Journal of Engineering and Applied Sciences 15 (5): 1205-1227, 2020
ISSN: 1816-949X

Advances, Challenges and Opportunities in Continuous Sign Language Recognition

Nada B. Ibrahim, Hala H. Zayed and Mazen M. Selim
Department of Computer Science, Faculty of Computers and Artificial intelligence,
Benha University, Banha, Egypt

Abstract: Sign Language (SL) is the hands spoken language assisting the deaf to understand each other. Understanding SL by vocal people not only paves the way to contribute deaf and dumb in the workforce but also provides a fertile environment for analyzing the human motion and gesturing. Consequently, translating SL sentence into written or spoken language, known as Continuous Sign Language Recognition (CSLR) will help in integrating the deaf and dumb in the society. Most of the surveys in the field of Sign Language Recognition (SLR) spotlight on isolated SLR that mainly deals with words, numbers and letters each in separate. Moreover, these systems are designed to operate in artificial settings for the background, signer dependency and limited vocabulary. Even though for real-life CSLR is the objective, till now there is not a complete survey on CSLR that provides researchers with a comprehensive study on the advances, challenges and opportunities in this field. The presented piece of work analyzes the articles published earlier and illustrates the core stumbling blocks related to CSLR including: the dynamic hand detection and tracking, facial expression recognition, movement epenthesis detection and recognition methods as well as a comparative study on the available benchmark databases. An inventory of the applications which stand to benefit from CSLR are also brightened up. The conclusions and recommendations of this research can be a milestone for developing evolved and efficient CSLR systems.

Key words: Continuous sign language recognition, gesture recognition, movement epenthesis, dynamic gesture, presented, dependency

INTRODUCTION

According to the World Health Organization (WHO) (WHO., 2018), 466 million people all over the world suffer from hearing loss (432 million adults and 34 million children). Nearly one-third of people over the age of 65 years suffers from disabling hearing loss.

The predominance in this age group is from South Asia, Asia Pacific and Sub-Saharan Africa. People are said to be "Deaf" if they are affected by either very little or hearing problem. Deaf invent sign language for communication as the human usually debrief his contemplations and sentiments with the assistance of his facial expressions, body language and hands movement. A sign language is a language depending on the use of manual communication to convey meaning, counter to sound patterns in spoken language. Sign language is considered a natural language as well as the spoken language. Although, they both differ from each other, they utilize the same language faculty by having the same linguistic properties (Stokoe *et al*., 1976; Stokoe Jr, 2005).

As a consequence of the high population of deaf and dumb, the world is showing continuously increasing interest in removing obstacles faced by them in communicating with the community and contributing in the workforce. The most challenge that faces this integration is the communication between deaf and dumb people and the hearing people. The presence of a human interpreter between the two will be expensive and may be unavailable in many situations. Sign language recognition systems play the role of a moderator between the deaf and dumb community and the vocal people by interpreting sign language into either a spoken language or a written one. In the last decades, Sign Language Recognition (SLR) researches reach a peak point, as there is a global trend to integrate the deaf and dumb in the community. Sign language can be divided into three main categories to be recognized: alphabets and numbers, words and sentences. Alphabets and numbers are mainly static gestures. Words are mainly dynamic gestures. But sentences are a mixture between static and dynamic gestures that must be separated to be correctly translated. The recognition of the first two categories are called isolated recognition as one gesture is performed at a time. While recognition of the sentences is called continuous recognition as more than one gesture is performed in a sequence without boundaries.

Corresponding Author: Nada B. Ibrahim, Department of Computer Science, Faculty of Computers and Artificial Intelligence, Benha University, Banha, Egypt

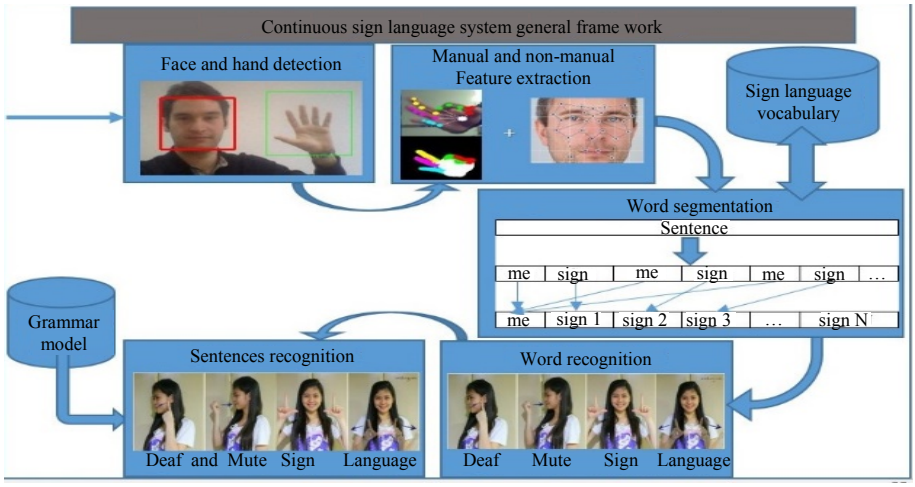

Fig. 1: The general framework for a continuous sign language recognition

Most of the work spotlight on isolated sign language recognition in artificial settings for background, signer Dependency and Limited Vocabulary, although, for real-life continuous recognition is the target. The researches on isolated sign language recognition are reviewed in many surveys like (Kakde *et al*., 2016; Agrawal *et al*., 2016; Sutarman *et al*., 2013) and a review on sign language from hand gesture point of view is presented by Cheok *et al*. (2019). However, there is not a complete survey on CSLR. Kakde *et al*. (2016) presented comprehensive analysis of the surveys and articles published earlier related to CSLR could be used for the design, development and implementation of evolved, robust and efficient CSLR systems for aiding the deaf community. In this research, the challenges and recent advances in the field of CSLR will be discussed to help the researchers in standing on the scope of their future work toward a practical CSLR system.

A general framework for CSLR: CSLR is an automatic system for interpreting sign language sentence to a written or spoken language. This system follows four essential steps. First, the face and the hands of the signer are detected from an image sequence. Then, the description of the sign represented in manual and non-manual features are extracted. After that word segmentation and recognition techniques are applied to divide the sentence into isolated words. Finally, the isolated words are subjected to a grammar-model to achieve sentence recognition. The interpreted sentence can be a written sentence or an audio clip that depends upon the specifications of the system. Figure 1 illustrates the overall stages of CSLR. Detailed discussion of each stage will be illustrated in section 3.

For sign language, the face and the hand hold the dominant information about the sign. The most popularly used algorithm for face detection is Viola-Jones face detector (Viola and Jones, 2001) which relies on Ada-Boosting technology. By the availability of large-scale datasets for face and non-face images, machine learning algorithms were used to extract facial features and correctly classify the face. Deformable Parts-Model methods are considered the latest face detector algorithms (Yan *et al*., 2014). A complete survey on the recent advances in real-world face detection is proposed by Zafeiriou *et al*. (2015). Signs are performed by using one-hand (dominant hand) or by two hands where the second hand is called non-dominant hand. For hand detection both hands of the signer must be detected and tracked. CSLR deals with dynamic hand tracking, rather than static one. Section 3.1 will discuss the methods used for dynamic hand tracking with their challenges and recent solutions.

Feature extraction stage is responsible for extracting informative features vectors that can distinguish efficiently between similar signs. The feature vector is a combi-nation of manual and non-manual features representing the sign. Manual features may be the hand: trajectory, velocity, moment, shape or the finger spelling. Non-manual features can be represented as facial expression, head pose and lip shape. Detecting the landmarks of the face facilitate the extraction of the non-manual features. Facial feature extraction is illustrated in section 3.2.

Word segmentation is the analysis of the sentence to its building units. In spoken language, the silence and phoneme define the words boundaries. But in sign language, there is no explicit definition to the boundaries between the words. Movement epenthesis between the signing words are the hand movement from the end of the sign to the beginning of the next sign. Extracting of movement epenthesis is the core of the word segmentation. The methods tailored for defining movement epenthesis is covered in section 3.3.

Neural networks and statistical-based methods have widely explored recognition of isolated sign language. The ability of the Hidden Markov Model (HMM) to deal with unequal length signs, makes it popular for SLR. But the efficiency of these methods is directly proportional to the size of the available datasets. Till now, the available databases for the CSLR are either small-scaled or medium-scaled (See section 3.4). So, HMM is successfully applied for word-level recognition but failed for sentence-level. Adapting HMM to suit for CSLR has attracted the attention of many researchers. Parallel HMM (Pa-HMM) and enhanced Level-Building HMM (LB-HMM) are the outcome of these researches. Dynamic Time Wrapping (DTW) is famous for the ability of finding an optimal alignment between two given time-dependent sequences under certain restrictions which is suitable for SLR. Moreover, Conditional Random Field (CRF) is a probabilistic model for segmenting and labeling sequence of data. CRF and DTW are the state-of-the-art recognition methods dedicated to CSLR. Most used recognition methods are pointed out at section 3.5.

The grammar of any language is the study of the way the sentences of a language are constructed. This includes phonology, morphology, syntax and pragmatics. Sign language as a language has its own grammar that differs from spoken language. For English spoken language, the sentence is structured as subject-verb-object or subject-verb. While in American Sign Language (ASL), the structure follows the order time-subject-verb object or time-subject-verb where a time frame is inserted before the sentence to indicate past and future events. Additionally in ASL, BE verbs (am, is, are, was, were) or anything to indicate the state of being aren't used as well as the articles (a, an, the) (2012). All these aspects emphasize the need for a grammar-model to correctly interprets the sign language to an understandable spoken or written interpreted language.

MATERIALS AND METHODS

Challenges and recent solutions: The recognition of continuous naturally performed sign language is a challenging problem due to its multi-modal nature. Recognizing of sign language should make use of manual and non-manual features. Manual features are related to hand shape, position and movement. Non-manual features refer to facial expression, head pose and lip shape. Well-defined manual and non-manual features are used in parallel to interpret a sign. Before that, segmenting the sentence of the sign language to its sub-units by handling movement epenthesis should be cured. On the other hand, the accuracy and applicability of any CSLR system must be test on a large-scale dataset. Till now, this dataset does not exist which resulted in the lake of the applicability of most available systems. This section outlines the challenges faced by CSLR system including dynamic hand recognition and tracking, facial expression recognition, movement epenthesis detection and benchmark databases in addition to, the available recent solutions to defeat these challenges.

Dynamic hand recognition and tracking: The sign language depends mainly on hands to represent the words they express. To translate the hands movements to meaningful words, the features of the hands must be considered. These features are called manual features. Dynamic hand recognition became an active field of research in the last few years. Hand can be recognized by either: device-based, vision-based or depth-based techniques (Hint: depth-based methods can be included into vision-based technique).

In the device-based approach, the signer should wear a mechanical glove that has sensors which recorded every motion of every part of the hand e.g., data gloves (Shukor *et al*., 2015), power gloves (Mohandes and Buraiky, 2007), cyber gloves (Caplier *et al*., 2004; Mohandes, 2013) and dexterous master gloves (Hoshino, 2006). Available devices have been proposed with slight functional or mechanical design differences but they all are characterized by easy calibration, stretchable fabrics and a glove-like design adapting different hand sizes (Chossat *et al*., 2015). Throughout the years, great efforts were done to reduce the connectivity between the gloves and the computer. These efforts yield to the appearance of the Bluetooth enabled sensors that appear in wearables such as smart-watches. Recently, Myo armband was developed by Thalmic labs and released for public in 2014. Myo armband is a gesture control device that fits arounds the meaty part of the forearm. It relies on the use of Electromyogram (EMG) signals to read the electrical activity of the muscles, arm rotation and muscles movement then translating them into real-time input by using Bluetooth. Using Myo have some limitations where the amount of the fat, sweat and hair of the arm affect dramatically on the EMG data collected as well as changing the position of the Myo during the performing of the gesture reflected negatively on the accuracy of the features. By Paudyal *et al*. (2016), two Myo devices were used to recognize twenty randomly chosen signs from the ASL that have different orientations and locations of the hands. The evaluation proved the high accuracy of the system that reaches 97%, the applicability of the system for real-time and the ability of the system to be extended. By Abreu *et al*. (2016), Myo was used to recognize LIBRAS (Brazilian Sign Language) excluding the letters H, J, K, X, Y and Z that involve some sort of movement. One-versus-all Support Vector Machine (SVM) classifier acts to distinguish between different gestures. This method failed to differentiate between similar gestures or

that have similar tension on the same fingers. Although, the accuracy of the outmoded device-based approaches, they suffer from high cost and less normality compared to other methods.

In vision-based approach, the signer uses either colored gloves (Wang and Popovic, 2009) or bare hand. A complete survey on vision-based hand gesture taxonomies, representations and recognition techniques used till 2012 is proposed by Rautaray and Agrawal (2015). The colored gloves are normal gloves that have different color for each finger and the wrist. These predefined colors are then detected from an image captured by a 2D camera (e.g. web camera or HD camera) and segmented by using image processing techniques. By Mazumdar *et al.* (2015), a predefined color glove is worn on the palm part of the user's hand. The HSV color space is being chosen to segment the gloved hand from the scene. But this method also lake for normality as the user must wear these gloves all the time. On the other hand, using bare hands techniques have low cost and high mobility as the signer has no connection with any device. Despite that a lot of image or video processing is needed which is computationally expensive resulting in increasing the response time and making this technology not suitable for real-time systems. Dynamic skin detectors (Brancati *et al.*, 2017; Michal *et al.*, 2014) are used to detect the face and the hands from the scene as they considered skin blobs. Hsv and YCrCb were proved to be the most suitable color spaces for skin detection according to (Shaik *et al.*, 2015). In HSV was used to extract the hand from a web camera image after removing the face by using Haar classifier. Then the largest component is labelled as the hand region. As the bare-hand techniques suffer from changing in illumination and hand occlusion with either each other or the face, till now it still an open field of research. This pave the way to the appearance of depth-based techniques that may be considered as some sort of the vision-based approaches. As these techniques have become the trend they will be covered in more details.

In depth-based techniques, the 3D hand surface is acquired by using a camera and a beam of light (depth mapping). This leads to having the depth at every pixel, so, the shape of the object can be measured even if an occlusion occurs. These methods are less affected by changing in illumination and have low cost compared with device-based methods. On addition, it has high normality in motion as no devices is connected to the signer. Depth-based techniques combine the advantages of device-based and vision-based methods. The core of this technology is the RGB-D images that are captured by one of three main sensing technologies: stereo cameras, Time-of-Flight (ToF) cameras and structured light.

By using stereo cameras, the hand is captured by using two cameras from two different fields of view (stereo vision). Then, the shape of the hand is reconstructed from the depth map and tracked. The main advantage of this method is that it does not need any flashing lights that may distract the user (Sandbach *et al.*, 2012). Despite of that using two cameras is expensive than using one. One of the systems that uses the stereo vision is the leap motion which is primarily designed for hand gesture and finger position detection in interactive software applications. The facilitated API of the leap motion provides the user with the direct mapping of the hands and the fingers without the need to take a plunge in piles of raw data. Leap motion is integrated in many fields: education, healthcare, Virtual Reality (VR) and autos. However in certain configurations some visible fingers could be lost, especially, if the hand is not parallel to the camera. Furthermore, nearby objects can easily confuse fingers, like bracelet or sleeves edges. Leap motion was used widely in different SLR systems, like American SL (ASL) (Chuan *et al.*, 2014), Arabic SL (ArSL) (Mohandes *et al.*, 2014), Pakistan SL (PSL) (Raziq and Latif, 2016), Greek SL (GrSL) (Simos and Nikolaidis, 2016) and Mexican SL (MSL) (Najera *et al.*, 2016). Sometimes two leap motions can be used in perpendicular to each other to over-come the occlusion between the fingers or the two hands by mixing the field of view of each device (Mohandes *et al.*, 2015).

Structured light technology relies on projecting of one or more encoded light pat-terns on the hand. Then, a camera from a different direction is used to measure the deformation of the pattern on the hands surfaces to detect the hands shape and depth. The used encoded light patterns may be color, binary-coded or fringe patterns (Rautaray and Agrawal, 2015). It is also possible to use infrared light instead of visible light to reduce the distraction of the user at high speed systems. Structured light has low cost as one high speed camera with a projector is needed. It suitable for real-time acquisition of sequences of 3D hand surface. A major issue with structured light is that the depth images have holes because some areas cannot be seen by both the projector and the camera (Chen *et al.*, 2013). The Microsoft Kinect Xbox 360 (Kinect V1) is the most widely used system that is based on structured light. It can track the skeleton of the body and 20 joints of two bodies at a time. Detecting and tracking the face is robust and feasible by using the Kinect. It is mainly being used in body detection and tracking, action recognition and gamming.

Time-of-Flight (ToF) technology is based on measuring the time that light emitted by an illumination unit requires to travel to an object and back to the sensor array. ToF utilizes the Continuous Wave (CW) intensity modulation approach. Because their distance calculation is computationally simple, ToF cameras can achieve high frame rates which make them suitable for real-time applications. The main advantages of ToF cameras are

their dense depth map that covers every pixel, unlike structured light depth maps that may contain holes. Due to these pros, Microsoft uses this technology in its new version of Kinect (Kinect for Xbox One i.e. Kinect V2). The new Kinect is characterized by full HD color images, keeping track of 6 bodies at a time and its larger field of view compared to the Kinect Xbox 360. A comprehensive review of the application of the Kinect technology in many different areas and difference between the two versions of Kinect is demonstrated in Lun and Zhao (2015). Engaging the leap motion together with the Kinect increase the performance of the systems by integrating the features of fingers from leap motion and the depth information or the voice from the Kinect (Marin *et al*., 2014). A survey on 3D hand gesture recognition is proposed by Cheng *et al.* (2015a, b).

Non-manual features recognition: Beside the manual features (properties of the hand), non-manual features play an important role in recognizing and understanding of the gestures. They have turned to un-ignorable features in a lot of areas. Facial expression, head pose and lip shape are the most common non-manual features. In sign language, there may be two words that have the same manual features (hand movement and shape) but the facial expression gives them a different meaning, like married and divorced words in the Arabic sign language.

The face is a canvas which flashes hundreds of expressions every day (Ekman and Friesen, 1978). Facial expression become an active field of research in the last decades. Improving e-Learning experiences, improving gamming experiences, monitoring patient progress, measuring customer satisfaction and detection of truthfulness during police interrogations are all different scopes that benefit from facial expression recognition. Head pose and lip shape can be included under the field of research of Facial expression recognition. Lip shape is a milestone in the identification of a specific facial expression while head pose is one of the obstacles that affect its accuracy. There are six familiar facial expression counters to neutral: happy (joy), sad, anger, disgust, fear and surprise.

Facial expression is described by Face Action Coding System (FACS) (Ekman and Friesen, 1978) that refer to contraction or relaxation of one or more facial muscles as an Action Unit (AU). Single or mixed collections of AUs can annotate each expression. For example, the happy expression is annotated by AU (6+12) which represent the cheek raiser and the lip corner puller, respectively. To define AU; landmarks of the face (i.e., fiducial facial points) must be detected and tracked. Facial landmarks are the points that locate candidate regions of the face, like nose tip, eyes corners, lip corners and eyebrows. Keep tracking of these landmarks is informative in defining AUs that are considered as local facial features. On the other hand, global appearance features are based on standard feature descriptors extracted from the whole facial region. Combining globaland local features descriptors yields to increasing in the rate of recognizing a facial expression. Briefing, facial expression recognition has four main stages: face detection, landmarks localization, global feature extraction and classification. A comprehensive study on static and dynamic RGB, 3D and thermal approaches on facial expression recognition is proposed by Sandbach *et al.* (2012), Corneanu *et al.* (2016), Martinez and Valstar (2016). By Anil and Suresh (2016), a literature survey on face and face expression recognition is introduced. Facial landmarks detection can be achieved by using constrained local model-based methods or active appearance model-based methods or regression-based methods a brief discussion of these methods is covered by Wang *et al.* (2017). By adding the capabilities of the depth-based technologies to the 2D features of the face this leads to enhance the accuracy and reliability of facial expression detection (Aly *et al*., 2016).

Movement Epenthesis (ME) detection: The sentence of the sign language consists of a sequence of gestures performed one after another without any pauses, like spoken language. By Holt *et al.* (2011), the isolated words sign is divided into three movement phases: preparation, stroke and retraction. Preparation phase is the movement of the hand from rest position to the signing area in front of the body. Stroke phase is the movement of the hands describing the word (i.e., original sign). Retraction phase is the movement of the hand back to the rest state. In the case of continuous signing there is four movement phases: preparation, stroke, transition and retraction. The transition phase is the hand movement after the ending of one sign to the position from which the subsequent sign will be performed. The determination of the beginning and the ending of each sign is a challenging problem as the transitions between any two sequential signs are not defined and the possible hand movements are highly diversified (Li *et al*., 2016). The methods for detecting ME are summarized in Fig. 2. Each of which will be over viewed in detail.

Researchers has two ways to defeat this problem. The first is neglecting the nature of the transition phase and use the isolated-like methods. In this case, the whole sign sentence is treated as one gesture and the same methods used for isolated SLR is being applied. This method lacks sense as it will be needed to form all the possible sentences that can be achieved by small vocabulary that can detect them. This is practically impossible as the language contains thousands of words and billions of sentences. This method is applicable only for small-sized vocabulary.

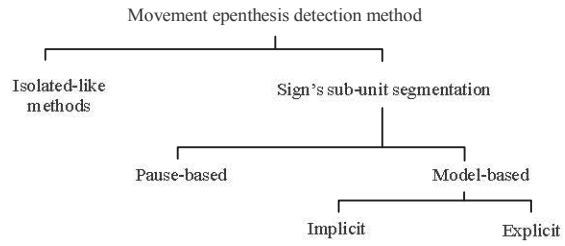

Fig. 2: Different methods for dealing with the ME problem

The second method is using signs sub-units in which the sign sentence is break down into the signs they originally formed from. This can be achieved by either a pause-based segmentation (Khan *et al*., 2014) or a model-based technique. Pause-based segmentation depending on forcing a go-stop scheme in which the signer is asked to put his/her hands down after each sign (Amin *et al*., 2015) or by making a pause for few seconds between the signs (Baranwal *et al*., 2015) and this is not the case in real-life. It can also rely on the changing in the velocity of the hand, like using the zero-crossing of the right-hand acceleration to divide the sentence into sub-units then use the dynamic time warping to threshold the similarity between the hand shape features of the consecutive sub-units to extract the words (Zadghorban and Nahvi, 2018). On the other hand, the model-based methods are the most popular, trending and challenging techniques. They can be further be divided into two main techniques: implicit modeling and explicit modeling.

Explicit Modeling of ME was first proposed by Volger (Vogler and Metaxas, 1997) in 1997. The problem was treated as spatial vector and K-means clustering with least distance criterion on the start and the end points of the isolated signs was used to produce the less possible combining models. But the ME is a temporal sequence of the vector. Modified K-means clustering with DTW criterion was applied by Gao *et al*. (2004) to make use of Temporal clustering. By combining the temporal and the spatial characteristics by Yang and Sarkar (2006), Spatiotemporal characteristics are used to extract key frames then CRF is applied to these frames to decide whether it is a ME or not. Later by Bhuyan *et al*. (2014), CRF is trained with one label for each gesture and an extra label is added to determine the non-gestural movements. Universal transition model being combined with HMM is used by Koller *et al*. (2015) Yang *et al*. (2007) to model ME. One of the cons of the explicit modeling of ME is the quadratically relation between the number of the MEs and the number of the signs in the dataset. One way to overcome this demand for large training dataset is using an analogous concept to phonemes in speech, thus, reducing the set of possible units to model (Yang *et al*., 2007). This yield to the appearance of implicit modelling of ME.

Implicit modeling of ME implies modeling of signs only and consider any other signs as MEs. By combining the temporal and the spatial characteristics by Hyeon-Kyu and Kim (1999), an HMM-based threshold model that calculates the likelihood threshold of an input pattern and provides a confirmation mechanism for the temporarily matched gesture patterns was provided. This work was extended by Kelly *et al*. (2009a, b) for calculating a probability distribution of a two-hand input sign using continuous multidimensional observations. By Li *et al*. (2010), the amplitude of 4-channel EMG signals of the right forearm where used to automatically detect signs and then decision tree and multi-stream HMM are applied to the recognition of sign language signs sub-units. Iterated Conditional Modes (ICM) is proposed by Nayak *et al*. (2009) where the starting point and the width of the sub-units are sampled and updated by using joint conditional distribution until the parameters converge to stable solution. By Yang *et al*. (2007), an enhanced level-building algorithm coupled with trigram grammar-model is used to segment the continuous series of signs. Then, Nested dynamic programming acts to improve this algorithm (Yang *et al*., 2010). Later by Simao *et al*. (2016), unsupervised method is used to segment the continuous motion features coming from Cyber II-glove sensor into static and dynamic patterns by using Genetic Algorithm (GA) generated threshold. Then, sudden inversions of movement direction are analyzed.

Benchmark databases: A sign language recognition system has one essential target which is interpreting sign language to spoken or written language. In case of spoken languages, translating between two different languages require the availability of a huge equivalent vocabulary from both languages to make this translation applicable. For example, in the Google Translate engine, it was clear that as more data become available for the training model the more efficient the Arabic-English statistical machine translation system works. To reach an accuracy of nearly 53% more than 200 billion words are used. On the other hand in case of sign language small-scale equivalent vocabulary are available. There are a few benchmark databases for different sign languages but they still have small-scale vocabulary as the largest database available is RWTH-Phoenix-Weather which includes only 1,558 sign words. This lack of databases makes any available system inapplicable and not suitable for real-time applications. The recent available continuous sign language benchmark datasets will be covered excluding these mentioned by Agrawal *et al*. (2016). A comparison between all the available datasets is pointed out in Table 1. The "SL" column stated the sign language used, the "CD" Column Declared the utilized capture device, the "#" is an abbreviation for number, the "123" announced the presence of numbers in dataset and the "ABC" shows that the dataset including alphabets.

Table 1: Comparison between CSLR available datasets

Dataset	SL	CD	No. of signers	No. of words	123	ABC
RWTH BOSTON-50	ASL	------	3	50	x	x
RVL SLLL	ASL	------	14	43	√	√
RWTH BOSTON-104	ASL	3 Black and white cameras+color camera	3	104	x	x
RWTH BOSTON-400	ASL	------	4	406	x	x
Signs World Atlas	ArSL	Canon power Shot A490	10	500	√	√
Arabic benchmark databas	ArSL	Kinect V2+leap motion+ordinary HD camera	4	1,216	√	√
Multi-modal gesture	ItSL	Kinect V1	27	13,858	x	x
RWTH-PHOENIX-Weather	GSL	------	9	1,558	x	x

RVL-SLLL American Sign Language database (Martinez *et al*., 2002) consists of 2576 videos corresponding to 14 Different native signers of ASL. Two different lighting conditions were used. The dataset is sectioned into two parts. The first part shows separate video clips for the motion primitives and handshapes that include alphabets, numbers from 1-20, handshapes and signs in isolation to show different motions. The second part consists of a carefully selected set of two or more sentences in a paragraph to show connected discourse. The technical report about the dataset is available at (Wilbur and Kak, 2006).

RWTH-BOSTON-400 (Dreuw *et al*., 2008a, b) is a video-based ASL database including 843 sentences that are split into 633 train sentences, 106 development sentences and 104 evaluation sentences. The vocabulary represents 406 unique words in total with 7768 running words excluding pronunciation information. Four signers performed the training dataset: 2 males and 2 females, where the males perform 454 videos and the rest were done by females. By annotating the position of the signer's both hands in 15 videos of the dataset, the RWTH-BOSTON-Hands database was created with 1119 frames in total.

RWTH-PHOENIX-Weather (Forster *et al*., 2012) is considered the largest continuous video-based sign language recognition and translation dataset till now. It is a weather forecast Ger-man public TV show recorded in German Sign Language (GSL). It comprises 1,980 GSL sentences including 911 solitary words carried out with seven signers. The videos are manually annotated to distinguish sentence boundaries, word boundaries, pronunciation, the utterances of the announcer and translation of the glosses into written German. A subset of the dataset acquires additional annotation for the center point of the hand palms and the nose tip. Recordings from years 2011-2013 were added to extend the dataset to comprise 6,861 sentences and 75,107 running glosses performed by 9 signers. The overall vocabulary reaches 1,558 different signs.

Signs world atlas (Shohieb *et al*., 2015) is an image/video based ArSL database. It comprises 500 manual and non-manual signs performed by ten signers including: hand shapes, alphabets, numbers, isolated words, continuous sentences, lip movement and facial expressions. The dataset was captured by Canon Power Shot A490 digital camera under controlled lighting conditions. This was the first attempt for developing a benchmark database for ArSL.

Arabic sign language benchmark database (Alfonse *et al*., 2015) has been captured by three different heterogeneous sensors: Kinect V2, leap motion and an ordinary HD camera. The number of signs/words is: 1,216 signs/words including the alphabets and numerical and comprises 531 sentences. Four sign language experts have captured the complete dictionary under different illumination conditions. The four sign experts are deliberately picked as: two left-handed signers and two right-handed signers.

Recognition: The main objective of the recognition step is increasing the recognition rate of the framework by matching the input with the most appropriated output. The following subsections will cover the state-of-art recognition methods dedicated for sign language recognition.

Hidden Markov Model (HMM): Hidden Markov Model (HMM) is a type of statistical model firmed in a Bayesian framework that has proved itself in recognizing speech, human activity, gesture, handwriting and sign language. The state-based nature of HMM makes it suitable for representing the signs by capturing the change and variation in the duration of signs over time (Pashaloudi and Margaritis, 2002). Starner and Pentland (1997) are the firsts who employ HMM in recognizing a set of ASL sentences with a recognition rate of 92% with no grammar. An HMM consists of a few states and transitions between them. The system is engaged in one of the HMM states at any applied time. Transition from one state to another takes place in discrete time intervals which are regularly spaced. Each state is distributed according to a previously calculated probability distribution function to generate anappropriate output as well as each transition is accompanied by a probability to be selected. Despite of that HMM has some cons like poor performance when training data, weighting features dynamically is not available and violations of the stochastic independence assumptions.

Regular HMMs can handle sequentially natured problems but they fail with those of parallel nature as sign

language. Researchers have proposed extensions to HMMs to model the parallelism such as Factorial Hidden Markov Models (FHMMs) or Coupled Hidden Markov Models (CHMMs) (Brand *et al*., 1997). But they must be trained on all the combinations of parallel process which make them practically suffer from scalability problems and computationally expensive. To solve this issue, parallel process of the sign is modeled independently by using Parallel HMMs (PaHMMs) (Vogler and Metaxas, 1999), then the probability of the outputs are combined to make a decision. Product HMM (PHMM) (Yu *et al*., 2011) is another effort where it lies between the completely synchronous fusion scheme and PaHMMs. It obligates the parallel process to be asynchronous in the model but synchronous at the model's boundary.

Coupled Hidden Semi-Markov Model (CHSMM) is used to handle three essential cons in regular HMM: model complex multi-scale structure by including a hierarchy of hidden states, the duration probability of a state by introducing explicit state duration models and multi-channel extensions by the dependency of the current state on all the previous states (Natarajan and Nevatia, 2007). However, CHSMMs are complex and computationally expensive. By the appearance of the concept of subunits, Gesture-Threshold HMM (GT-HMM) (Kelly *et al*., 2011) has been applied to develop a gesture subunit initialization technique to create HMM states which model gesture sub-patterns and implement a threshold HMM to compute a dynamic epenthesis likelihood of input gestures. On the other hand, Multi-Stream HMM (MSHMM) (Li *et al*., 2012) is employed for fusion the information of ACC and EMG to recognize single-handed gestures as well as modeling the sign sub-words per- formed by double hands.

Dynamic Time Wrapping (DTW): Dynamic Time Warping (DTW) is a time series analysis technique for measuring similarity between two temporal sequences which may vary in speed by warping the sequence non-linearly in the time dimension. It can be easily implemented with dynamic programming. In signing process, the sequences are the representations of two signs: a model sign and a query sign. DTW matches frame by frame the model and the query signs, described by the feature vectors, to find the minimal cost warping path between both signs.

DTW is an exemplar-based matching procedure that requires comparing of a query sign to all the model signs which is computationally exhaustive to the system and is time consuming. The statistical nature of the HMM makes it more suitable for sign language recognition than DTW, so, the use of the DTW has faded with time. The appearance of Statistical DTW (SDTW) (Lichtenaur *et al*., 2008) has proved its worth, achieving the highest recognition rates compared to HMM. The main difference between DTW and SDTW is that SDTW warp a signal onto a reference model and regard the time normalized signal as a fixed-size feature set where the features are selected statistically to remove irrelevant and redundant parts and dimensions. SDTW is computationally attractive as time warping is solved by dynamic programming and the classification step is even significantly less costly. By Zhang *et al*. (2014), DTW was hybridized with HMM to obtain the possible time interval to locate the endpoint of the sign in the sentence directly.

Conditional Random Fields (CRF): Conditional Random Fields (CRFs) are a type of discriminative undirected probabilistic graphical model used to encode known relationships between observations and construct consistent interpretations. Whereas a discrete classifier predicts a label for a single sample without considering neighboring samples, a CRF can take context into account. CRF has been used successfully to label and segment text sequential data, segment images and recognize whole body human movement.

Yang and Sarkar (2006) are the firsts to employ CRF in detecting the coarticulation in sign language. CRF is applied to label the set of key frames that describe the sign language sentence as a linear chain, to their corresponding labels as either: coarticulation or sign. The CRF Model uses pairwise probabilities over states and observations for each time instant. Each observation is associated with a state label. Violations of the stochastic independence assumptions of HMM prevent the representation of the relationship between the labels and the observation over time while CRF easily embed context dependence and hence is flexible. Experimental results prove the efficiency of the CRF over HMM.

A two-layer CRF consisting of a Threshold model with CRFs (T-CRFs) and a regular CRF is applied to a CSLRS by Yang and Lee (2010). The hands motion and location are employed by the T-CRF to select the appropriate label to a pattern as: discriminate signs, finger spellings and non-sign patterns. The regular CRF is used to recognize the common patterns between signs. To merge the SIGN sub-segments and recognize the signs, authors by Kong and Ranganath (2014) use a two-layer CRF Model that differs in structure from that of (Yang and Lee, 2013). The lower layer of the recognition module consists of four independent linear CRF Models to recognize and output sequences of phoneme labels in the component channels, from the corresponding SIGN sub-segment sequences. The corresponding phoneme labels from the four channels are then combined and input to the upper layer semi-Markov CRF for sign recognition.

Support Vector Machines (SVM): Support Vector Machines (SVM) are hypothetical classifiers that use

maximal-margin hyperplane to split the input variable space into two classes. The hyperplane is learned from training data using an optimization procedure that maximizes the margin. To search for the coefficients of the hyperplane, Sequential Minimal Optimization (SMO) method is efficiently employed to break the problem down into sub-problems that can be solved analytically rather than numerically. Multiclass ranking SVMs, one-against-all classification and pairwise classification are several approaches to adopting SVMs to classification problems with three or more classes. SVM has been used in text and hypertext categorization, Handwriting recognition, classification and regression.

The signing problem is a multi-class classification problem. By Kelly *et al.* (2010), a one-against-all classification SVM is used to recognize 10 gestures and 23 Irish Sign Language (IrSL) alphabets. Set of SVMs is trained on data extracted from labeled images. Features are extracted from an unknown hand image and the SVMs use the data to estimate the most probable hand posture classification. For each posture class, two SVMs are used to calculate the probability that image posture belongs to a class. Regular SVM suffers from the high algorithmic complexity and extensive memory requirements of the required quadratic programming in large-scale tasks.

Latent variables have long been used to model observations in generative probabilistic models such as Hidden Markov Models. Latent SVM treats the desired state value as latent variables and consider different correlations into potential functions in a discriminatory manner. The main pros of latent SVM over regular SVM is its ability to weakly supervise latent parts within an element to be recognized. This ability nominated it for sign language recognition. By Sun *et al.* (2015), the desired state is the discriminating capability of each frame in videos. Three types of potential functions are specially formulated to encode the latent variables representing frames into a unified learning framework. The best configurations of latent variables for all frames are searched via. optimization and the sign language videos are classified. By Yang and Lee (2013), the SVM is used to recognize facial expressions as non-manual signals when the system fails to distinguish signs by detecting manual features and hand shape.

Deep Learning (DL): Artificial neural networks are biological brain-like computing systems that can process information from external inputs to give a convenient output. Neurons are the smallest building units of an ANN that retain highly interconnected with weighted connections to process information dynamically according to the variation in inputs to give a decision. History of ANN backs to the 1940's and according to (Goodfellow *et al.*, 2016) it has been evolved by three waves: perceptron in the 1940-1960's, back-propagation in the 1980-1990's and Deep Learning Network (DLN) in 2006 till now:

- Perceptron is a feed-forward uni-direction network consisting of a single input layer that is connected directly to the output layer through weighted connections
- Back-propagation networks use bi-direction training technique that acts to up-date the weighted connections to reduce the computed error
- Deep learning network is a huge-scaled supervised ANN capable of extracting meaningful features from quite complicated inputs with robust accuracy

DLN has been utilized in the last decade in solving problems popular for its complexity in some disciplines such as: computer vision (Aldoukhi *et al.*, 2019), speech recognition (Tu *et al.*, 2019), action recognition (Tran *et al.*, 2015) and natural language processing (Bacchi *et al.*, 2019). It is a model-free ANN that learns features and classifies the input data jointly with high accuracy without any pre-processing relying on thousands to millions labeled dataset entries. On the other hand, it is computationally expensive and training such networks may take weeks. High performance hardware and software infrastructure is critical for DL to afford high computational workload. General-Purpose Graphics Processing Units (GP-GPU) with their empirical capabilities in form of parallelism and memory bandwidth has proved its worth to be one of the leading platforms for DL. To make benefit of the high accuracy of the DLN, transfer learning (Lu *et al.*, 2015) has been used to train a related small dataset by transferring the parameters of a stabilized DLN that was applied to a same domain tobe the initial parameters of the small-scaled dataset DLN. This type of learning fine-tunes DLN to suit a novel problem with its prior learned set of features and dramatically decreases the training time. Recurrent Neural Network (RNN) and Convolution Neural Network (CNN) are special-scaled editions of DLN that have attracted SLR researcher's attention, since, 2016.

RNN is a sequence based DLN for processing temporal sequences, e.g., natural language. It acts as an internal memory to learn temporal dynamics in sequential data. For continuous gesture recognition, RNN outperforms non-recurrent models in predicting the beginning and ending of gesture in a sequence with high accuracy and can learn hierarchies of motion (Pigou *et al.*, 2018). From memory point of view, RNN can be categorized as: simple or Long-Short Term Memory (LSTM) or Bi-directional LSTM (BLSTM). Simple RNN (SRNN) is a simple-state neuron that remembers the previous input. Although, it is computationally cheap, it vanishes the gradient which

holds valuable information and leads to blowing up errors. To overcome gradient vanishing, LSTM (Hochreiter and Schmidhuber, 1997) is proposed by adding a cell state to the hidden state of SRNN to preserve long term dependencies. But this will increase parameters propagation, so, three gates are added to reduce the dimensionality (input, forget and update gates). Both SRNN and LSTM rely only on previous frames to predict the gesture ignoring upcoming frames that can empirically improve the recognition. BLSTM (Camgoz *et al*., 2017) is a two-layered LSTM where each layer is operating in opposite directions for preserving forward-backward dependencies. Although, this technique improves the recognition, it becomes inapplicable to on-line incidents. The main drawback of RNN is its disability to perform as an end-to-end model by depending on a preprocessing step of spatial or global spatio-temporal features extraction however, its high accuracy in temporal features extraction.

On the other hand, CNN is a large-scaled DLN dealing with a huge number of inputs represented in a matrix form e.g., high-resolution image. It is a frame-by-frame model affording spatial features only ignoring temporal features. To employ CNN in gesture recognition, different temporal feature extraction models have been integrated such as: Connectionist Temporal Classification (CTC) or RNN. CTC (Graves *et al*., 2006) is a probability-based sequence-to-sequence model. The probability of observing a sequence depends on broadening the vocabulary with silence and transition label, then ignore this label when the sequence is detected. However, CTC can deal with different length sequences, the number of corresponding extended vocabulary will expand as duration of the input sequence increases.

DLN has been used in spotting recognition strategies where the continuous sequence is segmented into isolated sub-sequences. A two-stream RNN (2S-RNN) (Chai *et al*., 2016) is used to spot the continuous gestures into isolated ones by the consistence of a hand-crafted hand detector. The RGB and depth sequences are fed to two parallel SRNN to extract hand features. Then, these features are combined with a fusion layer to be fed to an LSTM to model the context information of each gesture. Finally, the class label is generated by a probability-based soft max layer. Instead of employing raw RGB and depth images (Wang *et al*., 2017) uses Depth Dynamic Images (DDIs) and Depth Motion Dynamic Images (DMDIs) from a bidirectional rank pooling of an image sequence. Both images are fed to CNN while the RGB and the saliency sequence is fed to 3D-CNN LSTM. The maximum score in the average-score fused vector illustrates the probability of the test sequence while its index represents the predicted class label.

End-to-end DLN acts to solve the continuous recognition problem directly from the videos without any preprocessing steps. A 3D-CNN (Tran *et al*., 2015) evolves traditional CNN by its capability to extract arranged spatio-temporal features from multiple frames at a time in the absence of a separate temporal model. However, 3D-CNN suffers from overfitting and increasing of parameters per layer. Temporal convolutions (Pigou *et al*., 2018) defeats 3D-CNN by extracting hierarchies of motion in addition to temporal features from the first layer. By Cui *et al*. (2017), combining temporal CNN with BLSTM and detection network through a three-stages optimization process, positively affected the performance of the network by tuning feature extraction and sequence learning components. This method avoids the overfitting problem. A recurrent-3DCNN network with a CTC is employed by Molchanov *et al*. (2016) which correctly classify a gesture from multi-model frames without spotting the sequence. This network has proven its applicability for on-line recognition by the early classification of a gesture without processing an enormous number of frames with zero or negative lag. Video-sentence relevance for generating semantic sentences has been covered by Huang *et al*. (2018) where two stream 3D-CNN is used as a global-local feature extractor to represent the video and words are distinguished from a given vocabulary by one-hot vector. Words are then mapped to the same Laten Space (LS) to model video-sentence relevance by bridging the semantic gaps. Finally, the output of LS is fed to a Hierarchical Attention Network (HAN) to generate semantic sentences. Merging separately sequence-to-sequence pre-trained SubUNets Models and feed them to BLSTM and CTC layers impacts the performance of end-to-end network by self-tuning (Camgoz *et al*., 2017). A combination of HMM with DL algorithms outperforms traditional end-to-end net-work depending only on DL. A CNN-2BLSTM-HMM network has been introduced by Koller *et al*. (2017) where an iterative re-alignment using full frames is proposed to reach stability after a few iterations.

Evaluation: To evaluate the results, the criterion used in continuous speech recognition was employed where two primary metrics are used to estimate the performance of the recognition system: the Word-Error Rate (WER) and the Perplexity (PP) over test data. The former is the mostly used metric which is based on three error types: substitution where an incorrect gesture was substituted for the correct one; deletion where a correct gesture was omitted in the recognized sequence and insertion where an extra gesture was added in the recognized sequence. Deletion (D), Insertion (I) and Substitution(S) errors are possible in the unrestricted grammar systems. Many of the (I) errors correspond to signs with repetitive motion. The recognition rate was then calculated as (Zhang *et al*., 2014):

$$WER = 100 * \left(1 - \frac{S+I+D}{\#\text{of gesture}}\right) \quad (1)$$

Perplexity is an information-theoretic assessment of the system predictive power that is computationally inexpensive and can be computed separately from a recognition system. Perplexity (Klakow and Peters, 2002) is defined as:

$$pp = \exp\left(-\frac{1}{N}\sum_{i=1}^{N} \log p(w_i/h_i)\right) \quad (2)$$

where, w_i is the ith word in the test set and h_i is the preceding history. The appearance of DLN was accompanied by a novel metric called Jaccard Index. Which measures the similarity between ground truth sequence and predicted one. This index pays attention to true values in a sequence only without concerning deletion or insertion and it is tailored for continuous sequential data. According to (Wan *et al*., 2016), modified Jaccard Index for a sequence s with ls true labels is:

$$J_s = \frac{1}{I_s}\sum_{i=1}^{L} J_{s,i} \quad (3)$$

where, the Js, i is the Jaccard index of class i in a sequence s which can be computed as:

$$J_{s,i} = \frac{G_{s,i} \cap p_{s,i}}{G_{s,i} \cap p_{s,i}} \quad (4)$$

where, $G_{s,i} \cap P_{s,i}$ is the overlap between ground truth and predicted frames of a class i in the sequence s while $G^{s,i} \cup P^{s,i}$ is the overall frames of the ith class in both the ground truth and the predicted sequences.

RESULTS AND DISCUSSION

The field of CSLR has highly promoted, since, its beginning with the appearance of high technology capturing devices for motion and depth information. Some of the previous researches done in the field of CSLR were gathered to be analyzed according to the previous discussion over the last decade. The analysis comprises of: Sign Language (SL), tracking method of Dynamic Hand Movement (DHDM), overcoming Occlusion (Occl.), the features either its manual or non-manual, the movement epenthesis Extraction Methods (ME), the Grammar-Model (GM), the recognition method and the dataset used. Attempts have done to gather as much research as possible in this area with the help of Google Scholar trying to obtain a comprehensive survey of developments in this field. The exhaustive researches accompanied by their analysis are listed in Table 2.

Most of the treatises rely on the vision-based as well as the depth-based hand detection techniques. The former makes the signer moves freely and do not encumber him with a bunch of sensors and wires. Although, the projection of 3D world in a 2D image may cause the loss of some information and a lot of image or video processing operations are needed to extract an informative description of the sign. While the latter has attracted the researcher's attention after the release of Microsoft Kinect in 2010. The appearance of depth-based capturing technology enhanced the hand's detection and tracking in CSLR systems by providing the depth information corresponding to the joints of the body in additional to, the accurate location of the facial landmarks. The high accuracy of this information helps greatly in integrating non-manual features (like facial expression, head pose and lip movement) with the manual features from the hands to correctly classify similar gesture in hand

Table 2: Analysis of some vital literature related to CSLRS

References	S	DHDM	Occl	Features	ME	GM	Rec.	Dataset
(Sarkar *et al*., 2010)	ASL	Vision-based →bare hands	X	Relational distributions+ facial expression and head motion	Implicit→distances between SoPF+non-manual information	---	Simple distance measure	25 sent 39 words
(Yang and Lee, 2010)	ASL	Vision-based →bare hands	X	Global+local features	Implicit→Enhanced level building (Elb) algorithm	Tri-gram	Levenshtein distance	- sent - words
(Assaleh *et al*., 2010)	ArSL	Vision-based →bare hands	X	Zonal coded DCT coefficient of a thresholded image difference	Implicit	Unrestricted	HMM	40 sent 80 words
(Kelly *et al*., 2010)	IrSL	Vision-based →bare hands+ color glove	X	Eigenspace size functions+Hu moments	Implicit	---	SVM	10 gestures +23 alphapets
(Huang and Tsai, 2010)	TSL	Vision-based →color glove	X	Hand's shape and orientation+ 7 Hu moment	Implicit	Tri-gram	HMM	20 sent-words
(Pitsikalis *et al*., 2010)	ASL	Vision-based →bare hands	Linear forward	x, y coordinates of the dominant	Implicit→data driven sub-unit	Unrestricted	HMM+ k-means	BU-400

Table 2: Continue

Ref.	S	DHDM	Occl	Features	ME	GM	Rec.	Dataset
			backward prediction and template matching	hand centroid using as reference point+the centroid of the signer's head and its aforemetioned products	construction		clustering	
(Yang and Lee, 2010)	ASL	Vision-based →bare hands	X	Hand's motion shape and location	Implicit	---	Two-layer CRFs	515 words 50 finger spellings
(Yu *et al*., 2011)	TSL	Vision-based →color glove	X	Hand's position, shape and orientation	Implicit→hand location-based coarse segmentation +hand shape-based fine segmentation	---	PHMM	3 sent 18_23 words
(Kelly *et al*., 2011)	-----	Vision-based →color glove	X	Hand's position +face and eye positions+face width	Implicit→gesture model	---	GT-HMM	160 sent 8 words
(Li *et al*., 2011)	CSL	Device-based →ACC and sEMG	X	Hand shape, orientation and movement	Implicit→amplitudes of EMG signals	Bigram+ trigram	GMM and MSHMM	40 sent 175 words
(Sarkar *et al*., 2011)	ASL	Vision-based →bare hands	X	Hand hypothesis and its relative movement to face+facial expressions+head motion	Implicit→DTW signeme extraction	---	Enhanced level building	25 sent 65 words
(Yang and Lee, 2011)	ASL	Vision-based →bare hands	X	Hand shape, motion and location+facial expression	Implicit	---	CRF+AAM	98 sent 24 words
(Gweth *et al*., 2012)	GSL	Vision-based →bare hands	Linear prediction +template matching	Hand shape, direction and velocity	Implicit→raw canonical subunits models	---	HMM	GSL corpus
(Li *et al*., 2012)	CSL	Device-based →ACC and semg	X	Hand shape, orientation and movement	Implicit→average energy of sEMG	Unconstrain	GMM and MSHMM	200 sent 120 words
(Maraqa *et al*., 2012)	ArSL	Vision-based →color glove	X	Fingertips relative positions and orientations to the wrist and to each other	Implicit→correlation value or output of the neural network for every two successive frames	---	Recurrent neural networks	30 gesture
(Theodorakis *et al*., 2012)	GrSL	Vision-based →bare hand	Linear prediction +template matching	Hand shape, position, direction and velocity	Implicit→raw canonical phonetic movement+hand shape subunits	---	HMM	52 sent 142 words
(Tolba *et al*., 2013)	ArSL	Vision-based →bare hands	X	Image signature using PCNN	Implicit	---	Graph matching algorithm	30 sent 100 words
(Forster *et al*., 2013)	GSL	Vision-based →bare hands	X	Hand shape+ hand and head orientation and position+mouth and eye openings +degrees of eyebrow raise	Implicit	---	HMM	SIGNUM+ PHOENIX
(Yang and Lee, 2013)	GSL	Depth-based→ stereo cameras	X	Hand motion+ fingerspelling+ facial expressions	Implicit→hierarchical CRF	---	Boost map embedding+ SVM	98 sent
(Sun *et al*., 2013)	ASL	Depth-basedt →kinec	X	HOG+body pose +hand's shape and motion	Implicit→extract discriminative and representative frames	---	Latent SVM	1971 sent 73 words
(Zhang *et al*.,	ASL	Depth-based	X	Skeleton	Implicit→DHMM	---	Threshold-	180 sent

Table 2: Continue

Ref.	S	DHDM	Occl	Features	ME	GM	Rec.	Dataset
2014)		→kinect		information +depth map for hands and elbows			based HMM-DTW	34 words
(Yang *et al*., 2014)	ASL	Depth-based →kinect	X	Skeleton information+depth map for hands and elbows	Implicit→hierarchical CRF	---	Boost map	180 sent 34 words
(Kong and Ranganath, 2014)	ASL	Device-based→ cyber glove+ polhemus fastrack	X	Handshape, palm orientation and hand position	Implicit→fusing the outputs of independent SVM and CRF classifiers by a BN	---	Two-layer CRF	74 sent 107 words
(Theodorakis *et al*., 2014)	ASL + GrSL	Vision-based →bare hands	Employing a forward backward linear prediction for the estimation of each body part ellipse's parameters	Centroid of the head+hand's centroid, velocity and instantaneous direction +concatenated histograms of each level in the 3-level pyramids visual vocubulary	Implict→datadriven unsupervised dynamic/static sequentiality	---	PaHMMs	ASLLVD +GrSL lemmas+ BU400
(Ong *et al*., 2014)	GSL+ GSL	Depth-based →kinect	X	Binary features vector from (Ong *et al*., 2012)	Implict→hierarchical sequential inetrval patten tress	---	HSP-forestts	3 different datasets
(Koller *et al*., 2015)	GSL	Vision-based →bare hand	Spring-like function centered on the face postion	HOG-3D featuress, trajectories and handedness of the hands+high level face features	Explict→learning a universal transition model, being a background HMM	Tri-gram	HMM	SIGNUM+ RWTH PHOENIX WEATHER
(Tubaiz *et al*., 2015)	ArSL	Decice-based →DG5-V hand data gloves+ vision-based	X	Bend in each finger+hand acceleration and orientation	Explicit→manually assign the sign boundaries	---	MKNN	40 sent 80 words
(Zhang *et al*., 2015)	CSL	Depth-based →kinect	X	3D postion of the hand and the both elbows	Implicit→warping templates for each signs	---	Dynamic programming	180 sent 30 words
(Sun *et al*., 2015)	ASL	Depth-based →kinect	X	HOG+OP+hands postions, shape and motion+body pose	Implicit	Tri-gram	New latent SVM	63 sent. 28 words
(Cheng *et al*., 2015)	CSL	Device-based accelerometer (ACC) and surface electromyography (sEMG) sensors	X	Palm orientation +hand shape and movement	Pause-based →natural relax for about 12 sec +implicit→gesture coding	---	DTW+ HMM	223 charactors
(Ghotkar and Kharate 2015)		Depth-based →kinect	X	Skeleton joint information +quadrant information+ motion pattern	Implicit→ keywords	Grammar model using Skeleton information	DTW+ inverted indexing	100 sent
(Tripathi and Nandi, 2015)	InSL	Vision-based →bare hands	X	Orientation histogram +PCA	Implicit→gradient based key frame extraction	---	Correlation or Euclidean distance	10 types of sent
(Li *et al*., 2016)	CSL	Device-based →pair of digital gloves	X	Hand shape, position and orientation	Explicit→universal transition modeling using Gaussian mixture based-HMM	---	HMM	1,024 sent 510 words
(Kishore *et al*., 2016)	InSL	Vision-based→ bare hands	X	Horn schunck optical flow+chan -vese active contours	Implicit→velocity vectors computed using HSOF algorithm	---	Back-propagation neural network	1 sent 58 words
(Yang *et al*., 2016)	CSL	Depth-based →kinect	X	Skeleton information from kinect+motion	Implicit→ threshold-HMM	Bigram ---	Fast-HMM -based level building	100 sent 21 words

Table 2: Continue

Ref.	S	DHDM	Occl	Features	ME	GM	Rec.	Dataset
				trajectory descriptor				
(Koller *et al*., 2016)	GSL	Vision-based →bare hands	X	Dynamic programming based tracking of the right hand	Implicit→CNN	---	CNN-HMM	RWTHPHOENIX Weather+SIGNUM +RWTHPHOENIX -weather multi -signer
(Chai *et al*., 2016)	---	Depth-based →kinect	X	Spatio-temporal	Implicit→RNN	---	2S-RNN	ChaLearn LAP ConGD
(Molchanov *et al*., 2016)	---	Depth-based →kinect+stereo cameras	X	Spatio-temporal	Implicit→RNN	---	Recurrent 3D-CNN	ChaLearn LAP ConGD
(Rao and Kishore, 2018)	InSL	Vision-based →bare hands	X	Eigen vector from PCA uniquely represents DCT energy of the hand shape and head	Implicit→ disappearing of the hand	---	Back-propagation neural network	1 sent 18 words
(Wang *et al*., 2017)	---	Depth-based →kinect	X	Long-term spatiotemporal features	Implicit→2S-CNN	---	3D-CNN	ChaLearn LAP ConGD
(Cui *et al*., 2017)	GSL	Vision-based →bare hands	X	Spatio-temporal feature	Implicit→stacked temporal convolution	---	Temporal CNN + BLSTM	RWTH-PHOENIX-Weather multi-signer
(Koller *et al*., 2017)	GSL	Vision-based →bare hands	X	Spatio-temporal feature	Implicit→hybrid CNN-BLSTM	---	Recurrent CNN-HMMs	RWTH-PHOENIX weather+ SIGNUM
(Huang *et al*., 2018)	Diff. SL	Vision-based →bare hands	X	Hand locations /motions+one-hot vector	Implicit→3D CNN	---	LS-HAN	RWTHPHOENIX -Weather sent 178 words
(Koller *et al*., 2019)	GSL	Vision-based →bare hands	X	Sign gloss+mouth shape+hand shapes	Implicit	---	Hybrid multi Stream CNN-LSTM HMM	PHOENIX

movement or to identify a negation in the sentence. The most used manual features are hand's position, shape and movement.

Although, the occlusion affects negatively on the performance of CSLR, most of the researchers neglecting it by considering the occluded hand with either the head or the other hand as one object. For the device-based systems, occlusion is not considered as an obstacle where sensors give the exact position of each hand and the fingertips. On the other hand, vision-based systems employ either a template matching approaches or a forward-backward prediction that depends on the hand's trajectory to overcome occlusion. By the releasing of depth-based devices, depth information has been applied to partially solve the occlusion problem. Yet, to the best of our knowledge, interlocking and crossing of the hands when performing the sign are still unsolved issues that affect dramatically on the hand tracking.

As for the movement epenthesis detection, different methods have been used but it is obviously that researchers tend to use implicit model-based methods more than others. Explicit modeling of the MEs acts to model all the possible hand transition between the signs and accompany it with all the sign models to get a precise presentation of the sign. This is computationally expensive and time consuming. The mainly used implicit model depends on segmenting the SL sentence to sub-signs which is the part of the sign that is robust against the variation of the adjacent signs and the associated ME. These sub-signs are then linked together to be labeled as signs or MEs. Sub-signs extraction can be done manually or automatically.

As shown in Fig. 3b, HMM is ranked as the most preferred recognition method used by researchers compared to SVM, CRF and DTW. HMM was not used in its native state. A set of modifications were applied to HMM to make it suitable for solving the continuity signing problem of CSLR systems like: CHSMM, PHMM, GT-HMM, MSHMM, CNN-HMM and PaHMM. Sometimes, HMM is hybridized with other methods like: DTW, graph matching or K-means clustering to increase the recognition rate and overcome some cons of the native HMM. There were two methods that had promising results compared to that of HMM: enhanced level building that depends on DP and HSP-forests. These two methods cover many aspects of CSL and is tailored to solving the issue of segmenting the SL sentence into its corresponding sub-segments.

Evolution of GPU's computational capabilities resurrects re-evolving neural networks to be more deeper and act as an end-to-end problem solver. For the last 3 years, DLN has attracted researches attention and an intensive workload is done due to its quickly improvement response to state-of-art techniques. The

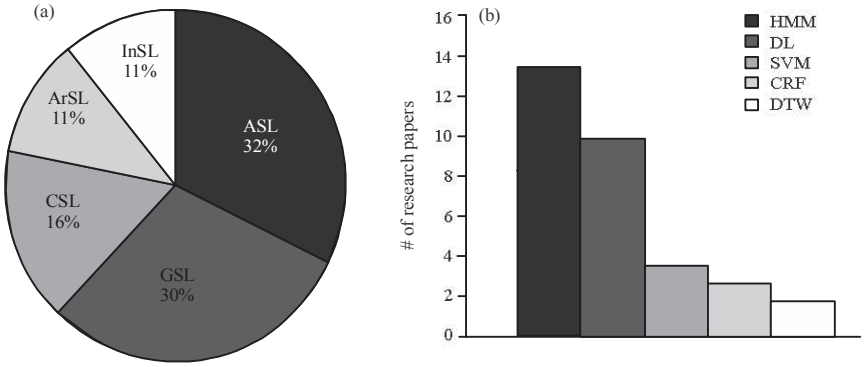

Fig. 3(a, b): Statistics on the wok done in the field of CSLR in the last decade, (a) The SL used and (b) The recognition

main components of DL-depending algorithms are RNN and CNN. These two DLN are merged in different forms to achieve reasonable performance. Lately, the combination between DLN and HMM with a weakly supervised algorithmachieves promising results and considered as state-of-art method for CSLR problem.

CSLR systems act to translate the SL into spoken language. Each SL has its own grammar and syntax. An additional step after recognizing the signs is to put these signs into a reasonable form encountered by a grammar model to enhance the communication between hearing impaired people and the community. N-gram grammar models are employed to put the recognized signs in well-structured understandable sentences. Bi-gram and tri-gram grammar models are rarely used where the probability of the sentence is calculated from the conditional probabilities of each sign given the n-1 preceding signs. The survey reflects researchers neglection of including a grammar model in most of the work done. This may be ascribable to the complexity and high computations required for implementing such models.

Statistics showed that ASL had received more attention by researchers than any other sign language by nearly 32% of the total work done as shown in Fig. 3a. Despite the availability of benchmark datasets for different SL (under variance capturing techniques, lighting conditions, backgrounds, lab environment and real-life circumference), the survey indicates that most of the work done prioritize building their own datasets. The researchers escape to this solution to facilitate the segmentation of the MEs by either controlling the speed of the signing or by obligating a pause for 1-2 sec between signs or to enhance the recognition by avoiding complicated occlusion. The most challenging vision-based dataset is RWTH-PHOENIX-Weather multi-signer as it is considered the first multi-signer dataset without any constraints on the signer. The promising recognition results for this dataset is evolved by the appearance of DLN.

Applications domain of SLR: SLR systems, since, its early days of research and development have found pivotal applications to a wide range of real-life scenarios. The interpreting, gamming and education are the main domains that attract researchers. This section provides a brief overview on the main application domains of SLR systems.

Interpreting is the essential target of the SLR. Virtual sign (Escudeiro *et al*., 2015) is a real-time bidirectional translator of Portuguese Sign Language (PSL). It has three modes of operations: PSL to text, text to PSL and game modes. In PSL to text mode, the PSL is translated to its equivalent text by collecting input data from a Microsoft Kinect de-vice and a pair of data gloves. On the other hand, text to PSL mode acts to convert a Portuguese written text to its equivalent PSL that is performed by an avatar. A serious game for improving the learning of PSL is proposed in the game mode. This game comprises a map with different scattered objects where a character controlled by the user collect these objects. The collected objects represent several gestures from the PSL that is performed by the character (Escudeiro *et al*., 2014). Visualcomm (Chai *et al*., 2013) is a real-time bidirectional translator of Chinese Sign Language (CSL), that has two modes of operation: translation mode and communication mode. Translation mode is desired for grasping isolated Chinese sign words with their corresponding translation. In the communication mode, continuous CSL is translated into text or speech and the text or speech can also be converted to a continuous CSL performed by an avatar. This system relies on the Kinect technology to record the 3D trajectory of the hand and then be used in the recognition.

Gamming is the second domain of SLR. All the available games for deaf are serious games that aim to enhance their capabilities in memorizing and education. CopyCat (Zafrulla *et al*., 2011) is the first interactive educational game to facilitate learning of ASL based on colored gloves with wrist mounted accelerometer with a single firewire camera. The child interacts with the

hero with signing to identify the place of an object or to warn of villain. SMILE (Adamo-Villani and Anasingaraju, 2017) is the first virtual learning game for science and math targeting deaf children. The virtual environment is a fictional city inhabited by 3D phantom characters that the children interact with. The application is designed primarily for display in stationary projection systems where participant learns specific math or science concepts by performing hands-on activities using ASL. The evaluation is based on three assessments: expert panel-based, formative and summative. Kinect-Sign (Gameiro *et al*., 2013) is a game to encourage listeners to learn PSL based on the Kinect technology. The game has two modes: school mode and competition mode. In the school mode, the user is enrolled in short lesson in a classroom-environment. While in the competitive mode, two games are included: quiz game and lingo game. The first is a TV-show styled game in a form of question and the user must sign the correct answer. While the second game is a spelling check game for the words was taught. By Lotfi *et al.* (2015), an interactive web-based serious game for teaching ArSL based on the leap motion controller was presented. The game consists of 2D and 3D game environments. The 2D-level is a letter recognition simple game while the 3D-level is choosing letters in order to form a word of a randomly displayed picture. By Chebka and Essalmi (2015), an ArSL crosswords game for deaf is proposed. This game has two interfaces: learner's interface and teacher's interface. The learners interface let the child to play the game, consult the words and related images, then see its corresponding video. On the other hand, the teachers interface allows the teacher to control the displayed words and the availability of adding new words to the database. The system is evaluated by testing the knowledge of 38 children before and after playing the game. By Uluer *et al.* (2015), humanoid robot-based interactive game aids in teaching of Turkish Sign Language (TSL) by producing and recognizing TSL using an RGB-D camera. The robot in its humanoid form provides the children with suitable environment for communication. And not only afford assistant to deaf children but also to those with either partial or no deafness who want to learn the TSL. Mathsigner (Adamo-Villani and Wilbur, 2010) is a 3D animation interactive software package designed for ASL. It targets K-6 children to teach them math concept, signs and corresponding english terminology by introducing sets of activities with implementation guidelines that are additionally helpful for parents and teachers.

In the field of education, tutors for the SL facilitates the learning and the participation of the deaf children in a communication. Holographic signing avatars (Adamo-Villani and Anasingaraju, 2017) aids in real-time interpreting of spoken English to ASL to help deaf children in learning math. The signing avatar is displayed as a 3D hologram and viewed through the AR glasses worn by the child. Mobile-based Filipino Sign Language (FSL) tutor is illustrated by Garcia *et al.* (2016). It has three modes of operations: dictionary, challenge and help modes. The dictionary holds 50 FSL signs listed in order whenever a sign name is hit, its corresponding sign video is displayed. The challenge mode comprises of four levelfor assessing. The main goal of the assessment is to test the ability of the child to remember the meaning of a sign or group of signs.

CONCLUSION

The field of automatic sign language recognition is categorized under the area of gesture recognition as the dominant component of the sign language is the hand motion. The state-of-the-art CSLR systems have proposed different techniques which only remain functional in controlled-lab environment. A minor variation in these conditions leads to a deadlock to the entire framework. The lack of a practical CSLR system indicates that the operating conditions are not sufficient to deal with the specifications of a practical environment.

As mentioned before there are nearly four challenges that affect the CSLR field: dynamic hand recognition and tracking, facial expression recognition, movement epenthesis detection and the lack of large-scale benchmark databases. Defeating these challenges can be the scope of the future work in this field. For dynamic hand recognition and tracking, the accuracy of EMG-based devices as they are promising in the field of sign and gesture recognition can be enhanced by engaging them with depth-based methods to achieve high recognition rates by making use of the pros of both methods.

The unsupervised modeling of the movement epenthesis can be improved to decrease the time needed for recognition. Work on providing a large-scale benchmark database; to benefit from the elegant technology of deep learning of the neural network. The grammar structure of the sign language differs from that of spoken language, so, up-coming researches must give more attention to model this grammar to enhance the result of the recognition process. All available systems are designed for being used by single user at a time. Generalization to multi-signer systems will require the innovation of a fast-real-time feature analysis and recognition algorithms.

REFERENCES

Abreu, J.G., J.M. Teixeira, L.S. Figueiredo and V. Teichrieb, 2016. Evaluating sign language recognition using the Myo armband. Proceedings of the 2016 18th International Symposium on Virtual and Augmented Reality (SVR), June 21-24, 2016, IEEE, Gramado, Brazil, ISBN:978-1-5090-4149-7, pp: 64-70.

Adamo-Villani, N. and R. Wilbur, 2010. Software for Math and Science education for the deaf. Disability Rehabil. Assistive Technol., 5: 115-124.

Adamo-Villani, N. and S. Anasingaraju, 2017. Holographic signing avatars for deaf education. Proceedings of the 3rd International Conference on E-Learning, E-Education and Online Training (eLEOT'16), August 31-September 2, 2016, Dublin, Ireland, pp: 54-61.

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

Aldoukhi, A.H., H. Law, K.M. Black, W.W. Roberts, J. Deng and K.R. Ghani, 2019. PD04-06 Deep learning computer vision algorithm for detecting kidney stone composition: Towards an automated future. J. Urology, 201: e75-e76.

Alfonse, M., A. Ali, A.S. Elons, N.L. Badr and M. Aboul-Ela, 2015. Arabic sign language benchmark database for different heterogeneous sensors. Proceedings of the 2015 5th International Conference on Information & Communication Technology and Accessibility (ICTA'15), December 21-23, 2015, IEEE, Marrakech, Morocco, pp: 1-9.

Aly, S., A.L. Abbott and M. Torki, 2016. A multi-modal feature fusion framework for kinect-based facial expression recognition using Dual Kernel Discriminant Analysis (DKDA). Proceedings of the 2016 IEEE Winter Conference on Applications of Computer Vision (WACV'16), March 7-10, 2016, IEEE, Lake Placid, New York, USA., pp: 1-10.

Amin, O., H. Said, A. Samy and H.K. Mohammed, 2015. HMM based automatic Arabic sign language translator using Kinect. Proceedings of the 2015 10th International Conference on Computer Engineering & Systems (ICCES'15), December 23-24, 2015, IEEE, Cairo, Egypt, pp: 389-392.

Anil, J. and L.P. Suresh, 2016. Literature survey on face and face expression recognition. Proceedings of the 2016 International Conference on Circuit, Power and Computing Technologies (ICCPCT'16), March 18-19, 2016, IEEE, Nagercoil, India, pp: 1-6.

Assaleh, K., T. Shanableh, M. Fanaswala, F. Amin and H. Bajaj, 2010. Continuous Arabic sign language recognition in user dependent mode. J. Intell. Learn. Syst. Appl., 2: 19-27.

Bacchi, S., L. Oakden-Rayner, T. Zerner, T. Kleinig, S. Patel and J. Jannes, 2019. Deep learning natural language processing successfully predicts the cerebrovascular cause of transient ischemic attack-like presentations. Stroke, 50: 758-760.

Baranwal, N., K. Tripathi and G.C. Nandi, 2015. Possibility theory based continuous Indian Sign Language gesture recognition. Proceedings of the TENCON 2015 IEEE Region 10 Conference, November 1-4, 2015, IEEE, Macao, China, pp: 1-5.

Bhuyan, M.K., D.A. Kumar, K.F. MacDorman and Y. Iwahori, 2014. A novel set of features for continuous hand gesture recognition. J. Multimodal User Interfaces, 8: 333-343.

Brancati, N., G.D. Pietro, M. Frucci and L. Gallo, 2017. Human skin detection through correlation rules between the YCb and YCr subspaces based on dynamic color clustering. Comput. Vision Image Understanding, 155: 33-42.

Brand, M., N. Oliver and A. Pentland, 1997. Coupled hidden Markov models for complex action recognition. Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, June 17-19, 1997, IEEE, San Juan, Puerto Rico, USA., pp: 994-999.

Camgoz, N.C., S. Hadfield, O. Koller and R. Bowden, 2017. Subunets: End-to-end hand shape and continuous sign language recognition. Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV'17), October 22-29, 2017, IEEE, Venice, Italy, pp: 3075-3084.

Caplier, A., L. Bonnaud, S. Malassiotis and M.G. Strintzis, 2004. Comparison of 2D and 3D analysis for automated cued speech gesture recognition. Proceedings of the 9th International Conference on Speech and Computer (SPECOM'04), September 20-22, 2004, Saint-Petersburg, Russia, pp: 35-41.

Chai, X., G. Li, X. Chen, M. Zhou, G. Wu and H. Li, 2013. Visualcomm: A tool to support communication between deaf and hearing persons with the kinect. Proceedings of the 15th International ACM International SIGACCESS Conference on Computers and Accessibility (ASSETS'13), October 21-23, 2013, ACM, Bellevue, Washington, USA., pp: 1-2.

Chai, X., Z. Liu, F. Yin, Z. Liu and X. Chen, 2016. Two streams recurrent neural networks for large-scale continuous gesture recognition. Proceedings of the 2016 23rd International Conference on Pattern Recognition (ICPR'16), December 4-8, 2016, IEEE, Cancun, Mexico, pp: 31-36.

Chebka, R. and F. Essalmi, 2015. A crosswords game for deaf. Proceedings of the 2015 5th International Conference on Information & Communication Technology and Accessibility (ICTA'15), December 21-23, 2015, IEEE, Marrakech, Morocco, pp:1-6.

Chen, L., H. Wei and J. Ferryman, 2013. A survey of human motion analysis using depth imagery. Pattern Recognit. Lett., 34: 1995-2006.

Cheng, H., L. Yang and Z. Liu, 2015a. Survey on 3D hand gesture recognition. IEEE. Trans. Circuits Syst. Video Technol., 26: 1659-1673.

Cheng, J., X. Chen, A. Liu and H. Peng, 2015b. A novel phonology-and radical-coded Chinese sign language recognition framework using accelerometer and surface electromyography sensors. Sensors, 15: 23303-23324.

Cheok, M.J., Z. Omar and M.H. Jaward, 2019. A review of hand gesture and sign language recognition techniques. Int. J. Mach. Learn. Cybern., 10: 131-153.

Chossat, J.B., Y. Tao, V. Duchaine and Y.L. Park, 2015. Wearable soft artificial skin for hand motion detection with embedded microfluidic strain sensing. Proceedings of the 2015 IEEE International Conference on Robotics and Automation (ICRA'15), May 26-30, 2015, IEEE, Seattle, Washington, USA., pp: 2568-2573.

Chuan, C.H., E. Regina and C. Guardino, 2014. American sign language recognition using leap motion sensor. Proceedings of the 2014 13th International Conference on Machine Learning and Applications, December 3-6, 2014, IEEE, Detroit, Michigan, USA., pp: 541-544.

Corneanu, C.A., M.O. Simon, J.F. Cohn and S.E. Guerrero, 2016. Survey on RGB, 3D, thermal and multimodal approaches for facial expression recognition: History, trends and affect-related applications. IEEE. Trans. Pattern Anal. Mach. Intell., 38: 1548-1568.

Cui, R., H. Liu and C. Zhang, 2017. Recurrent convolutional neural networks for continuous sign language recognition by staged optimization. Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, July 21-26, 2017, Honolulu, Hawaii, USA., pp: 7361-7369.

Dreuw, P., C. Neidle, V. Athitsos, S. Sclaroff and H. Ney, 2008a. Benchmark databases for video-based automatic sign language recognition. Proceedings of the International Conference on Language Resources and Evaluation (LREC'08), May 26-June 1, 2008, Marrakech, Morocco, pp: 1-7.

Dreuw, P., J. Forster, T. Deselaers and H. Ney, 2008b. Efficient approximations to model-based joint tracking and recognition of continuous sign language. Proceedings of the 2008 8th IEEE International Conference on Automatic Face & Gesture Recognition, September 17-19, 2008, IEEE, Amsterdam, Netherlands, pp: 1-6.

Ekman, P. and V.W. Friesen, 1978. Facial Action Coding System. Psychologist Press Inc., San Francisco, California.

Escudeiro, P., N. Escudeiro, R. Reis, J. Lopes and M. Norberto *et al*., 2015. Virtual sign-A real time bidirectional translator of Portuguese sign language. Procedia Comput. Sci., 67: 252-262.

Escudeiro, P., N. Escudeiro, R. Reis, M. Barbosa and J. Bidarra *et al*., 2014. Virtual sign game learning sign language. Proceedings of the 5th International Conference on Computers and Technology in Modern Education, April 23-25, 2014, Renaissance Kuala Lumpur Hotel, Kuala Lumpur, Malaysia, pp: 26-33.

Forster, J., C. Schmidt, T. Hoyoux, O. Koller, U. Zelle, J.H. Piater and H. Ney, 2012. RWTH-PHOENIX-Weather: A Large Vocabulary Sign Language Recognition and Translation Corpus. Proceedings of the International Conference on Language Resources and Evaluation (LREC'12), May 2012, Istanbul, Turkey, pp: 3785-3789.

Forster, J., O. Koller, C. Oberdorfer, Y. Gweth and H. Ney, 2013. Improving continuous sign language recognition: Speech recognition techniques and system design. Proceedings of the 4th International Workshop on Speech and Language Processing for Assistive Technologies, August 2013, Grenoble, France, pp: 41-46.

Gameiro, J., T. Cardoso and Y. Rybarczyk, 2013. Kinect-sign: Teaching sign language to listeners through a game. Procedia Technol., 17: 384-391.

Gao, W., G. Fang, D. Zhao and Y. Chen, 2004. Transition movement models for large vocabulary continuous sign language recognition. Proceedings of the 6th IEEE International Conference on Automatic Face and Gesture Recognition, May 19, 2004, IEEE, Seoul, South Korea, pp: 553-558.

Garcia, M.G., C.I.S. Luis and M.J.C. Samonte, 2016. E-tutor for Filipino sign language. Proceedings of the 2016 11th International Conference on Computer Science & Education (ICCSE'16), August 23-25, 2016, IEEE, Nagoya, Japan, pp: 223-227.

Ghotkar, A.S. and G.K. Kharate, 2015. Dynamic hand gesture recognition and novel sentence interpretation algorithm for Indian sign language using microsoft kinect sensor. J. Pattern Recognit. Res., 1: 24-38.

Goodfellow, I., Y. Bengio and A. Courville, 2016. Deep Learning. MIT Press, Massachusetts, United States, ISBN:9780262337373, Pages: 800.

Graves, A., S. FernAndez, F. Gomez and J. Schmidhuber, 2006. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. Proceedings of the 23rd International Conference on Machine Learning (ICML'06), June 25-29, 2006, ACM, Pittsburgh, Pennsylvania, USA., pp: 369-376.

Gweth, Y.L., C. Plahl and H. Ney, 2012. Enhanced continuous sign language recognition using PCA and neural network features. Proceedings of the 2012 IEEE International Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW'12), June 16-21, 2012, IEEE, Providence, Rhode Island, USA., pp: 55-60.

Hochreiter, S. and J. Schmidhuber, 1997. Long short-term memory. Neural Comput., 9: 1735-1780.

Holt, G.A.T., A.J.V. Doorn, M.J.T. Reinders, E.A. Hendriks and H.D. Ridder, 2011. Human-inspired search for redundancy in automatic sign language recognition. ACM. Trans. Applied Percept. (TAP), Vol. 8, No. 2. 10.1145/1870076.1870083

Hoshino, K., 2006. Dexterous robot hand control with data glove by human imitation. IEICE. Trans. Inf. Syst., 89: 1820-1825.

Huang, C.L. and B.L. Tsai, 2010. A vision-based Taiwanese sign language recognition. Proceedings of the 2010 20th International Conference on Pattern Recognition, August 23-26, 2010, IEEE, Istanbul, Turkey, pp: 3683-3686.

Huang, J., W. Zhou, Q. Zhang, H. Li and W. Li, 2018. Video-based sign language recognition without temporal segmentation. Proceedings of the 32nd AAAI International Conference on Artificial Intelligence (AAAI'18), February 2-7, 2018, New Orleans, Louisiana, USA., pp: 1-8.

Hyeon-Kyu, L. and J.H. Kim, 1999. An HMM-based threshold model approach for gesture recognition. IEEE Trans. Pattern Anal. Mach. Intell. (PAMI), 21: 961-973.

Kakde, M.U., M.G. Nakrani and A.M. Rawate, 2016. A review paper on sign language recognition system for deaf and dumb people using image processing. Int. J. Eng. Res. Technol. (IJERT), 5: 590-592.

Kelly, D., J. McDonald and C. Markham, 2009a. Recognizing spatiotemporal gestures and movement epenthesis in sign language. Proceedings of the 2009 13th International Conference on Machine Vision and Image Processing, September 2-4, 2009, IEEE, Dublin, Ireland, pp: 145-150.

Kelly, D., J. McDonald and C. Markham, 2010. A person independent system for recognition of hand postures used in sign language. Pattern Recog. Lett., 31: 1359-1368.

Kelly, D., J. McDonald and C. Markham, 2011. Recognition of Spatiotemporal Gestures in Sign Language Using Gesture Threshold Hmms. In: Machine Learning for Vision-Based Motion Analysis, Wang L., G. Zhao, L. Cheng and M. Pietikainen (Eds.). Springer, London, UK., ISBN: 978-0-85729-056-4, pp: 307-348.

Kelly, D., J.R. Delannoy, J. McDonald and C. Markham, 2009b. A framework for continuous multimodal sign language recognition. Proceedings of the 2009 International Conference on Multimodal Interfaces (ICMI-MLMI'09), November 02-04, 2009, ACM, Cambridge, Massachusetts, USA., pp: 351-358.

Khan, S., D.G. Bailey and G.S. Gupta, 2014. Pause detection in continuous sign language. Int. J. Comput. Appl. Technol., 50: 75-83.

Kishore, P.V.V., M.V.D. Prasad, D.A. Kumar and A.S.C.S. Sastry, 2016. Optical flow hand tracking and active contour hand shape features for continuous sign language recognition with artificial neural networks. Proceedings of the 2016 IEEE 6th International Conference on Advanced Computing (IACC'16), February 27-28, 2016, IEEE, Bhimavaram, India, pp: 346-351.

Klakow, D. and J. Peters, 2002. Testing the correlation of word error rate and perplexity. Speech Commun., 38: 19-28.

Koller, O., C. Camgoz, H. Ney and R. Bowden, 2019. Weakly supervised learning with multi-stream CNN-LSTM-HMMs to discover sequential Parallelism in sign language videos. IEEE. Trans. Pattern Anal. Mach. Intell., Vol. 2019, 10.1109/TPAMI.2019.2911077

Koller, O., J. Forster and H. Ney, 2015. Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers. Comput. Vision Image Understanding, 141: 108-125.

Koller, O., O. Zargaran, H. Ney and R. Bowden, 2016. Deep sign: Hybrid CNN-HMM for continuous sign language recognition. Proceedings of the British Conference on Machine Vision (BMVC'16), September 19-22, 2016, New York, USA., pp: 1-12.

Koller, O., S. Zargaran and H. Ney, 2017. Re-sign: Re-aligned end-to-end sequence modelling with deep recurrent CNN-HMMs. Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, IEEE, Honolulu, Hawaii, pp: 4297-4305.

Kong, W.W. and S. Ranganath, 2014. Towards subject independent continuous sign language recognition: A segment and merge approach. Pattern Recognit., 47: 1294-1308.

Li, K., Z. Zhou and C.H. Lee, 2016. Sign transition modeling and a scalable solution to continuous sign language recognition for real-world applications. ACM. Trans. Accessible Comput. (TACCESS), Vol. 8, No.2. 10.1145/2850421

Li, Y., X. Chen, J. Tian, X. Zhang, K. Wang and J. Yang, 2010. Automatic recognition of sign language subwords based on portable accelerometer and EMG sensors. Proceedings of the International Joint Conference and Workshop on Multimodal Interfaces and the Machine Learning for Multimodal Interaction (ICMI-MLMI'10), November 08-10, 2010, ACM, Beijing, China, pp: 1-17.

Li, Y., X. Chen, X. Zhang, K. Wang and J. Yang, 2011. Interpreting sign components from accelerometer and sEMG data for automatic sign language recognition. Proceedings of the 2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, August 30-September 3, 2011, IEEE, Boston, Massachusetts, USA., pp: 3358-3361.

Li, Y., X. Chen, X. Zhang, K. Wang and Z.J. Wang, 2012. A sign-component-based framework for Chinese sign language recognition using accelerometer and sEMG data. IEEE. Trans. Biomed. Eng., 59: 2695-2704.

Lichtenaur, J.F., E.A. Hendriks and M.J.T. Reinders, 2008. Sign language recognition by combining statistical DTW and independent classification. IEEE. trans. pattern anal. machine intell., 30: 2040-2046.

Lotfi, E., B. Amine and B. Mohammed, 2015. Teaching Arabic sign language through an interactive web based serious game. Int. J. Comput. Appl., 116: 12-18.

Lu, J., V. Behbood, P. Hao, H. Zuo, S. Xue and G. Zhang, 2015. Transfer learning using computational intelligence: A survey. Knowl. Based Syst., 80: 14-23.

Lun, R. and W. Zhao, 2015. A survey of applications and human motion recognition with microsoft kinect. Int.J. Pattern Recognit. Artif. Intell., Vol. 29, No. 05. 10.1142/S0218001415550083

Maraqa, M., F. Al-Zboun, M. Dhyabat and R.A. Zitar, 2012. Recognition of Arabic Sign Language (ArSL) using recurrent neural networks. J. Intell. Learning Syst. Appl., 4: 41-52.

Marin, G., F. Dominio and P. Zanuttigh, 2014. Hand gesture recognition with leap motion and kinect devices. Proceedings of the 2014 IEEE International Conference on Image Processing (ICIP), October 27-30, 2014, IEEE, Paris, France, pp: 1565-1569.

Martinez, A., B. Wilbur, R. Shay and A. Kak, 2002. Purdue RVL-SLLL ASL database for automatic recognition of American Sign Language. Proceedings of the 4th IEEE International Conference on Multimodal Interfaces, October 14-16, 2002, IEEE Computer Society Washington, DC., USA., pp: 167-172.

Martinez, B. and M.F. Valstar, 2016. Advances, Challenges and Opportunities in Automatic Facial Expression Recognition. In: Advances in Face Detection and Facial Image Analysis, Kawulok, M., M.E. Celebi and B. Smolka (Eds.). Springer, Cham, Switzerland, ISBN: 978-3-319-25956-7, pp: 63-100.

Mazumdar, D., M.K. Nayak and A.K. Talukdar, 2015. Adaptive Hand Segmentation and Tracking for Application in Continuous Hand Gesture Recognition. In: Recent Trends in Intelligent and Emerging Systems, Sarma, K.K., M.P. Sarma and M. Sarma (Eds.). Springer, New Delhi, India, ISBN: 978-81-322-2406-8, pp: 115-124.

Michal, K., N. Jakub and K. Jolanta, 2014. Skin Detection and Segmentation in Color Images. In: Advances in Low-Level Color Image Processing, Celebi, M.E. and B. Smolka (Eds.). Springer, Berlin, Germany, ISBN: 978-94-007-7583-1, pp: 329-366.

Mohandes, M., S. Aliyu and M. Deriche, 2014. Arabic sign language recognition using the leap motion controller. Proceedings of the 2014 IEEE 23rd International Symposium on Industrial Electronics (ISIE), June 1-4, 2014, IEEE, Istanbul, Turkey, pp: 960-965.

Mohandes, M., S. Aliyu and M. Deriche, 2015. Prototype Arabic sign language recognition using multi-sensor data fusion of two leap motion controllers. Proceedings of the 2015 IEEE 12th International Multi-Conference on Systems, Signals & Devices (SSD'15), March 16-19, 2015, IEEE, Mahdia, Tunisia, pp: 1-6.

Mohandes, M.A. and S. Buraiky, 2007. Automation of the Arabic sign language recognition using the powerglove. AIML. J., 7: 41-46.

Mohandes, M.A., 2013. Recognition of two-handed Arabic signs using the CyberGlove. Arabian J. Sci. Eng., 38: 669-677.

Molchanov, P., X. Yang, S. Gupta, K. Kim, S. Tyree and J. Kautz, 2016. Online detection and classification of dynamic hand gestures with recurrent 3d convolutional neural networks. Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR'16), June 27-30, 2016, IEEE, Las Vegas, Nevada, USA., pp: 4207-4215.

Najera, L.O.R., M.L. Sanchez, J.G.G. Serna, R.P. Tapia and J.Y.A. Llanes, 2016. Recognition of mexican sign language through the leap motion controller. Proceedings of the International Conference on Scientific Computing and (CSC'16), July 25-28, 2016, Las Vegas, Nevada, USA., pp: 147-151.

Natarajan, P. and R. Nevatia, 2007. Coupled hidden semi markov models for activity recognition. Proceedings of the 2007 IEEE Workshop on Motion and Video Computing (WMVC'07), February 23-24, 2007, IEEE, Austin, Texas, USA., pp: 10-10.

Nayak, S., S. Sarkar and B. Loeding, 2009. Automated extraction of signs from continuous sign language sentences using iterated conditional modes. Proceedings of the 2009 IEEE International Conference on Computer Vision and Pattern Recognition, June 20-25, 2009, IEEE, Miami, Florida, USA., pp: 2583-2590.

Ong, E.J., H. Cooper, N. Pugeault and R. Bowden, 2012. Sign language recognition using sequential pattern trees. Proceedings of the 2012 IEEE International Conference on Computer Vision and Pattern Recognition, June 16-21, 2012, IEEE, Providence, Rhode Island, USA., pp: 2200-2207.

Ong, E.J., O. Koller, N. Pugeault and R. Bowden, 2014. Sign spotting using hierarchical sequential patterns with temporal intervals. Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, June 23-28, 2014, IEEE, Columbus, Ohio, USA., pp: 1923-1930.

Pashaloudi, V.N. and K.G. Margaritis, 2002. Hidden markov model for sign language recognition: A review. Proceedings of the 2nd Hellenic Conference on Artificial Intelligence (SETN'02), April 11-12, 2002, Thessaloniki, Greece, pp: 343-354.

Paudyal, P., A. Banerjee and S.K. Gupta, 2016. Sceptre: A pervasive, non-invasive and programmable gesture recognition technology. Proceedings of the 21st International Conference on Intelligent User Interfaces (IUI'16), March 7-10, 2016, ACM, Sonoma, California, USA., pp: 282-293.

Pigou, L., V.D.A. Oord, S. Dieleman, M. Van Herreweghe and J. Dambre, 2018. Beyond temporal pooling: Recurrence and temporal convolutions for gesture recognition in video. Int. J. Comput. Vision, 126: 430-439.

Pitsikalis, V., S. Theodorakis and P. Maragos, 2010. Data-driven sub-units and modeling structure for continuous sign language recognition with multiple cues. Proceedings of the LREC International Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies (LREC'10), May 17-23, 2010, Valletta, Malta, pp: 196-203.

Rao, G.A. and P.V.V. Kishore, 2018. Selfie video based continuous Indian sign language recognition system. Ain Shams Eng. J., 9: 1929-1939.

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

Raziq, N. and S. Latif, 2016. Pakistan sign language recognition and translation system using leap motion device. Proceedings of the International Conference on P2P, Parallel, Grid, Cloud and Internet Computing (3PGCIC'16), November 5-7, 2016, Springer, Berlin, Germany, pp: 895-902.

Sandbach, G., S. Zafeiriou, M. Pantic and L. Yin, 2012. Static and dynamic 3D facial expression recognition: A comprehensive survey. Image Vision Comput., 30: 683-697.

Sarkar, S., B. Loeding and A.S. Parashar, 2010. Fusion of Manual and Non-Manual Information in American Sign Language Recognition. In: Handbook of Pattern Recognition and Computer Vision, Chen, C.H. (Ed.). World Scientific, Singapore, pp: 477-495.

Sarkar, S., B. Loeding, R. Yang, S. Nayak and A. Parashar, 2011. Segmentation-robust representations, matching and modeling for sign language. Proceedings of the CVPR 2011 International Workshops, June 20-25, 2011, IEEE, Springs, Colorado, USA., pp: 13-19.

Shaik, K.B., P. Ganesan, V. Kalist, B.S. Sathish and J.M.M. Jenitha, 2015. Comparative study of skin color detection and segmentation in HSV and YCbCr color space. Proc. Comput. Sci., 57: 41-48.

Shohieb, S.M., H.K. Elminir and A.M. Riad, 2015. Signsworld atlas: A benchmark Arabic sign language database. J. King Saud Univ. Comput. Inf. Sci., 27: 68-76.

Shukor, A.Z., M.F. Miskon, M.H. Jamaluddin, F. bin Ali, M.F. Asyraf and M.B. bin Bahar, 2015. A new data glove approach for Malaysian sign language detection. Procedia Comput. Sci., 76: 60-67.

Simao, M.A., P. Neto and O. Gibaru, 2016. Unsupervised gesture segmentation by motion detection of a real-time data stream. IEEE. Trans. Ind. Inf., 13: 473-481.

Simos, M. and N. Nikolaidis, 2016. Greek sign language alphabet recognition using the leap motion device. Proceedings of the 9th Hellenic Conference on Artificial Intelligence (SETN'16), May 18-20, 2016, ACM, Thessaloniki, Greece, pp: 1-4.

Starner, T. and A. Pentland, 1997. Real-Time American Sign Language Recognition from Video Using Hidden Markov Models. In: Motion-Based Recognition, Shah M. and R. Jain (Eds.). Springer, Dordrecht, Netherlands, ISBN: 978-90-481-4870-7, pp: 227.

Stokoe Jr, W.C., 2005. Sign language structure: An outline of the visual communication systems of the American deaf. J. Deaf Stud. Deaf Educ., 10: 3-37.

Stokoe, W.C., D.C. Casterline and C.G. Croneberg, 1976. A Dictionary of American Sign Language on Linguistic Principles. 2nd Edn., Linstok Press, USA., ISBN: 9780932130013, Pages: 346.

Sun, C., T. Zhang and C. Xu, 2015. Latent support vector machine modeling for sign language recognition with Kinect. ACM. Trans. Intell. Syst. Technol. (TIST), Vol. 6, No. 2. 10.1145/2629481

Sun, C., T. Zhang, B.K. Bao and C. Xu, 2013. Latent support vector machine for sign language recognition with Kinect. Proceedings of the 2013 IEEE International Conference on Image Processing, September 15-18, 2013, IEEE, Melbourne, Australia.,-pp: 4190.

Sutarman, M.B.A. Majid and J.B.M. Zain, 2013. Vision-based sign language recognition systems: A review. Proceedings of the 2013 International Conference on Computer Science and Information Technology (CSIT'13), June 16-19, 2013, Yogyakarta, Indonesia, pp: 195-200.

Theodorakis, S., V. Pitsikalis and P. Maragos, 2014. Dynamic-static unsupervised sequentiality, statistical subunits and lexicon for sign language recognition. Image Vision Comput., 32: 533-549.

Theodorakis, S., V. Pitsikalis, I. Rodomagoulakis and P. Maragos, 2012. Recognitionwith raw canonical phonetic movement and handshape subunits on videos of continuous sign language. Proceedings of the 2012 19th IEEE International Conference on Image Processing, September 30- October 3, 2012, IEEE, Orlando, Florida, USA., pp: 1413-1416.

Tolba, M.F., A. Samir and M. Aboul-Ela, 2013. Arabic sign language continuous sentences recognition using PCNN and graph matching. Neural Comput. Appl., 23: 999-1010.

Tran, D., L. Bourdev, R. Fergus, L. Torresani and M. Paluri, 2015. Learning spatiotemporal features with 3D convolutional networks. Proceedings of the IEEE International Conference on Computer Vision (ICCV'15), December 7-13, 2015, IEEE, Santiago, Chile, pp: 4489-4497.

Tripathi, K. and N.B.G. Nandi, 2015. Continuous Indian sign language gesture recognition and sentence formation. Procedia Comput. Sci., 54: 523-531.

Tu, Y.H., J. Du, L. Sun, F. Ma, H.K. Wang, J.D. Chen and C.H. Lee, 2019. An iterative mask estimation approach to deep learning based multi-channel speech recognition. Speech Commun., 106: 31-43.

Tubaiz, N., T. Shanableh and K. Assaleh, 2015. Glove-based continuous Arabic sign language recognition in user-dependent mode. IEEE. Trans. Hum. Mach. Syst., 45: 526-533.

Uluer, P., N. Akalın and H. Kose, 2015. A new robotic platform for sign language tutoring. Int. J. Soc. Rob., 7: 571-585.

Viola, P. and M. Jones, 2001. Rapid object detection using a boosted cascade of simple features. Proceedings of the Computer Society Conference on Computer Vision and Pattern Recognition, Volume 1, December 8-14, 2001, Kauai, HI., USA., pp: 511-518.

Vogler, C. and D. Metaxas, 1997. Adapting hidden Markov models for ASL recognition by using three-dimensional computer vision methods. Proceedings of the 1997 IEEE International Conference on Systems, Man, and Cybernetics Computational Cybernetics and Simulation Vol. 1, October 12-15, 1997, IEEE, Orlando, Florida, USA., pp: 156-161.

Vogler, C. and D. Metaxas, 1999. Parallel hidden Markov models for American sign language recognition. Proceedings of the 7th IEEE International Conference on Computer Vision Vol. 1, September 20-27, 1999, IEEE, Kerkyra, Greece, pp: 116-122.

WHO., 2018. Center media deafness and hearing loss. World Health Organization, Geneva, Switzerland.

Wan, J., Y. Zhao, S. Zhou, I. Guyon, S. Escalera and S.Z. Li, 2016. Chalearn looking at people RGB-D isolated and continuous datasets for gesture recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, June 26-July 1, 2016, IEEE, Las Vegas, Nevada, USA., pp: 56-64.

Wang, H., P. Wang, Z. Song and W. Li, 2017. Large-scale multimodal gesture segmentation and recognition based on convolutional neural networks. Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCVW), October 22-29, 2017, IEEE, Venice, Italy, pp: 3138-3146.

Wang, R.Y. and J. Popovic, 2009. Real-time hand-tracking with a color glove. ACM. Trans. Graphics, 28: 1-63.

Wilbur, R. and A.C. Kak, 2006. Purdue RVL-SLLL American sign language database. Report, TR-06-12, Purdue University, West Lafayette, Indiana. https://engineering.purdue.edu/RVL/Database/ASL/asl-database-front.htm.bk.htm

Yan, J., Z. Lei, L. Wen and S.Z. Li, 2014. The fastest deformable part model for object detection. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, June 23-28, 2014, IEEE, Columbus, Ohio, USA., pp: 2497-2504.

Yang, H.D. and S.W. Lee, 2010. Robust sign language recognition with hierarchical conditional random fields. Proceedings of the 2010 20th International Conference on Pattern Recognition, August 23-26, 2010, IEEE, Istanbul, Turkey, pp: 2202-2205.

Yang, H.D. and S.W. Lee, 2011. Combination of manual and non-manual features for sign language recognition based on conditional random field and active appearance model. Proceedings of the 2011 International Conference on Machine Learning and Cybernetics (ICMLC'11) Vol. 4, July 10-13, 2011, IEEE, Guilin, China, pp: 1726-1731.

Yang, H.D. and S.W. Lee, 2013. Robust sign language recognition by combining manual and non-manual features based on conditional random field and support vector machine. Pattern Recognit. Lett., 34: 2051-2056.

Yang, R. and S. Sarkar, 2006. Detecting coarticulation in sign language using conditional random fields. Proceedings of the 18th International Conference on Pattern Recognition (ICPR'06), August 20-24, 2006, IEEE, Hong Kong, China, pp: 108-112.

Yang, R., S. Sarkar and B. Loeding, 2007. Enhanced level building algorithm for the movement epenthesis problem in sign language recognition. Proceedings of the 2007 IEEE International Conference on Computer Vision and Pattern Recognition, June 17-22, 2007, IEEE, Minneapolis, Minnesota, pp: 1-8.

Yang, W., J. Tao and Z. Ye, 2016. Continuous sign language recognition using level building based on fast hidden Markov model. Pattern Recognit. Lett., 78: 28-35.

Yu, S.H., C.L. Huang, S.C. Hsu, H.W. Lin and H.W. Wang, 2011. Vision-based continuous sign language recognition using product HMM. Proceedings of the 1st Asian Conference on Pattern Recognition, November 28, 2011, IEEE, Beijing, China, pp: 510-514.

Zadghorban, M. and M. Nahvi, 2018. An algorithm on sign words extraction and recognition of continuous Persian sign language based on motion and shape features of hands. Pattern Anal. Appl., 21: 323-335.

Zafeiriou, S., C. Zhang and Z. Zhang, 2015. A survey on face detection in the wild: past, present and future. Comput. Vision Image Understanding, 138: 1-24.

Zafrulla, Z., H. Brashear, P. Presti, H. Hamilton and T. Starner, 2011. CopyCat: An American sign language game for deaf children. Proceedings of the Face and Gesture 2011, March 21-25, 2011, IEEE, Santa Barbara, California, pp: 647-647.

Zhang, J., W. Zhou and H. Li, 2014. A threshold-based hmm-dtw approach for continuous sign language recognition. Proceedings of the International Conference on Internet Multimedia Computing and Service (ICIMCS'14), July 10-12, 2014, ACM, Xiamen, China, pp: 237-240.

Zhang, J., W. Zhou and H. Li, 2015. A new system for Chinese sign language recognition. Proceedings of the 2015 IEEE China Summit and International Conference on Signal and Information Processing (ChinaSIP), July 12-15, 2015, IEEE, Chengdu, China, pp: 534-538.