IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received June 15, 2019, accepted July 11, 2019, date of publication July 16, 2019, date of current version August 5, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2929174

Technical Approaches to Chinese Sign Language Processing: A Review

SUHAIL MUHAMMAD KAMAL[1,2,3], YIDONG CHEN[1,2], SHAOZI LI[1,2], (Senior Member, IEEE), XIAODONG SHI[1,2], AND JIANGBIN ZHENG[1,2]

[1]School of Informatics, Xiamen University, Xiamen 361005, China
[2]Xiamen Key Laboratory of Language and Culture Computation, Xiamen University, Xiamen 361005, China
[3]Department of Information Technology, Faculty of Computer Science and Information Technology, Bayero University, Kano 3011, Nigeria

Corresponding author: Yidong Chen (ydchen@xmu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61573294, in part by the National Social Science Foundation of China under Grant 16AZD049, in part by the Outstanding Achievement Late Fund of the State Language Commission of China under Grant WT135-38, and in part by the Fundamental Research Funds for the Central Universities under Grant 20720181002.

ABSTRACT As with the huge number of deaf-mute people in China is of concern, there is a growing need to integrate them into mainstream society through the use of efficient sign language processing technologies. Sign language processing entails the systematic recognition and translation of sign language images/videos to text or speech. This survey provides an overview of the most important work on Chinese sign language recognition and translation, discussed its classification, highlights the features explored in sign language recognition research, presents the datasets available, and provides trends for the future research.

INDEX TERMS Sign language, Chinese sign language recognition, deep learning, CSL datasets, vision-based, sensor-based, review.

I. INTRODUCTION

Sign language, as important as spoken language, is a visual gesture language among the deaf and hearing-impaired people. It is performed by the movement of the hands together with facial expressions in accordance with its grammatical rules. Each sign language has many thousands of signs, each differing from the next by minor changes in hand shape, motion, position and non-manual features (facial component movements such as eyebrow expressions). Moreover, the deaf and hearing-impaired people, with an estimated population of 20.54 million in China alone, are finding it difficult to be incorporated into the socioeconomic community, as they are in isolation due to their low status and literacy level [1]. In essence, Sign Language Recognition systems are being developed to alleviate this problem.

Sign Language Recognition (SLR) aims to develop algorithms and methods to correctly identify a sequence of produced signs and provides its meaning in the form of text or speech. There are two main categories of the SLR system: vision-based and sensor-based. Vision-based systems utilize images and videos of signs acquired from cameras. On the other hand, sensor-based systems require the use of wearable devices equipped with sensors to extract the hand shapes and motion of the signs.

Several SLR systems aimed at translating signs to texts have been developed to bridge the communication gap between the deaf and hearing people for various sign languages. Currently, there exist some review papers [2]–[7] that discusses the mainstream and state-of-the-art in general Sign Language Recognition system in the past few years with little emphasis on Chinese Sign Language (CSL). For instance, [6] highlighted some problems and challenges faced when building a sign language recognition system. Reference [5] provided an overview of sign capturing methods and classification techniques, problems with continuous SLR and, finally, they introduced Malaysian sign language. Reference [3] surveyed on vision-based and sensor-based approaches in Arabic Sign Language Recognition. [4] presented the existing recognition techniques in vision-based SLR. Reference [7] discussed the various techniques for hand segmentation and tracking, feature extraction, and classification of sign languages. They surveyed mostly traditional image processing approaches and focused more on Indian

The associate editor coordinating the review of this manuscript and approving it for publication was Changsheng Li.

Sign Language. Reference [8] discusses the feature extraction and classification methods of SLR, and also elaborated on challenges of SLR. Reference [9] surveyed the deep learning methodologies for SLR. They further discussed the available dataset for SLR in general.

This paper presents a technical overview of the current state of research in Chinese Sign Language Processing. We surveyed existing literature and provided the technical insights into the approaches in designing a Chinese Sign Language Recognition system (CSLR), and outlined the various modalities of the system. In addition, we further discussed the existing CSL dataset, respectively. We believe that this survey will be useful to researchers who are new to the field and inform themselves with state of the art in CSL.

The rest of the paper is organized as follows: Types of Chinese Sign Language Recognition systems are discussed in Section II. In Section III, Aspects of CSLR systems are elaborated. Datasets for CSL and other sign languages is presented in section IV. Finally, conclusions and trends for future research are given in Section V.

II. SIGN LANGUAGE RECOGNITION (SLR) SYSTEMS

An SLR system can be classified into three based on the type of SL it translates [3], i.e., Fingerspelling, Isolated word, and continuous sign sentence. Each of the categories can be realized using either vision- or sensor-based devices, as shown below.

A. FINGERSPELLING RECOGNITION

Fingerspelling is used in situations where new words, names of persons and places or words with no known signs are spelled in the air by hand movement. It typically consists of 30 hand shapes of CSL alphabets [10], as shown in figure 1. In the earliest work from [10], they used the data obtained from a Cyberglove equipped with a Polhemus tracker, as input to a simple 3-layer feedforward neural network to recognize the alphabets. SVMs were also used for the recognition [11]. More recently, [12] utilized a Leap motion controller to capture 3D coordinates from signers fingers for the alphabet recognition. Reference [13] proposed a vision-based alphabet sign language recognition model in a complex background scene. The model was designed to be adaptable to varying changes in skin color from different users and lighting conditions. Four kinds of feature descriptors, including bag-of-words, Hu moments, Fourier descriptors, and histogram of oriented gradients (HOG) are joined to describe the contours and the salient points of hand gestures. SVM is used in the classification stage to classify the signed alphabets. During recognition, an accuracy of 100% was obtained on their own constructed CSL dataset of 39000 images comprising 26 alphabets and 94% on the publicly available ASL dataset. Fingerspelling recognition can be considered as a basic image classification problem. Therefore, high accuracies are usually obtained because it only involves recognition of alphabet signs from static images. It has minimal focus among researchers than isolated word and continuous sentence recognition.

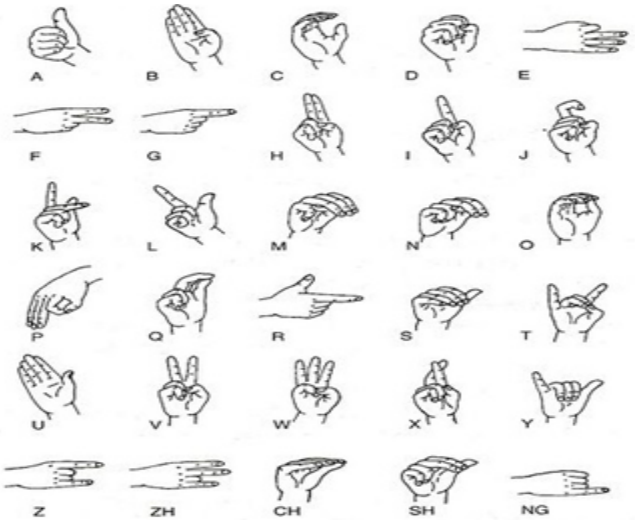

FIGURE 1. Standard Chinese sign language alphabets.

B. ISOLATED WORD RECOGNITION

In this category, isolated word recognition analyzes a sequence of images or signals of hand movements from sensor gloves representing a whole sign. Using a single type of sensor, [14] modeled a signer-independent Chinese Sign Language Recognition using SOFM, HMM, and Neural Network. It works both for isolated and continuous Signs. A recognition rate of 90.5% was observed on their dataset acquired by cybergloves. Reference [15] also presented a continuous Chinese sign language recognition using two CyberGloves and three Polhemus trackers. They utilized transition-movement models to solve the problem of transition parts between two adjacent signs that direct HMM and context-dependent model cannot deal with. An average accuracy of 91.9% was achieved over their large vocabulary of 5113 CSL signs.

In another work, [16] introduced an SLR using 3D CNNs. They applied a 3D CNN to extract the spatiotemporal features from the Kinect Video of signs and an MLP for classification. The architecture integrates multiple data as inputs; RGB, depth, and skeleton trajectory. After experimentation on their self-produced dataset of 675 words - 25 vocabularies with each word played by nine signers for three times - an accuracy of 94.2 % was achieved which is higher than the conventional GMM-HMM baseline model. Furthermore, [17] presented a Fast Sign Language Recognition Benefited from Low-Rank Approximation. They used the low-rank approximation to obtain the keyframes from the sign video and discard the redundant ones. Compared to traditional HMM, the number of hidden states is determined by the low-rank approximation which makes it a ''light-HMM''. In the experiments, they combined the HOG posture appearance features from RGB-D data and skeleton pair features from the skeleton data as

input to the light-HMM. An encouraging result for signer dependent and independent tests on 1000 vocabulary CSL words is obtained, which is faster and higher than HMM. In a related work, [18] developed an SLR system with LSTM. They used skeleton joint trajectories without color and depth information of the Kinect. Four skeleton joints (left and right hands, left and right elbow) were used as input to the LSTM neural network. For evaluation, they created two Chinese isolated datasets; dataset I with 100 daily words and; dataset II with 500 sign words. During testing, the performance of the system is compared with HMM, and a result of 86%, 63% is recorded for the datasets I and II.

In a similar work, [19] introduced an isolated word CSL recognition system with Grassmann covariance matrices (GCM). In this work, they fused the RGB-D and skeleton features of the Kinect video stream to a covariance matrix which encodes the long-term dynamics for sign representation. They selected the most significant singular vectors from a covariance matrix and further represented it as a compact GCM. The SVM classifier is used for sign classification. The method is tested on signer-dependent, and signer-independent Kinect datasets and good accuracy are reported.

Similarly, [20] developed a Chinese SLR with adaptive HMM. In their work, it contains 3 stages: Feature extraction; in which they introduced an enhanced context feature that captures the spatiotemporal features of the trajectory data. In the appearance based features they adopted HOG with PCA features; in the modeling stage, they employed adaptive HMM to determine the states based on the variation of the hand shapes; in classification, they proposed a combination of the probabilities of the trajectories and hand shapes. The method performs better than baselines in their CSL words Kinect-based dataset. Reference [21] also presented a Chinese SLR based on Hand shape and trajectory features. They used Histogram of Oriented Displacement (HOD) and Relative Distance Features (RDF) to describe the trajectory information of the sign words. They also adopt Histogram of Oriented Gradients (HOG) for the Handshapes. Linear Kernel SVM is used to model the HOD features, and validation HMM is adopted for modeling the HOG and RDF features. In the end, they introduced a normalization method to fuse the SVM and HMM probabilities for classification. They tested the model on their self-build Kinect dataset of 500 sign words and obtained a top 1, top 5, and top 10 error rates of 89%, 97%, and 99%.

Moreover, in another work, [22] presented a CSLR based on SHS descriptor and Encoder-Decoder LSTM model. They designed specific hand shape descriptors from a hand shape database acquired by Kinect. In the system, it extracts the SHS features and trajectory features which were then fed into the encoder-decoder LSTM network for recognition. The system was tested on 80-daily isolated CSL words acquired by Kinect, and good performance is reported. Comparisons were made with systems employing HOG and HMM methods.

Reference [23] presented a CSLR based on an optimized tree-structure framework using a combination of sEMG, ACC, and GYRO sensors. They categorized the CSL based on subwords according to one- or two-handed, hand orientation and hand amplitude components. An optimized tree-structure classification is adopted for the subwords recognition. Experimental results on their acquired 150 subwords dataset showed an accuracy of 94.31% in the user-specific test and 87.02% in user-independent tests.

In another work, [24] devised a single hand isolated word CSL recognition system using a combination of sensors. In the work, surface Electromyography (sEMG) and accelerometer are utilized to acquire hand movement signals from the right forearm, wrist and the back of the hand for 18 isolated CSL signs. Sliding windows are used to segment the data into several segments. For each segment window of the sEMG and ACC data, the features were extracted and combined to form a feature vector which was used to train the Linear Discriminant Analysis (LDA) classifier. After an experiment, a recognition rate of 91.4% was achieved.

In a related work, [25] presents a video-based isolated word Chinese sign language recognition using CNN. In the method, upper body images centered on the user's hands are extracted. The method simplifies the hand segmentation by background removal and transforming the color space of the hand region. A pre-trained CNN model is used as the classifier to recognize the 40 CSL vocabularies based on their self-built dataset. An accuracy of 99% was reported. Similarly, [26] proposed a combined CNN-LSTM model for recognition of CSL words. Different from previous methods, their work comprises of feeding the input video directly the CNN to generate a vector of information captured from the frames of the CSL video. LSTM, which has the capability for sequence learning task, received the generated vector for the recognition. An accuracy of 98.43% was achieved on their self-built CSL dataset which was achieved by localizing hand of the subject in the video during training to reduce the amount of computation.

In a recent work, [27] proposed a CSLR system with Sequence-to-Sequence Learning using Kinect. The proposed model consists of a CNN that extracts spatial features from the CSL image frames, and an LSTM encoder-decoder module for temporal modeling. It also consists of another encoder-decoder module that utilizes the auxiliary skeletal trajectory features of the CSL image frames. In the end, they proposed a probability combination method that fused probabilities of the image and skeletal models to get the final prediction. The dataset adopted in the experiment is captured by Kinect including RGB image and skeletal coordinates which consist of 90 CSL words widely used in our daily life. There are 100 samples for each word, which is comprised of 9000 samples in total. The experimental results from their self-built dataset demonstrated the effectiveness of the proposed method that results in an accuracy of 93.5% and 94.7% for the image model and the fused model. Furthermore, in another similar work, [28] proposed a novel Chinese Isolated sign language recognition approach based on Keyframe-Centered clips (KCC). The authors stated that

an isolated sign video contained many redundant frames. Including this redundant information, may hamper the efficiency and pose difficulty in modeling long-term dependency for Isolated SLR. The method includes extracting keyframes from an RGB-D video stream of CSL. A feature fusion layer fuses CNN features of RGB keyframes, histogram of oriented gradient (HOG), depth motion maps (DMMs), and trajectory features of skeleton joints for multi-modal KCC features. This feature fusion layer was then jointly trained with an LSTM encoder-decoder network, and the final sub-word output is estimated using softmax. Experiments were carried out on a self-built Kinect CSL words dataset, an accuracy of 89.87% and 91.18% is reported for different methods without KCC and with KCC, and also proves to be better than HMM, DTW, CNNs, and LSTMs.

Similarly, [29] presented a 3D Convolutional Neural Networks for Sign Language Recognition. Their architecture extracts both the spatial and temporal features from raw CSL video data, thus capturing motion information from dynamic gestures. The model is also deeper and iteratively integrated discriminative data representations from multi-modal data. They utilized single channels of infrared and contour data of the Kinect in their 3D CNN architecture. The model achieved 89.2% accuracy on their created Chinese sign language video in museum (SLVM) dataset (6800 samples of 20 vocabularies) and 92.4% on the Chalearn dataset.

Isolated sign word recognition is more useful than fingerspelling recognition, in that the target is recognizing Chinese sign words. There exist several works on this type of SL, with the majority employing vision-based approaches with encouraging performances. Further research in this category should be geared towards improving signer-independent word recognition and building an extensive dataset that contains all the current CSL words. With this in place, a first of its kind bidirectional CSL words to Chinese lexical e-dictionary can be realized with a competitive accuracy that will serve as a reference to CSL learners and the deaf as well.

C. CONTINUOUS SIGN SENTENCE RECOGNITION

This is the most important for real-life situations of hearing-to-deaf communication. However, as such research here is still limited partly due to unavailability of large scale standard CSL corpora. [30] developed a sign language recognition and translation system using Kinect. They used the 3D trajectory information of the Kinect and employed a Euclidean distance based trajectory matching algorithm to match the probe vectors and gallery trajectories to recognize and translate the signs. The system consists of two parts: Translation mode that translates signs to speech or text and; communication mode that does the vice-versa. The system was tested on a dataset of 239 CSL words with each word recorded 5 times. The rank-1 and rank-5 recognition rates are 83.51% and 96.32% respectively. In its sentence recognition part, it integrates individual recognized words to form a sentence using an LM. This will surely miss the rich grammatical structure of the SL as it is not just a combination of signs.

In a related work, [31] developed a continuous CSL recognition using level building based on fast HMM. They embedded HMM into the level building algorithm to improve sentence recognition accuracy. Also, grammar constraint and sign length constraint are employed to improve the recognition rate by reducing the insertion, deletion and substitution errors, and a coarse segmentation method is proposed to provide the maximal level number. In the experiments, they used only the skeleton motion features from Kinect of 100 sentences, which is composed of 5 signs. Their method shows superior performance than previous works. In a similar work, [32] introduced a Chinese SLR system using dynamic programming with warping templates obtained by dynamic time warping (DTW). They used skeleton information of the upper body from the Kinect sign video streams to learn the warping templates for the signs. They tested 180 sentences using the DP method and compared their results with HMM and CRF and obtained 85% accuracy.

Recently, [33] proposed a video-based sign language recognition without temporal segmentation using (LS-HAN) framework. In their work, the typical approach of decomposing a continuous CSL recognition by temporal segmentation, which may cause errors in recognition due to incorrect segmentation is addressed. A new two-stream 3-D CNN is designed to extract the global-local spatiotemporal features from sign videos. The features are paired with annotated sentences with each word encoded as a ''one-hot'' vector and mapped to the same latent space to bridge the semantic gap between the two features respectively. The Hierarchical Attention Network (HAN), which is an extension of LSTM with attention mechanism, is employed to generate the recognition results. Experiments were performed on RWTH-PHOENIX-Weather dataset that contains 7K weather forecasts sentences from 9 signers and their publicly available dataset acquired by Kinect. The accuracy reported was 61.7% and 82.7%.

Furthermore, in a recent similar work, [34] also presented a Hierarchical LSTM for Sign Language Translation. It solves the issue that conventional HMM and CTC may fail to tackle the cases with messing word order corresponding to visual content in sentences during recognition. They adopted the encoder-decoder framework, which respectively learns visual content and word embedding. They tested the method on their self-produced publicly available dataset and achieved a state of the art result better than previous work.

Although research on CSL continuous recognition is limited, there is a growing interest in the area. It is the most practical means of communication in real-life scenarios between the hearing people and the deaf. In contrast to Automatic speech recognition systems, that is widely available in commercial applications, continuous CSLR system is still by far not a matured technology. Availability of a large and well-annotated CSL-Chinese paired dataset, will be helpful and encouraging in realizing a sign language translation system, by adopting state-of-the-art Neural Machine Translation (NMT) models.

TABLE 1. A summary of the SLR system existing work based on its characteristics.

Author	Features	Isolated/ Continuous	Recognition Model	Acquisition Device	Dataset
Zhuang et al. [24]	Manual (wrist, back of the hand, forearm)	Isolated	LDA Classifier	sEMG, ACC	Self-built
J. Huang et al. [16]	Manual (hand shapes, depth & skeleton trajectory)	Isolated	MLP	Kinect	Self-built
He et al. [21]	Manual (Trajectory & Hand shape)	Isolated	Normalization late fusion of SVM and HMM probabilities	Kinect	Self-built
Gao et al. [14]	Manual	Isolated and continuous	SOFM, HMM, SRN	Cybergloves	Self-built
Guo et al. [34]	Manual	Continuous	Encoder-Decoder LSTM	Kinect	Public
W. Yang et al. [31]	Manual (skeleton trajectory)	Continuous	HMM	Kinect	Self-built
Zhang et al. [32]	Manual (skeleton trajectory)	Isolate and Continuous	Multi-SVM and Dynamic Programming	Kinect	Self-built
S. Yang et al. [26]	Manual (hand shapes)	Isolated	LSTM	Camera	Self-built
Wang et al. [17]	Manual (hand posture & skeleton trajectory)	Isolated	HMM	Kinect	Self-built
Zhang et al. [20]	Manual (hand shapes & trajectory features)	Isolated	HMM	Kinect	Self-built
Yang & Zhu [25]	Manual (hand feature)	Isolated	CNN	Camera	Self-built
Mao et al. [27]	Manual (hand shapes & skeleton trajectory)	Isolated	2-layer LSTM encoder-decoder	Kinect	Self-built
J. Huang et al. [33]	Manual (hand shapes – RGB modalities)	Continuous	Hierarchical attention network with latent space (LS-HAN)	Kinect	Public
S. Huang et al. [28]	Manual (hand shapes, depth motion & trajectory features)	Isolated	LSTM encoder-decoder network	Kinect	Self-built
Liang et al. [29]	Manual (infrared & contour data)	Isolated	3D CNN	Kinect	Public
Li at al. [22]	Manual (Hand shapes & trajectory features)	Isolated	LSTM encoder-decoder	Kinect	Self-built`
Liu et al. [18]	Manual (skeleton joint trajectory features)	Isolated	LSTM	Kinect	Self-built
Chai et al. [30]	Manual (Skeleton trajectory)	Isolated	Trajectory matching using Euclidean distance.	Kinect	Self-built
Pan et al. [13]	Manual (hand shapes)	CSL alphabets (fingerspelling)	SVM	Camera	Self-built
Wang et al. [19]	Manual (hand shape, Depth, and Skeleton trajectory features)	Isolated	SVM	Kinect	Self-built
X. Yang et al. [23]	Manual (hand motion signals)	Isolated	2 stream HMM classifier	sEMG, ACC & GYRO	Self-built

III. CHARACTERISTICS OF THE SLR SYSTEM

The following section discusses the aspects of a typical SLR system. Table 1 briefly outlines the aspects.

A. MANUAL AND NON-MANUAL FEATURES

Similar to [35] analysis of American sign language, features of CSL signs are based around the hands: shape, orientation,

location, and movement. The signs are distinguishable from each other by variations from those features. Therefore, the accurate detecting of the hands and tracking its movements is mandatory for a real-time SLR system. In standard SL, tracking of the hands is non-trivial as the hands are deformable objects, they move quickly, changing shape, location and also occlude each other in the process. In earlier work, signers wear colored gloves to simplify the segmentation task when tracking [36].

Almost majority of work on SLR systems utilizes only the hand features. However, the research in SLR has recently started appreciating the importance of non-manual parameters in some sign languages and is yet to be explored in depth. The non-manual features play an essential role in SL communication because they are related to the meaning of a sign or its syntax, and it leads to significant improvement in the quality of information exchanged. Facial Expression Recognition is an already well-developed research area, mainly due to its applicability in the construction of different types of AI-based systems. The non-manual features are generally centered at the face region. They include head movement, eye blinking, eyebrow movement, and mouth shape. To the best of our knowledge, no recent work incorporates these features in Chinese Sign language recognition so far. SLR systems in other SLs have adopted the non-manual features recently. For example, [37] recognized six facial expressions and used the Recursive Principal Component Analysis (RPCA) for feature extraction and MLP for classification of Arabic SL (ArSL) words. According to the authors, the integration of a facial expression recognition system improves classification results as there are some sign words, such as marriage and divorce, that are similar but are distinguished by an accompanying facial expression (happy or sad). They used an existing ArSL recognition system [38] that achieved an 88% classification rate on Arabic sign words. Test results showed that the classification rate increased to 98% after using 360 RPCA facial features and 20 hidden neurons in the MLP, resulting in a 10% increase in the original accuracy. More recently, [39] introduced a sign language recognition system based on human keypoints estimated by OpenPose; an open-source toolkit for real-time multiperson keypoint detection. The neural network model translates Korean sign language to natural language sentences by utilizing keypoints extracted from the hand, body parts, and face based on the sequence-to-sequence translation models. Experimental results on the effect of the use of keypoints information showed that the model achieved a ROUGUE-L score of 68.13 for hand and body; 64.93 for hand, body, and face; and 62.85 when utilizing hand and face keypoints. The performance drop was suspected to be due to an imbalanced number of keypoints from the various parts; 18 body keypoints, 21 keypoints from each hand, and 70 facial keypoints.

B. CSL ACQUISITION DEVICES

Ideally, an SLR system can be realized by employing vision- and sensor-based devices. The vision-based approach utilized the acquisition of images and videos for sign language recognition. Typically, there are two sources: Single/video camera are used for acquiring 2D images and videos; and Kinect Camera for acquiring RGB video with depth and trajectory information. This type of approach has gained more popularity in SLR research, as there is no need to wear or utilize cumbersome and expensive gloves or sensors. However, it might not be free of challenges ranging from lighting conditions, hand and face segmentation, complex background, and noise. For instance, some earlier works employed the use of colored gloves to ease the segmentation problem [36]. With the advent of deep learning, such as ResNet [40], the impressive success and progress being achieved in Computer Vision have reduced the effect of the challenges in SLR.

On the other hand, Sensor-based recognition methods employed wearable gloves equipped with sensors or measurement instruments attached to the hands, e.g., surface Electromyography (sEMG), Accelerometer (ACC) and Gyro sensors. These sensors provide information on the location, orientation, movement, and bending of the fingers. From data acquired from these devices, many sets of features can be extracted, and a classifier can be used to recognize the sign language words.

Recently, Leap motion controller and Intel's Real Sense has been used by researchers for sign recognition. Example, [41] used LMC and Kinect sensor for American Sign Language (ASL) recognition.

C. MODELS

In earliest work on CSLR system, handcrafted features are extracted and passed into classical sequential models such as HMM to recognize SLR words [14]. Besides, some recent work applied traditional recognition models for such a task. [13], [19] used SVM to classify CSL Alphabets and isolated words, and [30] performed a trajectory matching on the Kinetic trajectory features of CSL using Euclidean distance to classify the isolated words. In contrast to the above, recently, deep learning based models have achieved impressive success in computer vision and NLP tasks. CNNs [25], [29], LSTMs [18], [26], [34] or hybrid models [27], [33] have been adopted for isolated words and especially for continuous sentence recognition.

IV. BENCHMARK DATASETS

In SLR research, benchmark datasets are used as a standard reference and baseline for further research. The benchmark datasets allow comparisons of independent approaches. To the best of our knowledge, there exists only one comprehensive, and publicly available Chinese Sign Language dataset compiled by [33]. It contains a Compilation of the largest (as of September 2017) Chinese Sign Language (CSL) dataset for continuous SLR with sentence-level annotations. Apart from this, recent researchers usually create their customized dataset with Kinect or data glove sensors. For instance, [32] experimented their work on isolated and continuous. The dataset consists of 450 phrases over 30 isolated

TABLE 2. A summary of sign language datasets.

Dataset name	Language	Segmentation Type (Isolated/Continuous)	Content Type (Video/sensor data)	Availability	signers	sentences	Hrs of video	Vocabulary
J. Huang et al. [33]	Chinese	Isolated & Continuous	Kinect Video	Public	50	100	100+	178
DEVISGN [51]	Chinese	Isolated	Kinect Camera	On request	8			2K
SLVM [29]	Chinese	Isolated	Kinect	Public	17			20
RWTH PHOENIX-Weather 2014T [45]	German	Continuous	Camera video	Public	9	7K		1066
SIGNUM [46]	German	Continuous	Camera video	Public	25	15.6K	55.3	450
RWTH-PHOENIX-Weather 2014 [44]	German	Continuous	Camera video	Public	9	6861	10.73	1558
RWTH-BOSTON-104 [47]	English	Continuous	Camera Video	Public	3	201		104
RWTH-BOSTON-400 [48]	English	Continuous	Camera Video	On request (free)	5	843		406
ASLLVD [49]	English	Isolated	Camera Video	Public	4			3K
ATIS SL CORPUS [50]	Multilingual	Continuous	Camera Video	On request (free)	Several	680		400

signs from 5 signers, and 180 sentences randomly composed of the words from 3 signers. [28] created their dataset collection consisting of 310 isolated CSL sign words with Kinect. [21] collected the Kinect-based datasets from professional signers, which consists of 500 sign words. [20] built two datasets: Dataset I contains 100 sign words by one signer with 5 repetitions and 500 videos in total; and Dataset II, which contains 500 sign words by one signer with 5 repetitions and 2500 videos which were used to test performance on large vocabulary. Reference [22] constructed a dataset using Kinect of 80 commonly used words. Reference [42] collected two datasets with Kinect: Dataset I consists of 1000 vocabulary performed by 7 signers only once which is used for signer-independent tests; Dataset II consists of three repetitions of 1000 vocabulary signs performed by a single signer. Reference [16] used Kinect to record a CSL dataset that contains 25 vocabularies widely used in daily life. [18] produced a dataset using Kinect that contains 100 daily-used CSL words. Reference [31] create a dataset of 100 sentence samples over 20 sentences performed by 2 signers, where each sentence consists of 5 signs randomly selected from 21 isolated signs using Kinect. The authors also created a dataset that contains 714 sign samples over 21 isolated signs from 8 signers, where there are 34 samples for each sign. Reference [29] created a Kinect based dataset of CSL words. The dataset is focused on daily Chinese museum words which they named it as Sign Language Video in Museum (SLVM) dataset. It consists of 6800 samples that include 20 vocabularies, performed by 17 signers for 20 times.

DEVISIGN CSL dataset is a large isolated signs dataset created with Kinect but is only available upon request from the creators. References [23], [43] formed a self-acclaimed dataset of CSL subwords with 3D ACC, sEMG with the former author using an additional GYRO sensor.

Some other standard benchmark datasets for other sign languages include **RWTH-PHOENIX-Weather 2014** [44], **RWTH-PHOENIX-Weather 2014T** [45], **SIGNUM Corpus** [46], **RWTHBOSTON**-104 [47], **RWTH-BOSTON-400** [48], **ASLLVD** [49], **and ATIS Sign Language Corpus** [50].

Although there are Sign Language datasets available, there is an enormous scarcity of standard datasets for CSL research which is alarming. The only existing continuous sign language dataset available is recorded in a controlled environment with a limited vocabulary consisting of 100 daily-use sentences. In practice, a dataset is considered better if it is vast and contains annotations. Moreover, the annotation task is generally performed manually by experts of the SL and may consume much time and money for a large dataset.

On the contrary, the ASL and German SL corpus by RWTH has more sentences and vocabularies, and also rich gloss annotations. In all, we can finally ascertain that in order to build a high-accuracy Sign Language Translation system (SLT), attention should also be focused on creating an SL corpus that encompasses many spheres of life with rich annotations, human pose information, facial expression feature information and not only in a laboratory controlled environment. Those shortcomings are the factors inhibiting the end goal of SLT.

In table 2, the characteristics and comparisons of some of the standard continuous benchmark datasets are shown.

V. CONCLUSION AND TRENDS FOR FUTURE RESEARCH

CSLR is an on-going research that began decades ago, but till now there is no system deployed on a large scale. Realizing such a system will unquestionably have a tremendous impact on the deaf and hearing communication. This literature review provides an overview of the recent work on CSLR. Most papers reported achieving high recognition rates but with some challenges that hinder the systems from attaining a higher potential. The first issue to consider based on CSLR type, is that continuous sign language recognition is lagging than isolated word recognition. Though isolated word recognition showed good performance, continuous sign language recognition is the one that is needed the most in deploying a real-time communication system between deaf-mute persons and hearing individuals. With this, the bridging gap will be cleared. Unlike Isolated word recognition, which includes a sequence of image frames that can be recognized by a proper classifier, Continuous sign language recognition is non-trivial. It is not just a sequence of individual sign words. There is rich underlying grammatical and context information in it that must be captured in the system to accommodate real-time scenarios, which made it rather difficult.

The second issue of importance worth discussing is on CSL acquisition devices. Sensor gloves reported high accuracy but mostly used for isolated word recognition. Even though it is easier to extract the relevant features from the data acquired from the sensors, they are not popularly used. They are expensive and cumbersome for the signer. Majority of CSL research utilizes the Kinect, with the extra depth and trajectory information, accuracy is observed to be increased. Still, ordinary cameras should be realized for the task as the main aim is to develop an affordable and easy-to-use CSLR system. Currently, to the best of our knowledge, no work has been reported to include non-manual features in Chinese sign language recognition. Even though sign language revolves mostly on the hands, some context and grammatical information are also included from facial expressions, eye gaze and blink, and body postures. Future research should consider this aspect.

Thirdly, the dataset is also an issue for achieving a high-performance SLR system. From the list of datasets studied, there is only one continuous CSLR publicly available dataset and also does not contain much information as the state-of-the-art German sign language dataset [45]. The German SL is widely accepted among researchers [52] in testing their models as well as recognized as the benchmark dataset. Developing a standard CSLR dataset is open to research.

Finally, with the progress in deep learning, CSL is yet to be explored as a machine translation problem. Neural Machine translation (NMT) employs RNN-based sequence-to-sequence (seq2seq) architectures, which learn a statistical model to translate between different languages. Seq2seq [53] has seen success in translation between spoken languages. It consists of two RNNs, an encoder, and a decoder, that learn to translate a source sequence to a target sequence. Recently, [45] formulate a sign language translation based on the framework of Neural Machine Translation (NMT) with spatial and word embeddings for German sign language videos and German texts. The extracted non-linear frame from a sign video is converted into the spatial representation through a 2D CNN. The sequence-to-sequence (seq2seq) based deep learning methods learns how to translate the spatiotemporal representation of signs into speech or text.

In our final consideration in CSL research, we concluded that the research is open to three problems: Establishing a well-annotated public large corpus of CSL dataset which can also allow to improve and compare various state-of-the-art techniques; incorporating facial expressions, which are vitally important in SL and still left out in CSL that convey extra meaning in a sign language sentence, e.g., exclamation, questions, and emotions; lastly, with the tremendous success of seq2seq in speech recognition, further CSL research with NMT based models will boost its performance.

REFERENCES

[1] X. Xiao, X. Chen, and J. L. Palmer, "Chinese Deaf viewers' comprehension of sign language interpreting on television: An experimental study," *Interpreting*, vol. 17, no. 1, pp. 91–117, Jan. 2015.

[2] M. J. Cheok, Z. Omar, and M. H. Jaward, "A review of hand gesture and sign language recognition techniques," *Int. J. Mach. Learn. Cybern.*, vol. 10, no. 1, pp. 131–153, Jan. 2017.

[3] M. Mohandes, M. Deriche, and J. Liu, "Image-based and sensor-based approaches to arabic sign language recognition," *IEEE Trans. Human-Mach. Syst.*, vol. 44, no. 4, pp. 551–557, Aug. 2014.

[4] S. Joudaki, D. bin Mohamad, T. Saba, A. Rehman, M. Al-Rodhaan, and A. Al-Dhelaan, "Vision-based sign language classification: A directional review," *IETE Tech. Rev.*, vol. 31, no. 5, pp. 383–391, Oct. 2014.
[5] M. E. Al-Ahdal and M. T. Nooritawati, "Review in sign language recognition systems," in *Proc. IEEE Symp. Comput. Inform. (ISCI)*, Mar. 2012, pp. 52–57.
[6] S. Kausar and M. Y. Javed, "A survey on sign language recognition," in *Proc. Frontiers Inf. Technol.*, Dec. 2011, pp. 95–98.
[7] S. C. Agrawal, A. S. Jalal, and R. K. Tripathi, "A survey on manual and non-manual sign language recognition for isolated and continuous sign," *Int. J. Appl. Pattern Recognit.*, vol. 3, no. 2, pp. 99–134, 2016.
[8] A. Er-Rady, R. Faizi, R. O. H. Thami, and H. Housni, "Automatic sign language recognition: A survey," in *Proc. Int. Conf. Adv. Technol. Signal Image Process. (ATSIP)*, May 2017, 2017.
[9] L. Zheng, B. Liang, and A. Jiang, "Recent advances of deep learning for sign language recognition," in *Proc. Int. Conf. Digit. Image Comput., Techn. Appl. (DICTA)*, Nov./Dec. 2017, pp. 1–7.
[10] W. Jiangqin and G. Wen, "The recognition of finger-spelling for Chinese sign language," in *Proc. Int. Gesture Workshop*, vol. 1, 2002, pp. 96–100.
[11] Q. Yang, J. Peng, and L. Yulong, "Chinese sign language recognition based on gray-level co-occurrence matrix and other multi-features fusion," in *Proc. 4th IEEE Conf. Ind. Electron. Appl.*, May 2009, pp. 1569–1572.
[12] Y. Ji, C. Liu, S. Gong, and W. Cheng, "3D hand gesture coding for sign language learning," in *Proc. Int. Conf. Virtual Reality Vis. (ICVRV)*, Sep. 2016, pp. 407–410.
[13] T.-Y. Pan, L.-Y. Lo, C.-W. Yeh, J.-W. Li, H.-T. Liu, and M.-C. Hu, "Sign language recognition in complex background scene based on adaptive skin colour modelling and support vector machine," *Int. J. Big Data Intell.*, vol. 5, nos. 1–2, pp. 21–30, 2018.
[14] W. Gao, G. Fang, D. Zhao, and Y. Chen, "A Chinese sign language recognition system based on SOFM/SRN/HMM," *Pattern Recognit.*, vol. 37, no. 12, pp. 2389–2402, Dec. 2004.
[15] G. Fang, W. Gao, and D. Zhao, "Large-vocabulary continuous sign language recognition based on transition-movement models," *IEEE Trans. Syst., Man, Cybern. A, Syst., Humans*, vol. 37, no. 1, pp. 1–9, Jan. 2007.
[16] J. Huang, W. Zhou, H. Li, and W. Li, "Sign language recognition using 3D convolutional neural networks," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jun./Jul. 2015, pp. 1–6.
[17] H. Wang, X. Chai, Y. Zhou, and X. Chen, "Fast sign language recognition benefited from low rank approximation," in *Proc. 11th IEEE Int. Conf. Workshops Autom. Face Gesture Recognit. (FG)*, May 2015, pp. 1–6.
[18] T. Liu, W. Zhou, and H. Li, "Sign language recognition with long short-term memory," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2016, pp. 2871–2875.
[19] H. Wang, X. Chai, X. Hong, G. Zhao, and X. Chen, "Isolated sign language recognition with Grassmann covariance matrices," *ACM Trans. Accessible Comput.*, vol. 8, no. 4, p. 14, May 2016.
[20] J. Zhang, W. Zhou, C. Xie, J. Pu, and H. Li, "Chinese sign language recognition with adaptive HMM," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2016, pp. 1–6.
[21] J. He, Z. Liu, and J. Zhang, "Chinese sign language recognition based on trajectory and hand shape features," in *Proc. Vis. Commun. Image Process. (VCIP)*, Nov. 2016, pp. 1–4.
[22] X. Li, C. Mao, S. Huang, and Z. Ye, *Chinese Sign Language Recognition Based on SHS Descriptor and Encoder-Decoder LSTM Model* (Lecture Notes in Computer Science: Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 10568. Cham, Switzerland: Springer, 2017, pp. 719–728.
[23] X. Yang, X. Chen, X. Cao, S. Wei, and X. Zhang, "Chinese sign language recognition based on an optimized tree-structure framework," *IEEE J. Biomed. Health Inform.*, vol. 21, no. 4, pp. 994–1004, Jul. 2017.
[24] Y. Zhuang, B. Lv, X. Sheng, and X. Zhu, "Towards Chinese sign language recognition using surface electromyography and accelerometers," in *Proc. 24th Int. Conf. Mechatronics Mach. Vis. Pract. (M2VIP)*, Nov. 2017, pp. 1–5.
[25] S. Yang and Q. Zhu, "Video-based Chinese sign language recognition using convolutional neural network," in *Proc. IEEE 9th Int. Conf. Commun. Softw. Netw. (ICCSN)*, May 2017, pp. 929–934.
[26] S. Yang and Q. Zhu, "Continuous Chinese sign language recognition with CNN-LSTM," *Proc. SPIE*, vol. 10420, Jul. 2017, Art. no. 104200F.
[27] C. Mao, S. Huang, X. Li, and Z. Ye, "Chinese sign language recognition with sequence to sequence learning," in *Computer Vision*, vol. 771. Singapore: Springer, 2017, pp. 180–191.
[28] S. Huang, C. Mao, J. Tao, and Z. Ye, "A novel Chinese sign language recognition method based on keyframe-centered clips," *IEEE Signal Process. Lett.*, vol. 25, no. 3, pp. 442–446, Mar. 2018.
[29] Z. Liang, S.-B. Liao, and B.-Z. Hu, "3D convolutional neural networks for dynamic sign language recognition," *Comput. J.*, vol. 61, no. 11, pp. 1724–1736, 2018.
[30] X. Chai, G. Li, Y. Lin, Z. Xu, Y. Tang, and X. Chen, "Sign language recognition and translation with kinect," in *Proc. 10th IEEE Int. Conf. Autom. Face Gesture Recognit.*, Apr. 2013, pp. 22–26.
[31] W. Yang, J. Tao, and Z. Ye, "Continuous sign language recognition using level building based on fast hidden Markov model," *Pattern Recognit. Lett.*, vol. 78, pp. 28–35, Jul. 2016.
[32] J. Zhang, W. Zhou, and H. Li, "A new system for Chinese sign language recognition," in *Proc. IEEE China Summit Int. Conf. Signal Inf. Process. (ChinaSIP)*, Jul. 2015, pp. 534–538.
[33] J. Huang, W. Zhou, Q. Zhang, H. Li, and W. Li, "Video-based sign language recognition without temporal segmentation," in *Proc. 32nd AAAI Conf. Artif. Intell.*, Apr. 2018, pp. 1–8.
[34] D. Guo, W. Zhou, H. Li, and M. Wang, "Hierarchical LSTM for sign language translation," in *Proc. 32nd AAAI Conf. Artif. Intell.*, Apr. 2018, pp. 6845–6852.
[35] W. C. Stokoe, Jr., "Sign language structure: An outline of the visual communication systems of the American deaf," *J. Deaf Stud. Deaf Edu.*, vol. 10, no. 1, pp. 3–37, 1960.
[36] M. Mohandes and M. Deriche, "Image based arabic sign language recognition," in *Proc. 8th Int. Symp. Signal Process. Appl.*, vol. 1, Aug. 2005, pp. 86–89.
[37] A. S. Elons, M. Ahmed, and H. Shedid, "Facial expressions recognition for arabic sign language translation," in *Proc. 9th Int. Conf. Comput. Eng. Syst. (ICCES)*, Dec. 2014, pp. 330–335.
[38] M. F. Tolba, A. Samir, and M. Aboul-Ela, "Arabic sign language continuous sentences recognition using PCNN and graph matching," *Neural Comput. Appl.*, vol. 23, nos. 3–4, pp. 999–1010, Sep. 2013.
[39] S.-K. Ko, C. J. Kim, H. Jung, and C. Cho, "Neural sign language translation based on human keypoint estimation," 2018, *arXiv:1811.11436*. [Online]. Available: https://arxiv.org/abs/1811.11436
[40] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[41] G. Marin, F. Dominio, and P. Zanuttigh, "Hand gesture recognition with leap motion and Kinect devices," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2014, pp. 1565–1569.
[42] F. Yin, X. Chai, and X. Chen, *Iterative Reference Driven Metric Learning for Signer Independent Isolated Sign Language Recognition* (Lecture Notes in Computer Science: Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 9911. Cham, Switzerland: Springer, 2016, pp. 434–450.
[43] Y. Li, X. Chen, X. Zhang, K. Wang, and Z. J. Wang, "A sign-component-based framework for chinese sign language recognition using accelerometer and sEMG data," *IEEE Trans. Biomed. Eng.*, vol. 59, no. 10, pp. 2695–2704, Oct. 2012.
[44] O. Koller, J. Forster, and H. Ney, "Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers," *Comput. Vis. Image Understand.*, vol. 141, pp. 108–125, Dec. 2015.
[45] N. C. Camgoz, S. Hadfield, O. Koller, H. Ney, and R. Bowden, "Neural sign language translation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 7784–7793.
[46] U. Von Agris and K.-F. Kraiss, "Towards a video corpus for signer-independent continuous sign language recognition," in *Proc. Gesture Hum.-Comput. Interact. Simulation*, May 2007, pp. 10–11.
[47] M. Zahedi, P. Dreuw, D. Rybach, T. Deselaers, and H. Ney, "Geometric features for improving continuous appearance-based sign language recognition," in *Proc. BMVC*, vol. 3, 2006, pp. 1019–1028.
[48] P. Dreuw, C. Neidle, V. Athitsos, S. Sclaroff, and H. Ney, "Benchmark databases for video-based automatic sign language recognition," in *Proc. Sixth Int. Conf. Lang. Resour. Eval. (LREC)*, May 2008, pp. 1–6.
[49] V. Athitsos, C, Neidle, S. Sclaroff, J. Nash, A. Stefan, Q. Yuan, and A. Thangali, "The American sign language lexicon video dataset," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2008.
[50] J. Bungeroth, D. Stein, P. Dreuw, H. Ney, S. Morrissey, A. Way, and L. van Zijl, "The ATIS sign language corpus," in *Proc. Int. Conf. Lang. Resour. Eval. (LREC)*, 2008, pp. 2943–2946.

[51] X. Chai, H. Wang, and X. Chen, "The devisign large vocabulary of Chinese sign language database and baseline evaluations," Key Lab Intell. Inf. Process. Chin. Acad. Sci., Inst. Comput. Technol., CAS, Beijing, China, Tech. Rep. VIPL-TR-14-SLR-001, 2014.
[52] J. Pu, W. Zhou, and H. Li, "Dilated convolutional network with iterative optimization for continuous sign language recognition," in *Proc. IJCAI Int. Joint Conf. Artif. Intell.*, Jul. 2018, pp. 885–891.
[53] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 3104–3112.

SUHAIL MUHAMMAD KAMAL received the B.Eng. degree in computer engineering from Bayero University Kano (BUK) and the M.Eng. degree in electrical–computer and microelectronic system from Universiti Teknologi Malaysia. He is currently pursuing the Ph.D. degree in computer science and technology with Xiamen University, China. He is also a Lecturer with the Department of Information Technology, BUK. His research interests include sign language recognition, machine translation, deep learning, computer vision, image processing, and NLP.

YIDONG CHEN was born in Xiamen, Fujian, China, in 1977. He received the Ph.D. degree in artificial intelligence from Xiamen University, China, in 2008.

From 2006 to 2009, he was an Assistant Professor with the School of Information Science and Engineering, Xiamen University, where he has been an Associate Professor with the School of Information and Engineering, since 2009. He is the author of more than 50 articles and one invention. His research interests include machine translation (MT), question answering (QA), and natural language processing (NLP) applications.

Dr. Chen received the third prize of the Science and Technology Award of Fujian Province, in 2005, the second prize for the Outstanding Academic Papers of Natural Sciences of Fujian Province, in 2012, the first prize for the Qian Weichang Chinese Information Processing Award, in 2016, and the first prize of the Science and Technology Award of the Tibet Autonomous Region, in 2018. He acts as the Secretary-General of the Fujian Association for Artificial Intelligence, China, and is the reviewer for many top conferences in the area of NLP, such as EMNLP, COLING, and so on.

SHAOZI LI (SM'18) received the B.S. degree from Hunan University, the M.S. degree from Xi'an Jiaotong University, and the Ph.D. degree from the National University of Defense Technology. He currently serves as the Chair and a Professor with Cognitive Science Department, Xiamen University. His research interests include artificial intelligence and its applications, moving objects detection and recognition, machine learning, computer vision, and multimedia information retrieval. He has directed and completed more than 20 research projects, including several national 863 programs, the National Nature Science Foundation of China, and the Ph.D. Programs Foundation of Ministry of Education of China. He is a Senior Member of the ACM and the China Computer Federation (CCF). He is the Vice Director of the Technical Committee on Collaborative Computing of CCF and the Fujian Association of Artificial Intelligence.

XIAODONG SHI was born in Jiangyin, Jiangsu, China, in 1966. He received the B.S. degree in computer software from Nanjing University, in 1988, and the M.S. and Ph.D. degrees in computer software and theory from the National University of Defense Technology, in 1991 and 1994, respectively.

From 1994 to 1998, he was a Lecturer with the Department of Computer Science, National University of Defense Technology. From 1999 to 2002, he was the Vice Chief Engineer of the Sunshine Computer and Artificial Intelligence Development, Ltd. Since 2002, he has been a Professor with the Department of Cognitive Science, Xiamen University. He is the author of more than 30 papers and holds two patents. His research interests include machine translation, natural language processing, and artificial intelligence. He has been one of the Executive Council Member of the Chinese Information Processing Society of China, and also the Editor of the *Journal of Chinese Information Processing*.

Prof. Shi was a recipient of the first prize of the Qian Weichang Chinese Information Processing Award, in 2016, and the first prize of the Science and Technology Award of the Tibet Autonomous Region, in 2017.

JIANGBIN ZHENG is currently pursuing the master's degree with Xiamen University, specializing in intelligent science and technology. He is currently involved in cross-modal sign language translation research.