures either by excluding them explicitly or by not differentiating them from other signs. Subsequently, it becomes difficult to propose and evaluate techniques that incorporate other important non-manual articulators such as facial expressions, head and shoulder movements, and mouth shapes, which can provide extra information to enrich the meaning, provide clarity of similar signs with different meanings, represent grammatical markers, and show emotions and attitudes. This motivated us to propose a database for signs that combine manual and non-manual gestures to correctly recognize them. The proposed database involves two types of similarities: intra-class similarity and inter-class similarity. For intra-class similarity, each sign is performed under different conditions, including different signers. For inter-class similarity, each sign has various levels, such as "HUNGRY" and "VERY HUNGRY". This database will help also in studying the linguistic properties of ArSL using facial expressions as well as the relations and roles of manual and non-manual features. In addition, the proposed database will be freely available for researchers (https://faculty.kfupm.edu.sa/ICS/hluqman/mArSL.html (accesseed on 30 May 2021)).

3.2. Recording Setup

The mArSL database was recorded in an uncontrolled environment to resemble real-world scenarios. The database has been recorded in more than one session to ensure the variability of the signer's clothes and settings without any restrictions regarding clothes of specific colors. Figure 2a shows the recording room where signers are acting while sitting since the recorded signs require only the face and upper part of the signer's body. In addition, the distance between sensors and the signer was 1.5 m, which is enough to capture the signer's body and provide accurate skeleton information.

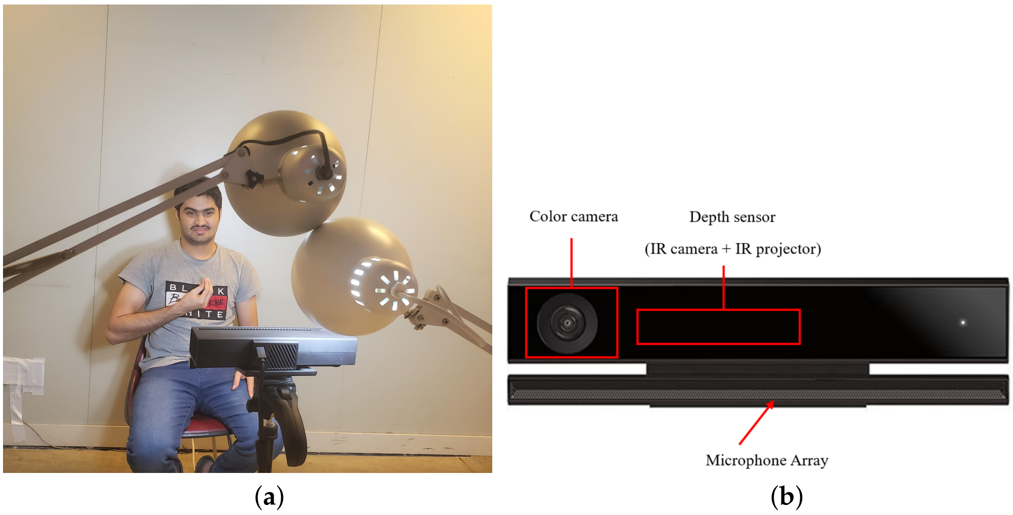


Figure 2. Acquisition environment and system: (**a**) Recording room, and (**b**) Kinect V2 device.

3.3. Sign Capturing Vision System

In order to obtain 3D multi-modal information of various signs of interest, we used Microsoft Kinect V2 as an acquisition device. The Kinect sensor is a motion sensing device that was initially designed by Microsoft for better user experience during video gaming and entertainment [43]. Two versions of Kinect were released by Microsoft in 2010 and 2015, respectively. Kinect V2 came with new features, such as the number of captured joint points, which was 25 in contrast to Kinect V1 that provides only 18 joint points.

The Kinect V2 sensor consists of color and infrared (IR) cameras. The color camera outputs a high resolution RGB video stream with a resolution of 1920 × 1080 pixels. The IR camera captures the modulated infrared light sent out by the IR projector/emitter to output depth images/maps that determine the distance between the sensor and each point of the scene based on the Time-of-Flight (ToF) and intensity modulation technique. The depth images are encoded using 16 bits and have a resolution of 512 × 424 pixels. In addition, the Kinect V2 sensor provides information about the signer's skeleton through 25 joint points and it is equipped with a microphone array to capture sounds, as shown Figure 2b.

The capturing software packaged with the Kinect device by Microsoft does not align with our requirements of having synchronized recordings of all data sources. To address this issue, we used the tool developed by Terven et al. [44] to capture several modalities synchronously. The recording system was developed using Matlab and provides five modalities for each sign gesture. These are color, depth, joint points, face, and face HD information. The color and depth information were saved in an MP4 format while other modalities were saved as a Matlab matrix. More information about each data modality will be discussed in the following subsections.

3.4. Database Statistics

The mArSL database consists of a set of signs that synchronously employ manual and non-manual gestures. The dominant non-manual component that appears with all database signs is the facial expressions, which rely on the movement pattern of eyes, eyebrows, lips, nose and cheeks to visually emphasize or express the person's mood or attitude. We intentionally selected these set of signs for words and phrases after studying a large number of signs in the sign language to focus on those requiring both manual and non-manual features.

The database signs can be divided into two parts based on the number of postures in the sign. The first part consists of signs with one posture, e.g., the sign for "Hungry" shown in Figure 3a. The second category of the database signs consists of more than one posture, such as the"VERY SMALL" sign shown in Figure 3b.

The proposed database consists of 6748 videos of 50 signs of ArSL. They were performed by four signers trained on ArSL. Each sign was repeated 30 times (except one signer who has more than 30 samples per sign) in different sessions. The duration of each sample differs from sign to sign and from signer to signer. The total number of frames of the entire database is 337,991 frames. Table 2 shows a comparison between mArSL and other available ArSL databases. As shown in the table, few databases are available for ArSL compared with other languages such as ASL. In addition, all other databases are not designed to include non-manual articulations since their focus was on the dominant manual articulations. The SignsWorld Atlas [45] is the only dataset that has some samples for facial expressions. This database is designed to include signs from different types of databases, such as fingerspelling and continuous ones. However, this database is not suitable for recognition systems since the number of samples per sign is not unified where some signs have only one sample while others have 10 samples. In addition, the non-manual gestures are represented using still images and they are not integrated with manual gestures; this can make them suitable for facial expression studies rather than sign language recognition.

Table 2. Comparisons of our mArSL database with existing ArSL databases.

Year	Dataset	Type	Manual	Non-Manual	#Signs	#Signers	#Samples
2007	Shanableh et al. [46]	RGB	✓		23	3	3450
2015	SignsWorld Atlas [45]	RGB	✓	✓	264+	10	535+
2020	KArSL [47]	Multi	✓		502	3	75,300
2021	mArSL	Multi	✓	✓	50	4	6748

3.5. Database Organization

Each captured sign is represented by five modalities: color, depth, joint points, face, and faceHD. An illustrative example is shown in Figure 4. A description of each of these modalities is provided below:

- Color data: Kinect V2 provides an RGB video stream of the captured gestures recorded at a frame rate of 30 fps (frames per second). This frame rate can keep the motion smooth and capture the gesture details efficiently. The color camera has a resolution of 1920 × 1080 pixels with a Field of View (FoV) of 84.1° × 53.8° . This RGB video stream of each sign sample is saved in an MP4 format.
- Depth map: Another datum that is provided by Kinect V2 is the depth map, which is captured simultaneously with the color data. A depth map describes, at each pixel, the distance to the signer from the front-facing camera. Each captured sign has a depth

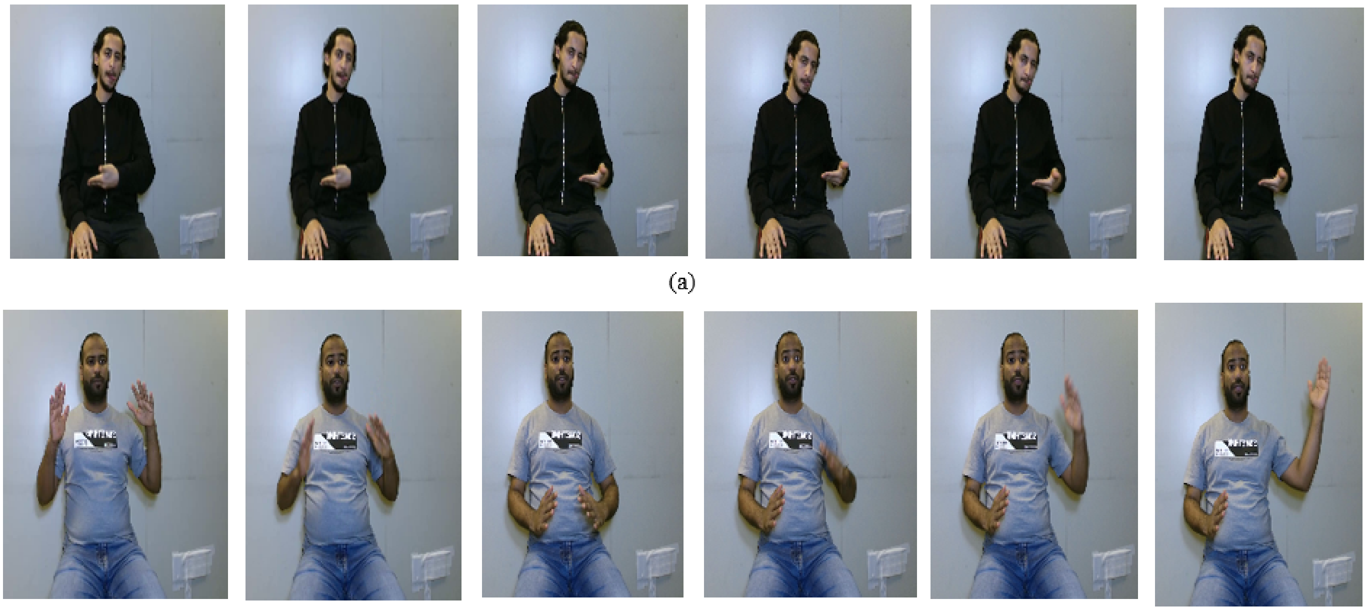
map which consists of a sequence of frames with a resolution of 512 × 424 pixels saved in an MP4 format.

- Skeleton joint points: Kinect captures the human skeleton using an integrated depth camera. This camera can detect and track the human skeleton and presents it using 25 joint points. The coordinates of each joint point are available in three spaces—color, depth, and camera. The color space describes the joint point coordinates (x and y) on the color image provided by the color camera. The depth space describes the 2D location of the joint point on the depth image. The coordinates of the joint point in the camera space are 3D (x, y, z) and are measured in meters. The coordinates (x, y) can be positive or negative, as they extend in both directions from the sensor while the z coordinate is always positive as it grows out from the sensor. In addition, the orientation information of each joint point is provided by Kinect as a quaternion which consists of four values (q_w, q_x, q_y, q_z) and is mathematically represented by a real part and 3D vector as follows: $Q = q_w + q_x i + q_y j + q_z k$, where i, j, and k are unit vectors in the direction of the x, y and z axes, respectively.
- Face information: One of the state-of-the-art capabilities of the Kinect device is face tracking. The Kinect sensor utilizes infrared and color cameras to track facial points. These points can be used in several applications related to facial expressions such as recognition and Avatar development. Two types of face information are provided for each tracked face: Face Basics and Face HD [44]. The former provides information about the signer face which includes:

  - Face Box: the coordinates of the face position in the color space (left, top, right, bottom);
  - Face Points: the center coordinates of the five face landmarks (left and right eyes, nose, right and left mouth boundaries);
  - Face Rotation: a vector containing pitch, yaw, and roll angles;
  - Face Properties: a 1 × 8 vector containing the detection result of certain face properties (0: unknown, 1: no, 2: maybe, 3: yes) of the following face properties: happy, engaged, wearing glasses, left eye closed, right eye closed, mouth open, mouth moved, and looking away.

  On the other hand, Face HD provides high definition face information with 94 facial points which are expressed in a camera space point. In addition to the Face Box and Rotation, each captured Face HD has the following information (https://docs.microsoft.com/en-us/previous-versions/windows/kinect/dn799271(v=ieb.10) (accesssed on 30 May 2021)):

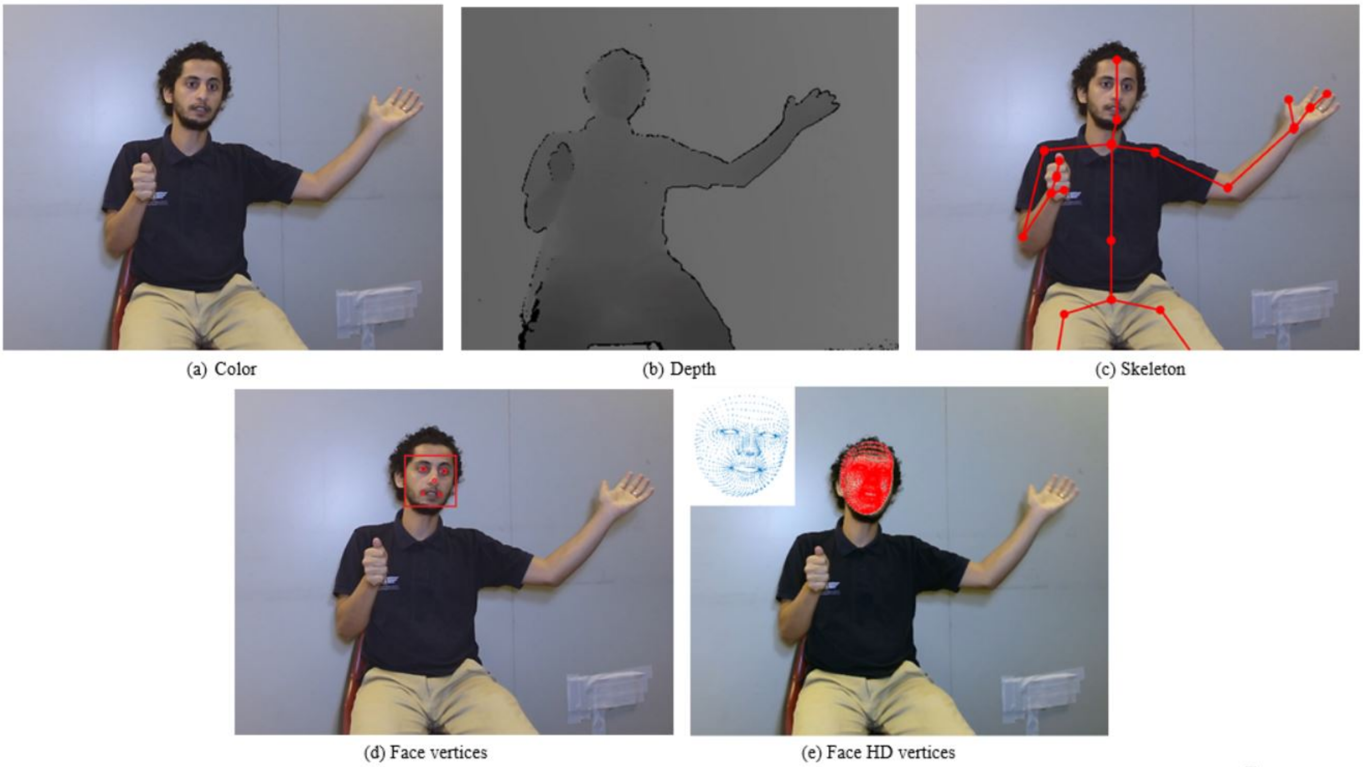
  - Head Pivot: the head center which the face may be rotated around. This point is defined in the Kinect body coordinate system;
  - Animation Units: a 1 × 17 vector containing the animation units. Most of these units are expressed as a numeric weight varying between 0 and 1;
  - Shape Units: a 1 × 94 vector containing the shape units which are expressed as a numeric weight that typically varies between −2 and +2;
  - Face Model: the coordinates of 1347 points of a 3D face model (Figure 4e).



(a)

(b)

Figure 3. Selected frames of two signs of mArSL database with different number of postures: (**a**) “HUNGRY” and (**b**) “VERY SMALL”.



(a) Color

(b) Depth

(c) Skeleton

(d) Face vertices

(e) Face HD vertices

Figure 4. An illustrative example for the five modalities provided for each sign sample.

4. Pilot Study and Benchmark Results

We evaluated manual and non-manual features for sign language recognition using the proposed dataset. We started by evaluating the manual gestures for automatic recognition of sign language. Then, we extended the experiments by fusing the manual gestures with facial expressions. Two evaluation settings were used to evaluate the proposed systems: signer-dependent and signer-independent settings. The signer-dependent mode evaluates the system on signer(s) already seen in the training data. In contrast, the signer-independent mode evaluates the system on signer(s) unseen in the training data. This type of evaluation is more challenging for machine learning algorithms compared with signer-dependent evaluations.

4.1. Manual Gestures

Sign language words are dynamic gestures where the motion is a primary part of the sign. Learning these gestures intuitively requires a sequential modeling technique such as Long short-term memory (LSTM) and Hidden Markov Model (HMM). These techniques are efficient for learning temporal data though they do not pay much attention to spatial information in the video stream. To address this issue, we used a convolutional neural network (CNN) to extract the spatial features from the gesture frames and feed them into a stacked LSTM.

We evaluated transfer learning using several state-of-the-art pre-trained models for spatial information extraction of images. We fine-tuned Inception-V3 [48], Xception [49], ResNet50 [50], VGG-16 [51], and MobileNet [52] models, which were pre-trained on ImageNet for large scale image classification with 14,197,122 images and 21,841 subcategories. In addition, we proposed a CNN model consisting of five layers with a number of kernels ranging from 16 to 64. The first two layers use a kernel of size 5 × 5, while other layers used a 3 × 3 kernel size. These layers were followed by a rectified linear (ReLU) activation function to remove the non-linearity of the input data to this function. We also used maximum pooling layers of size 2 × 2 to down-sample the feature maps. The extracted features were fed into a stacked LSTM consisting of two LSTM layers. Each layer consists of 1024 neurons followed by a dropout layer to reduce the overfitting. These layers were followed by a Softmax classifier. Adam optimizer was used with a learning rate of 0.0001. The framework of the proposed system is shown in Figure 5.

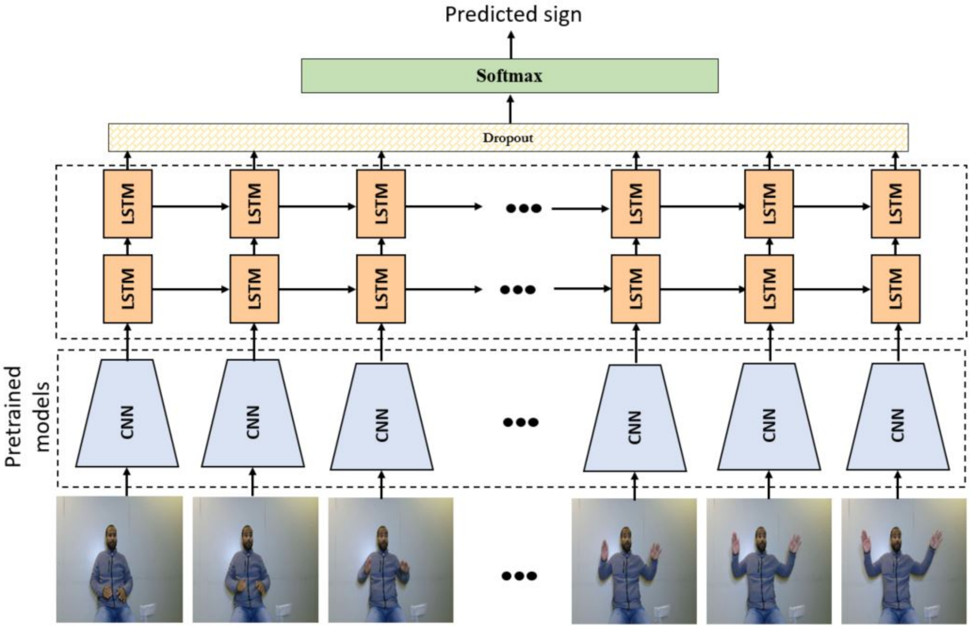


Figure 5. Framework of the proposed system where two variants of CNN are used in the pilot study: transfer learning with fine tuning and customized models.

Two data representations were used as inputs to the proposed systems: color and depth data. We fed 25 frames of each data input to the proposed models. These frames were selected by taking an interval between consecutive frames relative to the total number of sign frames. We replicated the last frame of sign samples that have less than 25 frames. Table 3 shows the obtained results using color and depth frames with signer-dependent and signer-independent modes. As shown in the table, the accuracy results of the signer-dependent mode using color data are higher than depth data with almost all the models. For signer-dependent and color images, the highest average accuracy of 99.7% was obtained with the MobileNet-LSTM model while the lowest accuracy of 94.9% was obtained using ResNet50-LSTM with both data representations. In contrast, the results dropped to 99.5% and 72.4% when using depth images with MobileNet-LSTM and ResNet50-LSTM, respectively. This can be caused by the signer information learnt with color data, leading to overfitting. This information is excluded with depth data, which explains the better performance of depth data representation.

Table 3. Recognition accuracies using color and depth data.

	Input Data	**Tested Signer**	**CNN-LSTM**	**Inception-LSTM**	**Xception-LSTM**	**ResNet50-LSTM**	**VGG-16-LSTM**	**MobileNet-LSTM**
Signer-Dependent	Color	1	0.969	1.0	1.0	0.993	1.0	1.0
		2	0.988	0.992	1.0	0.972	0.996	0.996
		3	0.993	1.0	0.989	0.935	0.982	1.0
		4	0.993	0.986	0.982	0.883	0.986	0.993
		All	0.989	0.995	0.994	0.964	0.991	0.996
		Average	0.986	0.995	0.993	0.949	0.991	0.997
	Depth	1	0.960	0.995	1.0	0.832	0.981	1.0
		2	0.952	0.984	0.992	0.712	0.964	0.988
		3	0.813	0.993	0.996	0.640	0.946	0.996
		4	0.979	0.986	0.993	0.714	0.926	0.993
		All	0.989	0.996	0.994	0.722	0.980	0.996
		Average	0.939	0.991	0.995	0.724	0.959	0.995
Signer-Independent	Color	1	0.060	0.289	0.287	0.074	0.330	0.566
		2	0.022	0.317	0.317	0.238	0.541	0.501
		3	0.137	0.369	0.295	0.165	0.271	0.361
		4	0.039	0.259	0.341	0.120	0.546	0.550
		Average	0.065	0.309	0.310	0.149	0.422	0.495
	Depth	1	0.042	0.382	0.331	0.158	0.305	0.541
		2	0.093	0.408	0.446	0.162	0.372	0.641
		3	0.137	0.325	0.357	0.132	0.245	0.476
		4	0.200	0.568	0.528	0.128	0.252	0.530
		Average	0.118	0.421	0.416	0.145	0.294	0.547

It is also noticeable in Table 3 that the signer-independent mode is more challenging than the signer-dependent one. The sharp decrease in the recognition accuracies with the signer-independent mode can be attributed to the models that started to overfit the signers during system learning. In addition, the variations between signs performed by signers have an effect on the recognition accuracy. To address this issue, we excluded the signer identity information and made the models focus only on the signer's motion through calculating the optical flow of the input data and fed them into the proposed models. Optical flow provides discriminative temporal information that helps in gesture recognition. We used the Dual TV-L1 algorithm, which is based on the total variation regularization and L1 norm, to compute the optical flow between two image frames [53,54]. Figure 6 shows a sequence of color frames with their optical flows. Table 4 shows the obtained recognition accuracies using the optical flow of color and depth data. As shown in the table, there is a significant improvement in the recognition accuracies of both data representations compared with raw color and depth data. It was also noticed that the optical flow of color data outperformed the optical flow of depth data. In addition, MobileNet-LSTM outperformed all other models with an average accuracy of 72.4% (compared to 54.7% without optical flow).

Table 4. Signer-independent recognition accuracies using the optical flow, computed either from color or from depth data.

Input Data	**Tested Signer**	**CNN-LSTM**	**Inception-SLSTM**	**Xception-LSTM**	**ResNet50-LSTM**	**VGG-16-LSTM**	**MobileNet-LSTM**
Color	1	0.558	0.608	0.636	0.025	0.387	0.490
	2	0.439	0.769	0.761	0.020	0.441	0.784
	3	0.449	0.757	0.795	0.020	0.336	0.863
	4	0.519	0.732	0.667	0.021	0.466	0.760
	Average	0.491	0.717	0.715	0.021	0.408	0.724
Depth	1	0.455	0.599	0.490	0.024	0.220	0.581
	2	0.474	0.641	0.609	0.020	0.174	0.638
	3	0.411	0.614	0.582	0.020	0.185	0.643
	4	0.519	0.678	0.696	0.014	0.221	0.726
	Average	0.465	0.633	0.594	0.020	0.200	0.647

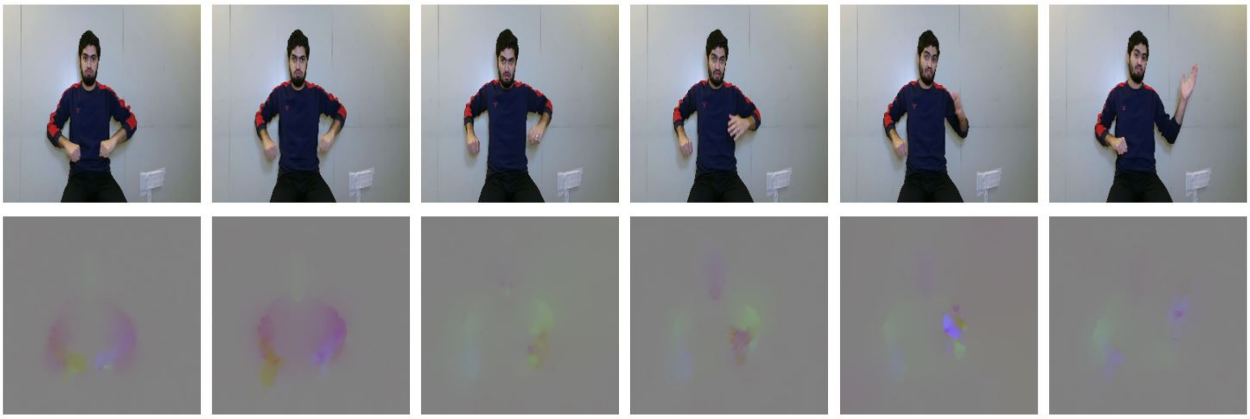


Figure 6. Selected frames of "VERY FAT" sign with their optical flow.

4.2. Non-Manual Gestures

In this subsection, we evaluated an important component of non-manual articulators, facial expressions. We used animation units (AUs) of the signer face provided by the Kinect sensor as an input to the proposed system (more information about face data can be found in Section 3.5).

We started by evaluating the facial expressions alone, and then we fused this information with the best model of the manual gesture recognition discussed in the previous section. We used a stacked LSTM model consisting of two LSTM layers with 1024 neurons to learn the temporal information of the facial landmarks. The extracted face features are fused at the classification level with the manual gesture features that were extracted in the previous section as shown in Figure 7.

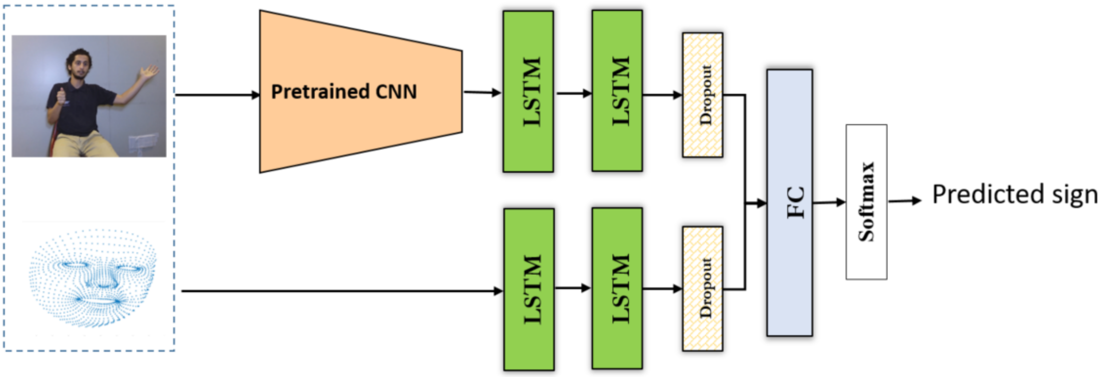


Figure 7. Manual and non-manual fused system.

Table 5 shows the obtained results with facial expressions using animation units (AU) in the signer-independent mode. We fused these features with the features that were extracted from the color and depth data of the manual gestures. The table compares the results before and after fusing manual and non-manual gestures. As shown in the table in the column labeled AU, using facial expressions alone for sign recognition does not give good results since similar facial expressions can be associated with multiple signs, especially when emphasizing something; hence, they are not necessarily linked to specific signs. In addition, a few signs in sign language depend on the non-manual gestures without manual gestures. Therefore, we concatenated these features with the best manual model reported in the previous section (namely, MobileNet-LSTM). As shown in the table, the highest recognition accuracy was obtained with optical flow of the color data and was improved by about 3% compared with the results without fusion. Based on these experiments, we can conclude that fusing manual gestures with facial expressions can improve the accuracy of the sign language recognition system. In addition, optical flow is efficient in calculating the motion between frames, which helps improve the accuracy of the signer-independent evaluation mode.

Table 5. Signer-independent recognition accuracies of manual gestures and the principal non-manual articulators using facial expressions (AU in the table refers to animation units representing facial expressions alone).

	Tested Signer	AU	Raw Images		Optical Flow	
			MobileNet-LSTM	MobileNet-LSTM+AU	MobileNet-LSTM	MobileNet-LSTM+AU
Color	1	0.174	0.566	0.603	0.490	0.589
	2	0.132	0.501	0.597	0.784	0.787
	3	0.142	0.361	0.280	0.863	0.858
	4	0.153	0.550	0.534	0.760	0.805
	Average	0.150	0.495	0.504	0.724	0.760
Depth	1	0.174	0.541	0.551	0.581	0.593
	2	0.132	0.641	0.620	0.638	0.622
	3	0.142	0.476	0.519	0.643	0.656
	4	0.153	0.530	0.599	0.726	0.664
	Average	0.150	0.547	0.572	0.647	0.634

5. Conclusions and Future Work

This paper introduced a new multi-modality video database for sign language recognition. Unlike existing databases, its focus is on signs that require both manual and non-manual articulators which can be used in a variety of studies related to sign language recognition. Though the signs are performed to match the guidelines of the Arabic sign language (ArSL), which is still in the developmental stage, the database can be beneficial for other researchers as well such as those working on pattern recognition and machine learning. Moreover, the paper presented a baseline pilot study to evaluate and compare six models based on the state-of-the-art deep learning techniques for spatial and temporal processing of sign videos in the database. Two cases are considered for the signer-dependent and signer-independent modes using manual and non-manual features. In the first case, we used color and depth images directly, whereas in the second case we used optical flow to extract more relevant features to the signs themselves not the signers. The best results are obtained when using MobileNet-LSTM with transfer learning and fine tuning with 99.7% and 72.4% for signer-dependent and signer-independent modes, respectively. As future work, more analysis on the effectiveness of each part of the non-manual gestures will be conducted. In addition, we are going to explore other deep learning approaches for isolated sign language recognition and investigate the generalization of the proposed techniques to other sign language datasets. Moreover, we will target continuous sign language recognition of ArSL as this problem has not been explored as deeply as isolated sign language recognition.

Author Contributions: Data curation, H.L.; Formal analysis, H.L. and E.-S.M.E.-A.; Investigation, H.L.; Methodology, H.L. and E.-S.M.E.-A.; Validation, E.-S.M.E.-A.; Writing—original draft, H.L. All authors have read and agreed to the published version of the manuscript.

Funding: The authors would like to acknowledge the support provided by King Fahd University of Petroleum and Minerals (KFUPM) for funding this work under project number SR191014.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Galindo, N.M.; Sá, G.G.d.M.; Pereira, J.d.C.N.; Barbosa, L.U.; Barros, L.M.; Caetano, J.Á. Information about COVID-19 for deaf people: An analysis of Youtube videos in Brazilian sign language. *Rev. Bras. Enferm.* **2021**, *74*. [CrossRef]
2. Makhashen, G.B.; Luqman, H.A.; El-Alfy, E.S. Using Gabor Filter Bank with Downsampling and SVM for Visual Sign Language Alphabet Recognition. In Proceedings of the 2nd Smart Cities Symposium (SCS 2019), Bahrain, Bahrain, 24–26 March 2019; pp. 1–6.
3. Sidig, A.A.I.; Luqman, H.; Mahmoud, S.A. Transform-based Arabic sign language recognition. *Procedia Comput. Sci.* **2017**, *117*, 2–9. [CrossRef]
4. Pisharady, P.K.; Saerbeck, M. Recent methods and databases in vision-based hand gesture recognition: A review. *Comput. Vis. Image Underst.* **2015**, *141*, 152–165. [CrossRef]
5. Luqman, H.; Mahmoud, S.A. Automatic translation of Arabic text-to-Arabic sign language. In *Universal Access in the Information Society*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 1–13.
6. Nair, A.V.; Bindu, V. A Review on Indian Sign Language Recognition. *Int. J. Comput. Appl.* **2013**, *73*, 33–38.
7. Gupta, P.; Agrawal, A.K.; Fatima, S. Sign Language Problem and Solutions for Deaf and Dumb People. In Proceedings of the 3rd International Conference on System Modeling & Advancement in Research Trends (SMART), Sicily, Italy, 30 May–4 June 2004.
8. Von Agris, U.; Knorr, M.; Kraiss, K.F. The significance of facial features for automatic sign language recognition. In Proceedings of the 8th IEEE International Conference on Automatic Face & Gesture Recognition, Amsterdam, The Netherlands, 17–19 September 2008; pp. 1–6.
9. Sidig, A.A.I.; Luqman, H.; Mahmoud, S.A. Arabic Sign Language Recognition Using Optical Flow-Based Features and HMM. In *International Conference of Reliable Information and Communication Technology*; Springer: Berlin/Heidelberg, Germany, 2017; pp. 297–305.
10. *LAS: Second Part of the Unified Arabic Sign Dictionary*; The League of Arab States & the Arab League Educational, Cultural and Scientific Organization: Tunis, Tunisia, 2006.
11. *LAS: First Part of the Unified Arabic Sign Dictionary*; The League of Arab States & the Arab League Educational, Cultural and Scientific Organization: Tunis, Tunisia, 2000.
12. Luqman, H.; Mahmoud, S.A. A machine translation system from Arabic sign language to Arabic. In *Universal Access in the Information Society*; Springer: Berlin/Heidelberg, Germany, 2019; pp. 1–14.
13. Ong, E.J.; Cooper, H.; Pugeault, N.; Bowden, R. Sign language recognition using sequential pattern trees. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 2200–2207.
14. Oszust, M.; Wysocki, M. Polish sign language words recognition with kinect. In Proceedings of the 6th IEEE International Conference on Human System Interactions (HSI), Sopot, Poland, 6–8 June 2013; pp. 219–226.
15. Kapuscinski, T.; Oszust, M.; Wysocki, M.; Warchol, D. Recognition of hand gestures observed by depth cameras. *Int. J. Adv. Robot. Syst.* **2015**, *12*, 36. [CrossRef]

16. Chai, X.; Wang, H.; Zhou, M.; Wu, G.; Li, H.; Chen, X. *DEVISIGN: Dataset and Evaluation for 3D Sign Language Recognition*; Technical Report; Beijing, China, 2015.
17. Neidle, C.; Thangali, A.; Sclaroff, S. Challenges in development of the american sign language lexicon video dataset (asllvd) corpus. In *5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon, LREC*. Available online: https://open.bu.edu/handle/2144/31899 (accessed on 30 May 2021).
18. Li, D.; Rodriguez, C.; Yu, X.; Li, H. Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. In Proceedings of the IEEE Winter Conference on Applications of Computer Vision, Snowmass Village, CO, USA, 1–5 March 2020; pp. 1459–1469.
19. Ravi, S.; Suman, M.; Kishore, P.; Kumar, K.; Kumar, A. Multi modal spatio temporal co-trained CNNs with single modal testing on RGB–D based sign language gesture recognition. *J. Comput. Lang.* **2019**, *52*, 88–102. [CrossRef]
20. Martínez, A.M.; Wilbur, R.B.; Shay, R.; Kak, A.C. Purdue RVL-SLLL ASL database for automatic recognition of American Sign Language. In Proceedings of the 4th IEEE International Conference on Multimodal Interfaces, Pittsburgh, PA, USA, 16 October 2002; pp. 167–172.
21. Zahedi, M.; Keysers, D.; Deselaers, T.; Ney, H. Combination of tangent distance and an image distortion model for appearance-based sign language recognition. In *Joint Pattern Recognition Symposium*; Springer: Berlin/Heidelberg, Germany, 2005; pp. 401–408.
22. Nandy, A.; Prasad, J.S.; Mondal, S.; Chakraborty, P.; Nandi, G.C. Recognition of isolated Indian sign language gesture in real time. In *International Conference on Business Administration and Information Processing*; Springer: Berlin/Heidelberg, Germany, 2010; pp. 102–107.
23. Kurakin, A.; Zhang, Z.; Liu, Z. A real time system for dynamic hand gesture recognition with a depth sensor. In Proceedings of the 2012 Proceedings of the 20th European Signal Processing Conference (EUSIPCO), Bucharest, Romania, 27–31 August 2012; pp. 1975–1979.
24. Ronchetti, F.; Quiroga, F.; Estrebou, C.; Lanzarini, L.; Rosete, A. LSA64: A Dataset of Argentinian Sign Language. In Proceedings of the XX II Congreso Argentino de Ciencias de la Computación (CACIC), San Luis, Argentina, 3–7 October 2016.
25. Ansari, Z.A.; Harit, G. Nearest neighbour classification of Indian sign language gestures using kinect camera. *Sadhana* **2016**, *41*, 161–182. [CrossRef]
26. Camgöz, N.C.; Kındıroğlu, A.A.; Karabüklü, S.; Kelepir, M.; Özsoy, A.S.; Akarun, L. BosphorusSign: a Turkish sign language recognition corpus in health and finance domains. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), Portoroz, Slovenia, 23–28 May 2016; pp. 1383–1388.
27. Özdemir, O.; Kındıroğlu, A.A.; Camgöz, N.C.; Akarun, L. BosphorusSign22k Sign Language Recognition Dataset. *arXiv* **2020**, arXiv:2004.01283.
28. Hassan, S.; Berke, L.; Vahdani, E.; Jing, L.; Tian, Y.; Huenerfauth, M. An Isolated-Signing RGBD Dataset of 100 American Sign Language Signs Produced by Fluent ASL Signers. In Proceedings of the LREC2020 9th Workshop on the Representation and Processing of Sign Languages: Sign Language Resources in the Service of the Language Community, Technological Challenges and Application Perspectives, Marseille, France 16 May 2020; pp. 89–94.
29. Yang, S.; Jung, S.; Kang, H.; Kim, C. The Korean Sign Language Dataset for Action Recognition. In *International Conference on Multimedia Modeling*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 532–542.
30. Krňoul, Z.; Hrúz, M.; Campr, P. Correlation analysis of facial features and sign gestures. In Proceedings of the IEEE 10th International Conference on Signal Processing, Beijing, China, 24–28 October 2010; pp. 732–735.
31. Caridakis, G.; Asteriadis, S.; Karpouzis, K. Non-manual cues in automatic sign language recognition. *Pers. Ubiquitous Comput.* **2014**, *18*, 37–46. [CrossRef]
32. Sabyrov, A.; Mukushev, M.; Kimmelman, V. Towards Real-time Sign Language Interpreting Robot: Evaluation of Non-manual Components on Recognition Accuracy. In Proceedings of the CVPR Workshops, Long Beeach, CA, USA, 16–20 June 2019.
33. Elons, A.S.; Ahmed, M.; Shedid, H. Facial expressions recognition for arabic sign language translation. In Proceedings of the 9th IEEE International Conference on Computer Engineering & Systems (ICCES), Cairo, Egypt, 22–23 December 2014; pp. 330–335.
34. Paulraj, M.; Yaacob, S.; Desa, H.; Hema, C.R.; Ridzuan, W.M.; Ab Majid, W. Extraction of head and hand gesture features for recognition of sign language. In Proceedings of the 2008 International Conference on Electronic Design, Penang, Malaysia, 1–3 December 2008; pp. 1–6.
35. Rao, G.A.; Kishore, P. Selfie video based continuous Indian sign language recognition system. *Ain Shams Eng. J.* **2018**, *9*, 1929–1939. [CrossRef]
36. Al-Rousan, M.; Assaleh, K.; Tala'a, A. Video-based signer-independent Arabic sign language recognition using hidden Markov models. *Appl. Soft Comput.* **2009**, *9*, 990–999. [CrossRef]
37. Kelly, D.; Reilly Delannoy, J.; Mc Donald, J.; Markham, C. A framework for continuous multimodal sign language recognition. In Proceedings of the 2009 international conference on Multimodal Interfaces, Wenzhou, China, 26–29 November 2009; pp. 351–358.
38. Sarkar, S.; Loeding, B.; Parashar, A.S. Fusion of manual and non-manual information in american sign language recognition. In *Handbook of Pattern Recognition and Computer Vision*; World Scientific: Singapore, 2010; pp. 477–495.
39. Quesada, L.; Marín, G.; Guerrero, L.A. Sign language recognition model combining non-manual markers and handshapes. In *International Conference on Ubiquitous Computing and Ambient Intelligence*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 400–405.

40. Kumar, P.; Roy, P.P.; Dogra, D.P. Independent bayesian classifier combination based sign language recognition using facial expression. *Inf. Sci.* **2018**, *428*, 30–48. [CrossRef]
41. Camgoz, N.C.; Koller, O.; Hadfield, S.; Bowden, R. Multi-channel transformers for multi-articulatory sign language translation. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 301–319.
42. Camgoz, N.C.; Hadfield, S.; Koller, O.; Ney, H.; Bowden, R. Neural sign language translation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7784–7793.
43. Lachat, E.; Macher, H.; Landes, T.; Grussenmeyer, P. Assessment and calibration of a RGB-D camera (Kinect v2 Sensor) towards a potential use for close-range 3D modeling. *Remote Sens.* **2015**, *7*, 13070–13097. [CrossRef]
44. Terven, J.R.; Córdova-Esparza, D.M. Kin2. A Kinect 2 toolbox for MATLAB. *Sci. Comput. Program.* **2016**, *130*, 97–106. [CrossRef]
45. Shohieb, S.M.; Elminir, H.K.; Riad, A. Signsworld atlas; a benchmark Arabic sign language database. *J. King Saud Univ. Comput. Inf. Sci.* **2015**, *27*, 68–76. [CrossRef]
46. Shanableh, T.; Assaleh, K. Arabic sign language recognition in user-independent mode. In Proceedings of the 2007 International Conference on Intelligent and Advanced Systems, Kuala Lumpur, Malaysia, 25–28 November 2007; pp. 597–600.
47. Sidig, A.A.I.; Luqman, H.; Mahmoud, S.; Mohandes, M. KArSL: Arabic Sign Language Database. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* **2021**, *20*. [CrossRef]
48. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the inception architecture for computer vision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 2818–2826.
49. Chollet, F. Xception: Deep learning with depthwise separable convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1251–1258.
50. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 770–778.
51. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
52. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv* **2017**, arXiv:1704.04861.
53. Zach, C.; Pock, T.; Bischof, H. A duality based approach for realtime tv-l 1 optical flow. In *Joint Pattern Recognition Symposium*; Springer: Berlin/Heidelberg, Germany, 2007; pp. 214–223.
54. Pérez, J.S.; Meinhardt-Llopis, E.; Facciolo, G. TV-L1 optical flow estimation. *Image Process. Line* **2013**, *2013*, 137–150. [CrossRef]