This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2022.3223260

Collaborative Multilingual Continuous Sign Language Recognition: A Unified Framework

Hezhen Hu, Junfu Pu*, Wengang Zhou, *Senior Member, IEEE*, Houqiang Li, *Fellow, IEEE*,

***Abstract*—Current continuous sign language recognition systems generally target on a single language. When it comes to the multilingual problem, existing solutions often build separate models based on the same network and then train them with their corresponding sign language corpora. Observing that different sign languages share some low-level visual patterns, we argue that it is beneficial to optimize the recognition model in a collaborative way. With this motivation, we propose the *first* unified framework for multilingual continuous sign language recognition. Our framework consists of a shared visual encoder for visual information encoding, multiple language-dependent sequential modules for long-range temporal dependency learning aimed at different languages, and a universal sequential module to learn the commonality of all languages. An additional language embedding is introduced to distinguish different languages within the shared temporal encoders. Further, we present a max-probability decoding method to obtain the alignment between sign videos and sign words for visual encoder refinement. We evaluate our approach on three continuous sign language recognition benchmarks, *i.e.,* RWTH-PHOENIX-Weather, CSL and GSL-SD. The experimental results reveal that our method outperforms the individually trained recognition models. Our method also demonstrates better performance compared with state-of-the-art algorithms.**

***Index Terms*—Continuous Sign Language Recognition, Multilingual.**

I. Introduction

As a class of natural languages, sign languages (SL) convey concrete semantic meanings through manual modality such as gesture and hand movements [1], [2], [3], together with non-manual information, *i.e.,* facial expressions, *etc.* It serves as an indispensable communication tool for deaf people in their daily life. To facilitate such communication, continuous sign language recognition (SLR) [4], [5], [6], [7] has been widely investigated, which aims to automatically recognize the sequential sign words performed by the signer in the video. Most existing SLR algorithms are designed for single-lingual sign language [8], [9], [10], [11], which limits the SLR system to recognize only the specific sign language.

Similar to natural language, sign languages in different countries are non-universal, with their specific grammars and lexicons. In other words, different sign languages are not mutually intelligible. To address the problem of multilingual sign language recognition, most existing works take it for granted that separate models based on the same network architecture are trained with the corresponding sign language corpus. Although encouraging results have been achieved, such a paradigm overlooks the fact that some low-level visual patterns are shared across different sign languages, despite their different linguistic rules in different countries.

(a) Sign word ‘good’ in German (left) and Chinese (right).

(b) Interrogative sign word ‘what’ in German (left) and Chinese (right).

Fig. 1: Sign word examples in different sign languages.

For instance, the word ‘good’ performed in German sign language and Chinese sign language share the same gestures, as shown in Figure 1a. Besides, when a signer asks a question in sign language, a question word at the end of the sentence is signed. Such interrogatives sometimes follow the same visual patterns in different sign languages. Take the word ‘what’ as an example, both expressions in German and Chinese sign language can be described as ‘to put both hands outward in the front with elbows bent, and spread hands’, as shown in Figure 1b. Such common visual patterns could be represented and learned by multilingual settings. In other words, it is beneficial to train the recognition model in a collaborative way, which will promote the recognition performance over separately trained models. As a byproduct of collaborative multilingual sign language learning, the involved training data is cumulatively augmented, which reduces overfitting in learning with limited data.

In this paper, to explore the shared visual patterns across different sign languages, we present a simple yet effective method to recognize multiple sign languages in a unified framework. We take advantage of multilingual sign videos to improve the recognition performance for all sign languages

Hezhen Hu, Junfu Pu, Wengang Zhou, and Houqiang Li are with the CAS Key Laboratory of Technology in Geo-spatial Information Processing and Application System, Department of Electronic Engineering and Information Science, University of Science and Technology of China, Hefei, 230027, China (E-mail: {alexhu, pjh}@mail.ustc.edu.cn, {zhwg, lihq}@ustc.edu.cn).

*Equal contribution with the first author.

Corresponding authors: Wengang Zhou and Houqiang Li.

Authorized licensed use limited to: University of Science & Technology of China. Downloaded on June 03,2023 at 05:01:09 UTC from IEEE Xplore. Restrictions apply.
© 2022 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2022.3223260

involved. In our method, we first use a common visual encoder to extract feature representations for all sign videos. Then, for each sign language, a sequential module is adopted to learn its distinctive characteristics. The shared visual patterns and common features among all sign languages are encoded by a sharing sequential module, which is initialized with language-identified embeddings. To further improve the performance, the visual encoder is refined with the max-probability alignment between the videos and sign words.

The contributions of this work are summarized as follows:

- To our best knowledge, we are the *first* to explore the multilingual topic in continuous SLR and propose a unified framework targeting this problem, which takes advantage of multilingual data to improve the recognition performance for all involved sign languages, especially for low-resource corpus with limited labeled sign videos.
- We propose a max-probability decoding method based on the target sequence probability matrix to obtain the alignments between videos and sign words for further visual encoder refinement.
- Evaluated on the RWTH-PHOENIX-Weather, CSL and GSL-SD benchmarks, our proposed method performs favorably against the individually optimized recognition models and achieves state-of-the-art performance on both datasets.

The rest of this paper is organized as follows. Section II gives a review of the related work on deep-learning-based continuous sign language recognition methods. In Section III, we elaborate the proposed framework including the unified multilingual sign language recognition network and refinement algorithm. After that, in Section IV, we provide the experimental results as well as discussions. Section V concludes this paper, provides further insights and discusses the future directions for this new multilingual topic.

II. Related Work

In this section, we briefly review existing continuous sign language recognition methods and elaborate their key modules, *i.e.,* feature extraction and sequential correspondence learning.

In the task of continuous sign language recognition (CSLR), each sign video corresponds to a sequence of glosses in an order consistent with the related sign actions. The problem of continuous SLR can be formulated as a mapping learning from a video sequence to a sign gloss sequence. Generally, a continuous SLR system consists of two key modules, *i.e.,* a visual encoder to extract video features and a sequence learning module to learn the correspondence between the visual features and sign glosses.

Feature representation for sign videos plays an important role in sign language recognition [4], [12], [13], [14], [15], [16], [17], [18], [19]. Early works utilize hand-crafted features as video representation. They describe hand motion, shape and appearance by using Volume Local Binary Patterns (VLBP) [20], HOG or HOG-3D [21], [22], SIFT [23], and motion trajectories [22], [23], [24]. Inspired by the great success of deep neural networks, there is a growing trend to utilize Convolutional Neural Networks (CNNs) for video representation learning. There exist many newly designed CNNs based on 2D convolutions [25], [26], [27], [28], [29], [30], [31], [32], 3D convolutions [33], [34], [35], [36], [37], and 2D/3D mixed convolutions [38], [39]. With this trend, researchers investigate suitable CNNs for sign language recognition. Oscar *et al.* [40] and Necati *et al.* [41] use 2D CNNs as the feature extractor for RGB images in an end-to-end way, and achieve remarkable performance in continuous SLR. To model temporal dependency in videos, some works are proposed [7], [42], [43]. SF-Net [44] uses mixed 2D/3D CNN to improve SLR performance. An alternative to the above-mentioned backbone for spatial-temporal representation is 2D CNN followed by 1D temporal convolution network (TCN) [6], [45], [46], which achieves the state-of-the-art performance in continuous SLR. It is noted that this kind of architecture (2DCNN-TCN) has become the mainstream method thanks to its simplicity and effectiveness.

As for sequence learning, the sequential models can be divided into three categories, *i.e.,* recurrent neural network with connectionist temporal classification (CTC) [47], Hidden Markov Model (HMM) or Hidden Conditional Random Fields (HCRF), and Encoder-Decoder network. The Recurrent Neural Networks (RNNs), *i.e.*, Long Short-Term Memory (LSTM) [48], [49], Gated Recurrent Unit (GRU) [50], have been successfully applied to sequential problems, such as speech recognition [51], machine translation [52], [53], gesture understanding [54], and video captioning [55], [56], [57], [58]. CTC has been successfully applied in speech recognition [59], handwriting recognition [60], and action recognition [61]. It has also been explored in continuous SLR. In [6], [7], [9], [41], bidirectional LSTM-CTC architecture is employed as a basic model for continuous SLR and becomes the most popular one. With the superiority of CTC, state-of-the-art performance is achieved on the RWTH-PHOENIX-Weather benchmark. HMM is effectively utilized in deep neural networks for SLR and related topics [4], [62]. Oscar *et al.* [4], [10], [40] propose hybrid CNN-HMMs to integrate 2D CNN with HMM to model the state transitions for statistical continuous sign language recognition. Similar to neural machine translation, some related works [42], [63] attempt to adopt an attention-aware encoder-decoder network to learn the correspondence mapping between visual features and sign words.

Due to the lack of temporal boundary labels for the sign gloss, continuous SLR can also be viewed as a weakly supervised learning problem. Recent works have demonstrated the importance of finding segment-gloss alignments, which serve as a pseudo label for the refinement of the visual encoder. In this way, the whole architecture can be optimized in an iterative way for performance boosting. Pu *et al.* [7] propose to use a soft Dynamic Time Warping (soft-DTW) alignment constraint, while the warping path indicates the possible alignment between input video clips and sign words. In [6], [64], Cui *et al.* suggest a similar pseudo label decoding method via dynamic programming with encouraging performance.

Some previous works [7], [43] focus on improving the network capability for learning the correspondence between video and sign gloss from different perspectives to boost performance. In other words, in [7], the soft Dynamic Time

Authorized licensed use limited to: University of Science & Technology of China. Downloaded on June 03,2023 at 05:01:09 UTC from IEEE Xplore. Restrictions apply.
© 2022 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2022.3223260

Warping is utilized as the alignment constraint, while in [43], dilated networks for temporal modeling are designed. In our work, we only choose the most representative components as our backbone for generality. Differently, we approach SLR from a novel perspective by exploring multilingual SLR in a unified framework, which contains a shared visual encoder, an independent sequential module for each language together with a shared sequential module. Along with the multilingual framework, we also propose an effective refinement strategy.

III. Our Approach

In this section, we first introduce three basic architectures to deal with multilingual continuous SLR. After that, we elaborate our proposed multilingual SLR framework followed by optimization and refinement strategies.

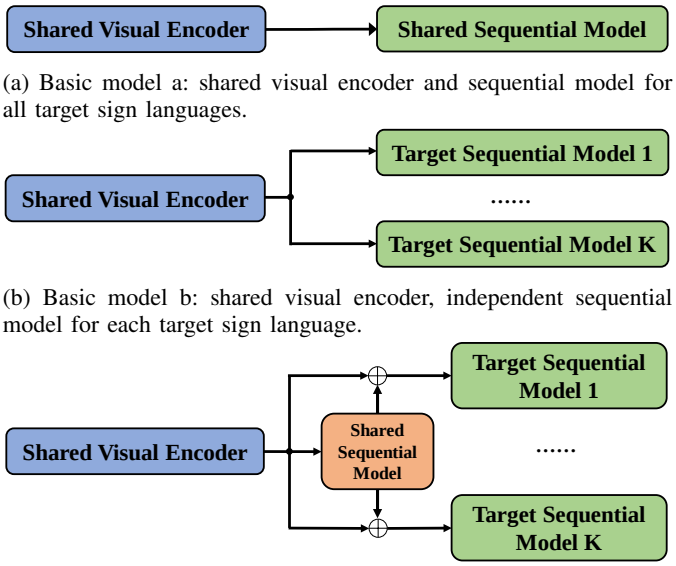
A. Basic Models

Similar to NMT, we consider three basic architectures to deal with multilingual sign language recognition shown in Figure 2. The simplest way is to use a shared visual encoder as well as a shared sequential model for all languages, illustrated in Figure 2a. This simple architecture can be easily implemented without making significant changes to the current basic continuous SLR system. However, it ignores the domain gap among different languages, which may lead to limited modeling capability. Another way is to use separate sequential models for different sign languages while sharing the same visual encoder, shown in Figure 2b. Each branch acts as same as the classical SLR system, without sharing information across each target sequential model. Such a paradigm is language-aware but fails to take the complementarity of different sign languages into consideration.

In order to take advantage of both basic architectures discussed above and avoid their limitations, the third alternative is to add an additional shared sequential model to learn the commonality among all involved sign languages, as well as separate target sequential models for each sign language, as shown in Figure 2c. In this paper, we select the third one for our framework.

B. Network Architecture

1) Framework Overview: Our multilingual continuous SLR system is employed on the third basic architecture illustrated in Figure 2c. The proposed continuous SLR system consists of a common CNN-TCN for feature extraction targeting all sign languages involved, shown in Figure 3. We use the separate bidirectional Long Short-Term Memory (BLSTM) as the sequential model to learn the correspondence between visual features and sign words for each sign language, respectively. Besides, as discussed in Section I, some similar sign words are sharing the same visual patterns among different sign languages. To model such shared patterns and common linguistic properties, we use an additional sequential BLSTM to model the commonality among all sign languages. Each branch of a specific sign language is optimized with CTC loss.



(a) Basic model a: shared visual encoder and sequential model for all target sign languages.

(b) Basic model b: shared visual encoder, independent sequential model for each target sign language.

(c) Basic model c: shared visual encoder, independent sequential model together with a shared sequential model.

Fig. 2: Proposed three basic architectures for multilingual SLR.

2) Visual Encoder: The target of continuous SLR is to learn the correspondence mapping between a video $\boldsymbol{X} = \{x_t\}_{t=1}^{N}$ and a sequence of sign glosses $\mathbf{s} = \{s_i\}_{i=1}^{L}$, where N and L are the number of frames in sign video and the number of total sign glosses, respectively.

For spatial-temporal representation, we choose the mainstream method, *i.e.,* 2DCNN-TCN, as the backbone. This architecture is first proposed in DNF [6] and achieves state-of-the-art performance on continuous SLR. From then on, it has become a popular method for comparison and has been adopted as a baseline such as in [65], [45], [46]. Specifically, the 2DCNN is implemented with GoogLeNet [66]. The TCN consists of two 1D temporal convolutional layers with a kernel size of 5, and each followed by a pooling layer with a kernel size of 2, with the combination like *conv1d-pool-conv1d-pool*. The strides for both convolutional layers and pooling layers are set to be 1. Thus, the length of output after the visual encoder is reduced to $N/4$, and the receptive field along the temporal dimension is calculated as 16 frames. Following [6], we also verify its stated kernel sizes as the optimal setting. Denote the mapping function of the visual encoder CNN-TCN as $E_v(\cdot)$, the feature representations $\boldsymbol{F}$ of the video after going through CNN-TCN can be written as follows,

$$\boldsymbol{F} = (f_1, \cdots, f_{N/4}) = \{E_v(x_t)\}_{t=1}^{N}. \tag{1}$$

3) Sequence Learning: Long Short-Term Memory (LSTM) units are adopted to explore the long-range temporal dependency. The LSTM unit at time step t can be represented with the cell state C_t and hidden state h_t. In LSTM, the gated structure is introduced to control the update of the cell state and hidden state for each time step. An LSTM unit has three kinds of gates as different information controllers, *i.e.*, input

Authorized licensed use limited to: University of Science & Technology of China. Downloaded on June 03,2023 at 05:01:09 UTC from IEEE Xplore. Restrictions apply.
© 2022 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2022.3223260

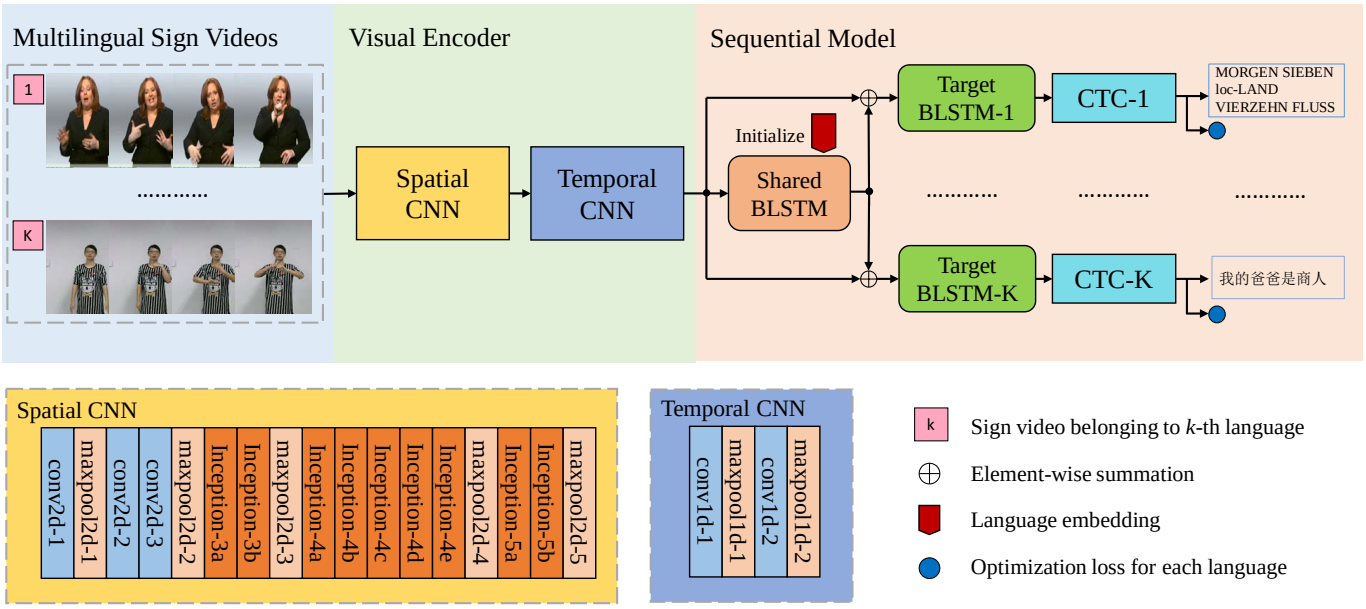


Fig. 3: Overview of our multilingual sign language recognition framework. The system consists of a common CNN-TCN visual feature extractor, language-independent BLSTM-CTC branches, together with a shared BLSTM initialized with language embeddings. In this work, the spatial CNN is implemented with GoogLeNet.

gate $g_t^{(i)}$, forget gate $g_t^{(f)}$, and output gate $g_t^{(o)}$ as follows:

$$g_t^{(i)} = \sigma(W_i \cdot [h_{t-1}, f_t] + b_i), \tag{2}$$

$$g_t^{(f)} = \sigma(W_f \cdot [h_{t-1}, f_t] + b_f), \tag{3}$$

$$g_t^{(o)} = \sigma(W_o \cdot [h_{t-1}, f_t] + b_o), \tag{4}$$

where σ is the sigmoid active function, t is the current time step, f_t is the input feature, W and b are parameters in linear projection. After taking the input feature, the cell state c_{t-1} and hidden state h_{t-1} from the previous time step, the current status and output of LSTM are updated as follows:

$$\widetilde{c}_t = \tanh(W_c \cdot [h_{t-1}, f_t] + b_c), \tag{5}$$

$$c_t = g_t^{(f)} \odot c_{t-1} + g_t^{(i)} \odot \widetilde{c}_t, \tag{6}$$

$$h_t = g_t^{(o)} \odot \tanh(c_t), \tag{7}$$

where $\odot$ means element-wise product.

In general, LSTM only models the temporal dependency along a single direction, *i.e.*, the output only depends on the current input and features from previous time steps. To model the temporal dependency with all input features from both forward and backward time steps, inspired by [41], we use bidirectional long short-term memory (BLSTM) to learn the long-range temporal dependency of the sign videos in an end-to-end manner. In addition, it also captures the correspondence between visual features and sign glosses, which benefits the following recognition. In our approach, two kinds of BLSTM units are adopted for different purposes. On the one hand, we expect to use an independent sequential model to learn the mapping between the visual features and sign words, since each sign language has its own linguistic rules. Separate sequential branches targeting different sign languages have the capability of capturing the characteristics of each specific sign language and could somehow reduce the disturbance. On the other hand, to encode similar visual patterns, *i.e.,* the same sign with the same meaning in different sign languages, and the commonality among all sign languages, a shared sequential model is introduced in our approach.

To embed the sign language identity to the shared sequential model, we use an embedding layer to encode the categories of the involved sign languages and initialize the hidden states and cell states of the shared BLSTM with language embeddings. Specifically, language embedding is a vector attached to a specific language and is learnable during training. It is computed via the common word-to-vector mapping, which embeds each language with a unique vector. And each vector belonging to a certain language is randomly initialized. The separate BLSTM branches are initialized with zero vectors, which is the same as the classical SLR systems.

With the visual features $\boldsymbol{F}$ extracted by CNN-TCN, the outputs of the shared sequential model $\boldsymbol{O}_s$ are represented as follows,

$$\boldsymbol{O}_s = \{o_s|t\}_{t=1}^{N/4} = BLSTM_s(\boldsymbol{F}; h_0 = e_k, c_0 = e_k), \tag{8}$$

where h_0 and c_0 are the initial hidden state and cell state, and e_k is the language embedding for the k-th sign language. We add the shared BLSTM outputs and CNN-TCN features for further correspondence mapping learning through the specific sequential model targeting different sign languages. The outputs of each separate target BLSTM branch for the corresponding sign language are formulated as follows,

$$\boldsymbol{O}_b^{(k)} = \{o_b^{(k)}|t\}_{t=1}^{N/4} = BLSTM_b^{(k)}(\boldsymbol{F}+\boldsymbol{O}_s; h_0 = \boldsymbol{0}, c_0 = \boldsymbol{0}), \tag{9}$$

where k indicates the k-th kind of sign language.

Authorized licensed use limited to: University of Science & Technology of China. Downloaded on June 03,2023 at 05:01:09 UTC from IEEE Xplore. Restrictions apply.
© 2022 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2022.3223260

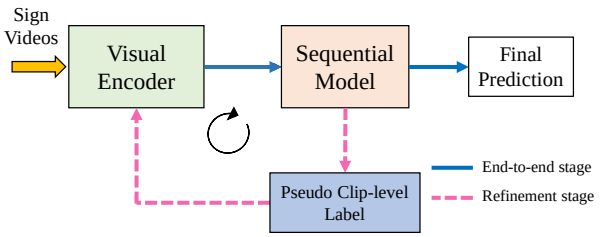


Fig. 4: Illustration of the optimization process. Two stages are alternately processed for better sentence-level prediction.

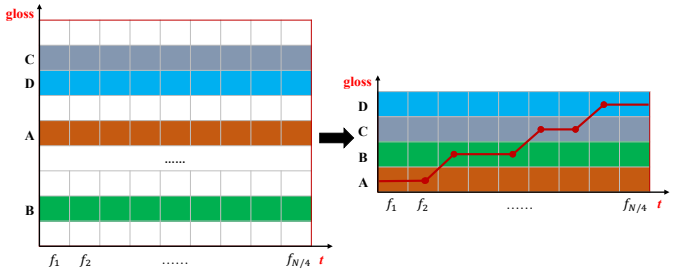


Fig. 5: Illustration of pseudo clip-level label generation for refinement. In the left figure, the abscissa represents the time (or index of frames), and the ordinate is gloss. The value in the coordinate (i, j) represents the predicted probability of the gloss j at the timestamp f_i. For the right figure, we extract the probability at each moment in the order of the ground-truth gloss sequence. Based on this probability matrix, we perform dynamic programming to get the alignment path with the maximum probability summation score.

C. Multilingual Optimization

To obtain the probability distribution corresponding to the target sequence of sign glosses, the outputs of each BLSTM branch are projected into categorical probabilities with a fully-connected and softmax layer, formulated as follows,

$$\boldsymbol{Y}^{(k)} = Softmax(Y_{t,s}^{(k)}) = Softmax(W_{fc}^{(k)} \cdot \boldsymbol{O}_b^{(k)} + b_{fc}^{(k)}), \tag{10}$$

where superscript k indicates the type of sign language, $Y_{t,s}$ is the probability of the t-th frame segment belonging to sign word s. To optimize the multilingual network, we employ connectionist temporal classification (CTC), with the objective of maximizing the posterior probability of the alignment from the source sequence to the target sequence. By introducing a blank label "$-$", CTC can deal with stillness, transitions, and reduplicated patterns. Define a many-to-one mapping $\mathcal{B}$ by removing the blank and repeated labels to obtain the final result, the posterior probability of target gloss sequence $\boldsymbol{s}$ is calculated as the sum of probabilities of all corresponding alignment paths:

$$p(\boldsymbol{s}|\boldsymbol{Y}^{(k)}) = \sum_{\pi \in \mathcal{B}^{-1}(\boldsymbol{s})} p(\pi|\boldsymbol{Y}^{(k)}), \tag{11}$$

where $\mathcal{B}^{-}1(\boldsymbol{s}) = \{\pi|\mathcal{B} = \boldsymbol{s}\}$ is the inverse mapping. The CTC loss of each separate branch for different sign languages is defined as follows,

$$\mathcal{L}_{\mathrm{CTC}}^{(k)} = -\ln p(\boldsymbol{s}|\boldsymbol{Y}^{(k)}). \tag{12}$$

For multilingual optimization, the total objective loss is the summation of the CTC losses for all branches, written as follows,

$$\mathcal{L}_{\mathrm{CTC}} = \sum_{k=1}^{K} \mathcal{L}_{\mathrm{CTC}}^{(k)}. \tag{13}$$

D. CNN-TCN Refinement

In continuous SLR, the encoded visual feature plays a vital role in final performance. However, during end-to-end training, the CTC objective is insufficient for optimizing the visual encoder due to the CTC spiky prediction and vanishing gradient in the low backbone layers. Drawing the experience of previous methods [7], [9], [40], [43], [67], we resort to the iterative training strategy (CNN-TCN refinement) to relieve this issue. The iterative training process is illustrated in Figure 4.

The idea of refinement is to explicitly fine-tune the visual encoder with clip-level labels, which are obtained via alignments between input videos and sign glosses. With this training setup, the fine-tuned CNN-TCN is utilized at the continuous SLR stage. These two stages can be iteratively performed for better CNN-TCN refinement. In this work, we propose a new method to explore the alignment between visual segment features from CNN-TCN and the corresponding sign glosses via dynamic programming.

We abbreviate $\boldsymbol{Y}^{(k)}$ to $\boldsymbol{Y}$ for simplification. As for the probability matrix $\boldsymbol{Y}$ calculated from Equation (10), we can re-organize $\boldsymbol{Y}$ with the ground-truth sign gloss sequence. For each sign gloss in the ground-truth label, we orderly pick up the probabilities of all frames/segments corresponding to the current sign gloss to form a new probability matrix $\boldsymbol{Y}^{'}$, illustrated in Figure 5. Based on $\boldsymbol{Y}^{'}$, we perform dynamic programming to find the alignment path with maximum decoding probability. With $P_{i,j}$ standing for the maximum decoding probability of subsequence $\{f_1, f_2, \cdots, f_i\}$ for visual features and subsequence $\{s_1, s_2, \cdots, s_j\}$ for sign glosses, the state transition equation of dynamic programming is defined as follows,

$$P_{i,j} = \boldsymbol{Y}_{i,j}^{'} + \max(P_{i-1,j}, P_{i-1,j-1}). \tag{14}$$

After progressive calculation over full visual features and glosses sequence, the clip-level alignments are derived and they serve as the pseudo labels to fine-tune the visual backbone. Although dynamic programming cannot ensure the maximum decoding result of each clip during alignment extraction, it can find an alignment path, of which the overall decoding probability sum is the largest and better fertilizes the visual encoder during fine-tuning.

IV. Experiments

In this section, we perform extensive experiments to evaluate our method. First, the datasets and evaluation metrics are introduced. Then we perform ablation studies on the effectiveness of modules in our architecture. Finally, we make comparisons with existing state-of-the-art methods.

Authorized licensed use limited to: University of Science & Technology of China. Downloaded on June 03,2023 at 05:01:09 UTC from IEEE Xplore. Restrictions apply.
© 2022 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2022.3223260

TABLE I: Statistical data on RWTH-PHOENIX-Weather, CSL and GSL-SD datasets.

Statistics	RWTH-PHOENIX-Weather			CSL		GSL-SD		
	Train	Dev	Test	Train	Test	Train	Dev	Test
#frames	799,006	75,186	89,472	963,228	66,529	781,414	140,138	114,603
#duration (h)	8.88	0.84	0.99	10.70	0.74	7.24	1.30	1.06
#vocabulary	1,231	460	496	178	20	310	310	310
#videos	5,672	540	629	4,700	300	8,189	1,063	1,043
#signers	9	9	9	50	50	7	7	7
out-of-vocab (%)	-	0.69	0.69	-	0	-	0	0

A. Dataset and Evaluation

We conduct experiments on three public datasets, *i.e.,* RWTH-PHOENIX-Weather multi-signer for German SLR [22] and CSL [42] for Chinese SLR and GSL-SD [67] for Greek SLR, respectively. RWTH-PHOENIX-Weather [22] dataset is one of the most popular benchmarks for continuous SLR. It provides RGB videos with corresponding annotations. All videos are recorded at 25 frames per second (FPS) with the resolution of 210×260. The videos contain 6,841 sentences performed by 9 signers with a vocabulary size of 1295. The dataset is divided into three subsets, *i.e.*, 5,672 instances for training, 540 for validation and 629 for testing. CSL [42] dataset contains 178 sign words in daily communication. 100 different sentences are performed by 50 signers, with totally 5,000 videos. The dataset is divided into two subsets, 4,700 videos for training and 300 for testing. GSL-SD dataset [67] is performed by 7 signers. The whole video instances are divided into three subsets, 8,189 for training, 1,063 for validation and 1,043 for testing. The detailed statistical data are summarized in Table I.

There exist many metrics for performance evaluation of continuous SLR. As a common metric, WER (Word Error Rate) is defined based on essentially an edit distance, which indicates the least operations of substitution, deletion and insertion to transform the predicted sentence into the reference sequence:

$$WER = \frac{n_i + n_d + n_s}{L}, \tag{15}$$

where n_i, n_d, and n_s are the number of operations for insertion, deletion, and substitution, respectively. L is the length of the reference sequence. Besides, we calculate the ratio of correct words to the reference words in the predicted sentence, denoted as Acc-w.

We also adopt semantics evaluation metrics widely used in NLP, NMT, and image description evaluation, *i.e.,* BLEU [68], CIDEr [69], ROUGE-L [70], and METEOR [71]. **BLEU** (Bilingual evaluation understudy) is one of the most popular metrics, which computes the modified precision metric using n-grams to measure the quality of machine-generated text. **CIDEr** (Consensus-based Image Description Evaluation) is a novel consensus-based evaluation protocol, which measures the similarity of a sentence to the majority, or the consensus of how most people describe. **ROUGE-L** (Recall-Oriented Understudy for Gisting Evaluation-LCS) measures sentence-level structure similarity naturally and identifies the longest co-occurring in sequence n-grams automatically. **METEOR** (Metric for Evaluation of Translation with Explicit ORdering) is based on the harmonic mean of unigram precision and recall, with recall weighted higher than precision.

B. Implementation Details

In our experiment, data augmentation is crucial for relieving over-fitting. We first randomly crop the video at the same spatial location across all frames, perform random horizontal flipping spatially, and temporally discard 20% of frames at random. The input size is 224×224. Besides, the hidden states of the 2-layer BLSTM are all set to 1024.

Following the previous methods [6], [7], we use a staged optimization strategy. First, the employed GoogLeNet backbone is pre-trained on ImageNet [72]. At the end-to-end training stage, the whole framework is trained end-to-end using the loss in Equation (13) for each language branch. During this stage, we use Adam optimizer with the learning rate of 5×10^{-5} and batch size of 3.

In continuous SLR, it is crucial for the visual encoder to produce discriminative feature representation. In the first stage, the CTC objective loss has limited contribution to low layers of visual encoder due to the vanishing gradient problem. Thus the visual encoder may not be fully optimized. To alleviate this issue, we explicitly train the visual encoder using the common classification method, which introduces the second refinement stage. We adopt the pre-trained weight from the first stage to decode pseudo labels for each clip using the method introduced in Section III-D. Then we add a fully-connected layer on top of the visual encoder to predict the corresponding class of each clip. It is trained with cross-entropy loss by stochastic gradient descent (SGD) optimizer.

We train the visual encoder for 35 epochs. The initial learning rate and weight decay are set to 5e-3 and 1e-4, respectively. The batch size is set to 32 and the clip size is 16. After that, initialized with pre-trained parameters from the refinement stage, the whole network is trained end-to-end as stated in the first stage. Through this iterative staged optimization strategy, our visual encoder is well trained and can cooperate with the sequential module for better performance. Our entire architecture is implemented by PyTorch and all the experiments run on NVIDIA Tesla V100.

C. Ablation Study

We perform an ablation study on the effectiveness of each proposed module in our framework. The ablation study is conducted on two benchmarks, *i.e.,* RWTH-PHOENIX-Weather and CSL datasets.

Authorized licensed use limited to: University of Science & Technology of China. Downloaded on June 03,2023 at 05:01:09 UTC from IEEE Xplore. Restrictions apply.
© 2022 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2022.3223260

TABLE II: Performance of different basic architectures on the RWTH-PHOENIX-Weather and CSL datasets (the lower the better). 'del' and 'ins' denote deletion and insertion, respectively.

Methods	RWTH-PHOENIX-Weather				CSL	
	Dev		Test		Test	
	del / ins	WER	del / ins	WER	del / ins	WER
Individual	7.8 / 3.5	23.8	7.8 / 3.4	24.4	9.8 / 0.0	31.9
Individual (pre-trained)	9.6 / 2.2	23.3	9.7 / 2.0	24.1	9.6 / 0.3	30.2
All shared (basic model a)	8.9 / 3.3	25.2	8.3 / 3.0	25.1	7.3 / 0.5	30.1
Ours (basic model b)	8.4 / 2.6	23.1	8.1 / 2.6	23.9	15.3 / 0.1	26.4
Ours (basic model c)	7.9 / 2.8	**22.7**	7.6 / 2.7	**23.6**	15.3 / 0.4	**20.2**

TABLE III: Ablation study on the effectiveness of our proposed refinement method on the RWTH-PHOENIX-Weather and CSL datasets (the lower the better).

Methods	Refinement			RWTH-PHOENIX-Weather				CSL	
				Dev		Test		Test	
	None	CTC-align	Max-prob (Ours)	del / ins	WER	del / ins	WER	del / ins	WER
Basic model b	✓			8.4 / 2.6	23.1	8.1 / 2.6	23.9	15.3 / 0.1	26.4
Basic model b		✓		8.4 / 2.7	22.5	8.0 / 2.3	23.1	14.9 / 0.1	21.5
Basic model b			✓	8.2 / 2.6	**21.0**	7.8 / 2.3	**21.7**	14.9 / 0.3	**19.7**
Basic model c	✓			7.9 / 2.8	22.7	7.6 / 2.7	23.6	15.3 / 0.4	20.2
Basic model c		✓		8.0 / 3.1	22.4	7.0 / 2.8	22.8	11.0 / 0.5	20.1
Basic model c			✓	7.0 / 2.8	**20.4**	6.9 / 2.8	**21.4**	14.2 / 0.4	**19.1**

Evaluation on basic architecture setting. As shown in Table II, we compare the effectiveness of different basic architectures for multilingual SLR. The first row shows the single lingual backbone training and testing on RWTH-PHOENIX-Weather and CSL dataset, respectively. 'Individual (pre-trained)' corresponds to the single lingual model first pre-trained on other datasets and then fine-tuned for each type of sign language. This setting is designed for fair comparison with multilingual settings and we pre-train the framework on other datasets to keep their training data scale consistent. To be specific, the model is pre-trained on CSL and fine-tuned on RWTH-PHOENIX-Weather dataset, and vice versa.

The basic model a indicates that we use both a shared visual encoder and a shared sequential module, corresponding to Figure 2a. Its WER results are inferior to the 'Individual (pre-trained)' on both benchmarks. This is due to the fact that visually similar words in different languages may disturb each other, and thus the shared sequential module may not be fully optimized for each language. The fourth row shows that an independent sequential module is utilized for each language, corresponding to Figure 2b. Benefiting from the enlarged training corpus data, the shared visual encoder produces robust feature representations, while separated sequential modules relieve the disturbance between different languages, which leads to better performance.

Compared with 'Ours (basic model b)', the basic model c contains a shared sequential module to transfer common linguistic rules across different languages, corresponding to Figure 2c. This architecture achieves the best performance on all benchmarks against others. Specifically, our model achieves 20.2% WER on CSL, contrasting the 30.2% WER of the Individual (pre-trained), which validates the effectiveness of our subtly designed framework.

Evaluation on refinement. Table III shows the performance with the refinement based on our proposed max-probability alignment and CTC alignment which is proposed in [6], [64]. It can be observed that a notable improvement has been made on both datasets after refinement, as shown in the third and last row. Furthermore, our max-probability alignment outperforms the CTC alignment by a large margin. Besides, the refinement module can be generally applied to different architectures proposed in our work. To further demonstrate the effectiveness of our refinement method, we visualize one sample in each benchmark dataset, as shown in Figure 6. The sample of RWTH-PHOENIX-Weather indicates that the baseline method fails to detect some sign words, such as 'DANN' and 'BISSCHEN'. With our architecture added, some missing words are complemented, but with a wrong substitution of the word 'AUCH'. Further, this wrong substitution can be corrected with refinement. A similar phenomenon is also observed in the example of CSL dataset. The baseline method misses several keywords, such as 'sister' and 'nurse', and our method fails to offer any complement, if no refinement is added. After further refinement introduced, however, all the mistakes get corrected. Such phenomenon can be explained as follows. With more data engaged in the training process, the visual encoder generates more robust feature representation. Besides, similar visual patterns are able to get enhanced among different languages, which facilitates the entire architecture's capability to capture missing words. With our proposed alignment method, the mappings between visual patterns and words are further refined, leading to the correction of wrong words in the sentence.

D. Comparison with the State-of-the-art Methods

We compare our approach with the existing state-of-the-art methods on three public datasets, *i.e.,* RWTH-PHOENIX-Weather, CSL and GSL-SD. All 3 languages are involved during training, including German, Chinese and Greek SL.

Evaluation on RWTH-PHOENIX-Weather. As shown in Table IV, we compare our approach with other methods on RWTH-PHOENIX-Weather. CMLLR [22] and 1-Million-Hand [73] utilize hand-crafted features with a traditional

Authorized licensed use limited to: University of Science & Technology of China. Downloaded on June 03,2023 at 05:01:09 UTC from IEEE Xplore. Restrictions apply.
© 2022 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2022.3223260

(a) An example on the the dev set of RWTH-PHOENIX-Weather dataset.

(b) An example on the the test set of CSL dataset.

Fig. 6: Qualitative illustrations of the effectiveness of the refinement method on two benchmark datasets. In each sub-figure, the first row shows the corresponding video sequences. The following three rows indicate the CTC predictions of the baseline method, our method without refinement and ours with refinement, respectively. The last three rows show the final generated sentence in the same orders as the CTC predictions. Note that the red boxes denote the failure cases. 'SUB' and 'DEL' indicate substitution and deletion, respectively.

HMM-based model. In SubUNets [41] and CNN-LSTM-HMM [4], hand patches are used as a cue to assist the full-frame RGB modality. However, it leads to doubled model sizes. FCN [46] utilizes fully convolutional networks, while in [74], it utilizes Transformer as the backbone. In our setting, we only utilize the full frame and our method achieves 20.3% and 20.9% WER on the dev and test set, which is a new state-of-the-art result on RWTH-PHOENIX-Weather.

Evaluation on CSL. As discussed in the works [42], [63], CSL dataset originally provides 2 splits for evaluation. For Split II, it is an unseen sentence setting: the training and testing sets share the same signers but without overlap of the same sentences. Generally, it is much more difficult for SLR on unseen sentences (Split II). Therefore, we evaluate our approach on this split, as shown in Table V. We compare our method with LSTM&CTC [47], S2VT [79], HLSTM [63], HRF [80], IAN [7], SL-Transformer [74] and CMA [45]. HLSTM proposes a hierarchical-LSTM (HLSTM) encoder-decoder model with visual content and word embedding. It also utilizes the temporal attention mechanism to balance the intrinsic relationship. In CMA, it exhibits a significant improvement as pseudo-video-text pairs are introduced. Even so, our method still largely outperforms CMA by 5.4% on WER. Besides, the results of the precision and semantic metrics of our method also show consistent improvement over the best competitor, *e.g.,* 4.8% improvement on Acc-w, *etc.* With lingual characteristics learned from different languages, the performance on the dataset which has a relatively small vocabulary size gets boosted more substantially.

Evaluation on GSL-SD. As shown in Table VI, DNF (RGB) [6] utilizes the RGB full frame as the input modality, which is our baseline method. Our method achieves 32.7% and 33.5% WER on the dev and test set, respectively, which is new state-of-the-art performance. These experimental results demonstrate the generalization capability of our method among different sign languages.

Authorized licensed use limited to: University of Science & Technology of China. Downloaded on June 03,2023 at 05:01:09 UTC from IEEE Xplore. Restrictions apply.
© 2022 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2022.3223260

TABLE IV: Evaluation on RWTH-PHOENIX-Weather (the lower the better).

Methods	Data				Dev		Test	
	Full	Hand	Face	Pose	del / ins	WER	del / ins	WER
CMLLR [22]		✓	✓	✓	21.8 / 3.9	55.0	20.3 / 4.5	53.0
1-Million-Hand [73]		✓	✓	✓	16.3 / 4.6	47.1	15.2 / 4.6	45.1
CNN-Hybrid [75]		✓			12.6 / 5.1	38.3	11.1 / 5.7	38.8
SubUNets [41]	✓	✓			14.6 / 4.0	40.8	14.3 / 4.0	40.7
Re-sign [40]	✓				-	27.1	-	26.8
RCNN [9]		✓			13.7/7.3	39.4	12.2/7.5	38.7
Hybrid CNN-HMM [10]		✓				31.6		32.5
Dilated [43]	✓				8.3 / 4.8	38.0	7.6 / 4.8	37.3
CTF [76]	✓				12.8 / 5.2	37.9	11.9 / 5.6	37.8
CNN-LSTM-HMM [4]	✓	✓			-	26.0	-	26.0
IAN [7]	✓				12.9 / 2.6	37.1	13.0 / 2.5	36.7
DNF (RGB) [6]	✓				7.8 / 3.5	23.8	7.8 / 3.4	24.4
SMC [65]	✓	✓	✓	✓	7.8 / 3.8	22.7	7.4 / 3.5	22.4
FCN [46]	✓				-	23.7	-	23.9
CMA [45]	✓				7.3 / 2.7	21.3	7.3 / 2.4	21.9
SFLM [77]	✓				10.3 / 4.1	24.9	10.4 / 3.6	25.3
SL-Transformer [74]	✓				5.8 / 4.7	23.1	5.4 / 4.6	24.2
CMJLS [31]	✓				-	23.9	-	24.0
VAC [78]	✓				7.9 / 2.5	21.2	8.4 / 2.6	22.3
Ours	✓				6.4 / 2.9	**20.3**	6.2 / 2.8	**20.9**

TABLE V: Evaluation on CSL dataset according to common semantic evaluation metrics. (↑ indicates the higher the better, while ↓ indicates the lower the better.)

Methods	Acc-w ↑	BLEU-1 ↑	BLEU-2 ↑	BLEU-3 ↑	CIDEr ↑	ROUGE-L ↑	METEOR ↑	WER ↓
LSTM&CTC [47], [48]	0.332	0.343	0.124	0.039	0.241	0.362	0.111	0.757
S2VT [79]	0.457	0.466	0.258	0.135	0.479	0.461	0.189	0.670
S2VT (3-layer) [79]	0.461	0.475	0.265	0.145	0.477	0.465	0.186	0.652
HLSTM (SYS sampling) [63]	0.459	0.463	0.293	0.185	0.476	0.462	0.173	0.630
HLSTM [63]	0.482	0.487	0.315	0.195	0.561	0.481	0.193	0.662
HLSTM-attn [63]	0.506	0.508	0.330	0.207	0.605	0.503	0.205	0.641
HRF-Fusion [80]	0.445	0.450	0.238	0.127	0.398	0.449	0.171	0.672
IAN [7]	0.670	0.724	-	-	3.946	0.716	0.383	0.327
SL-Transformer [74]	0.661	0.694	0.504	0.398	1.992	0.702	0.331	0.335
CMA [45]	0.747	0.784	-	-	3.006	0.782	0.390	0.245
Ours	**0.809**	**0.852**	**0.779**	**0.743**	**5.799**	**0.843**	**0.481**	**0.181**

TABLE VI: Evaluation on GSL-SD (the lower the better).

Methods	Dev		Test	
	del / ins	WER	del / ins	WER
SubUNets [29]	-	52.8	-	54.3
IAN [7]	-	61.9	-	68.5
DNF (RGB) [6]	-	43.5	-	48.5
Ours	12.5 / 2.2	**32.7**	12.5 / 2.7	**33.5**

V. Conclusion & Future Work

Current solutions to sign language recognition (SLR) treat each language independently and learn separate SLR models. To the best of our knowledge, we are the *first* to explore the multilingual sign language recognition topic. To this end, we propose a unified framework, which consists of a shared visual encoder, and an independent sequential module for each language together with a shared sequential module. The shared visual encoder and shared sequential module benefit from large training data of different languages and are able to promote each independent module for its corresponding language task. Besides, a max-probability decoding scheme is proposed to align the videos and sign glosses for further visual encoder refinement. Extensive experiments have validated the effectiveness of our method, which achieves new state-of-the-art performances on both three challenging benchmarks, *i.e.,* RWTH-PHOENIX-Weather, CSL and GSL-SD datasets.

The multilingual problem is a new research topic of great importance to the community. We argue that multilingual sign languages share common language-agnostic visual patterns. In other words, it is beneficial to explore the collaborative representation learning paradigm under this insight. There are several potential directions for further research. Firstly, since annotated sign data requires expert knowledge, the data scale of the current sign corpus is limited. It is meaningful to design a self-supervised learning framework by leveraging a large amount of multilingual sign data without annotation. With this kind of data involved, the recognition performance is expected to further boost. Secondly, it is worth exploring more advanced unified models which can automatically recognize the input sign video to its corresponding sign language. Besides, it is also desirable to study the interpretability, which is conducive to the analysis of the correlation among multilingual data, such as the common language-agnostic and language-specific components in different languages, which will help the sign community understand sign language better.

Acknowledgment. This work was supported by the National Natural Science Foundation of China under Contract U20A20183 and 62021001. It was also supported by GPU cluster built by MCC Lab of Information Science and Technology Institution, USTC.

Authorized licensed use limited to: University of Science & Technology of China. Downloaded on June 03,2023 at 05:01:09 UTC from IEEE Xplore. Restrictions apply.
© 2022 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2022.3223260

REFERENCES

[1] Z. Ren, J. Yuan, J. Meng, and Z. Zhang, "Robust part-based hand gesture recognition using Kinect sensor," *IEEE Transactions on Multimedia (TMM)*, vol. 15, no. 5, pp. 1110–1120, 2013.

[2] C. Wang, Z. Liu, and S.-C. Chan, "Superpixel-based hand gesture recognition with Kinect depth camera," *IEEE Transactions on Multimedia (TMM)*, vol. 17, no. 1, pp. 29–39, 2014.

[3] S. Tornay, M. Razavi, and M. M. Doss, "Towards multilingual sign language recognition," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6309–6313.

[4] O. Koller, C. Camgoz, H. Ney, and R. Bowden, "Weakly supervised learning with multi-stream CNN-LSTM-HMMs to discover sequential parallelism in sign language videos," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2019.

[5] H. Wang, X. Chai, and X. Chen, "A novel sign language recognition framework using hierarchical Grassmann covariance matrix," *IEEE Transactions on Multimedia (TMM)*, vol. 21, no. 11, pp. 2806–2814, 2019.

[6] R. Cui, H. Liu, and C. Zhang, "A deep neural framework for continuous sign language recognition by iterative training," *IEEE Transactions on Multimedia (TMM)*, vol. 21, no. 7, pp. 1880–1891, 2019.

[7] J. Pu, W. Zhou, and H. Li, "Iterative alignment network for continuous sign language recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 4165–4174.

[8] F. Yin, X. Chai, and X. Chen, "Iterative reference driven metric learning for signer independent isolated sign language recognition," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016, pp. 434–450.

[9] R. Cui, H. Liu, and C. Zhang, "Recurrent convolutional neural networks for continuous sign language recognition by staged optimization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 7361–7369.

[10] O. Koller, S. Zargaran, H. Ney, and R. Bowden, "Deep sign: Enabling robust statistical continuous sign language recognition via hybrid CNN-HMMs," *International Journal of Computer Vision (IJCV)*, vol. 126, no. 12, pp. 1311–1325, 2018.

[11] Z. Liu, X. Qi, and L. Pang, "Self-boosted gesture interactive system with st-net," in *Proceedings of the ACM International Conference on Multimedia (ACM MM)*, 2018, p. 145–153.

[12] D. Li, C. Xu, X. Yu, K. Zhang, B. Swift, H. Suominen, and H. Li, "Tspnet: Hierarchical feature learning via temporal semantic pyramid for sign language translation," *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 12 034–12 045, 2020.

[13] G. Varol, L. Momeni, S. Albanie, T. Afouras, and A. Zisserman, "Read and attend: Temporal localisation in sign language videos," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 16 857–16 866.

[14] T. Jin and Z. Zhao, "Contrastive disentangled meta-learning for signer-independent sign language translation," in *Proceedings of the ACM International Conference on Multimedia (ACM MM)*, 2021, pp. 5065–5073.

[15] B. Shi, D. Brentari, G. Shakhnarovich, and K. Livescu, "Fingerspelling detection in american sign language," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 4166–4175.

[16] L. Hu, L. Gao, Z. Liu, and W. Feng, "Temporal lift pooling for continuous sign language recognition," in *European Conference on Computer Vision (ECCV)*, 2022, pp. 511–527.

[17] Y. Chen, R. Zuo, F. Wei, Y. Wu, S. Liu, and B. Mak, "Two-stream network for sign language recognition and translation," 2022, pp. 1–18.

[18] P. Xie, M. Zhao, and X. Hu, "Pisltrc: Position-informed sign language transformer with content-aware convolution," *IEEE Transactions on Multimedia (TMM)*, vol. 24, pp. 3908–3919, 2022.

[19] H. Hu, W. Zhao, W. Zhou, Y. Wang, and H. Li, "SignBERT: Pre-training of hand-model-aware representation for sign language recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 11 087–11 096.

[20] L.-C. Wang, R. Wang, D.-H. Kong, and B.-C. Yin, "Similarity assessment model for Chinese sign language videos," *IEEE Transactions on Multimedia (TMM)*, vol. 16, no. 3, pp. 751–761, 2014.

[21] P. Buehler, A. Zisserman, and M. Everingham, "Learning sign language by watching TV (using weakly aligned subtitles)," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009, pp. 2961–2968.

[22] O. Koller, J. Forster, and H. Ney, "Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers," *Computer Vision and Image Understanding (CVIU)*, vol. 141, pp. 108–125, 2015.

[23] T. Pfister, J. Charles, and A. Zisserman, "Large-scale learning of sign language by watching TV (using co-occurrences)," in *Proceedings of the British Machine Vision Conference (BMVC)*, 2013.

[24] G. D. Evangelidis, G. Singh, and R. Horaud, "Continuous gesture recognition from articulated poses," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2014, pp. 595–607.

[25] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 2014, pp. 568–576.

[26] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool, "Temporal segment networks: Towards good practices for deep action recognition," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016, pp. 20–36.

[27] W. Huang, L. Fan, M. Harandi, L. Ma, H. Liu, W. Liu, and C. Gan, "Toward efficient action recognition: Principal backpropagation for training two-stream networks," *IEEE Transactions on Image Processing (TIP)*, vol. 28, pp. 1773–1782, 2018.

[28] H. Hu, W. Zhou, X. Li, N. Yan, and H. Li, "Mv2flow: Learning motion representation for fast compressed video action recognition," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 16, no. 3s, pp. 1–19, 2020.

[29] H. Hu, W. Wang, W. Zhou, W. Zhao, and H. Li, "Model-aware gesture-to-gesture translation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 16 428–16 437.

[30] H. Yang, C. Yuan, L. Zhang, Y. Sun, W. Hu, and S. J. Maybank, "STA-CNN: Convolutional spatial-temporal attention learning for action recognition," *IEEE Transactions on Image Processing (TIP)*, vol. 29, pp. 5783–5793, 2020.

[31] I. Papastratis, K. Dimitropoulos, D. Konstantinidis, and P. Daras, "Continuous sign language recognition through cross-modal alignment of video and text embeddings in a joint-latent space," *IEEE Access*, vol. 8, pp. 91 170–91 180, 2020.

[32] H. Hu, W. Zhou, and H. Li, "Hand-model-aware sign language recognition," in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2021, pp. 1558–1566.

[33] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 4489–4497.

[34] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 6299–6308.

[35] Z. Qiu, T. Yao, and T. Mei, "Learning spatio-temporal representation with pseudo-3D residual networks," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 5533–5541.

[36] Z. Qiu, T. Yao, C.-W. Ngo, X. Tian, and T. Mei, "Learning spatio-temporal representation with local and global diffusion," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 12 056–12 065.

[37] H. Hu, W. Zhou, J. Pu, and H. Li, "Global-local enhancement network for nmf-aware sign language recognition," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 17, no. 3, pp. 1–19, 2021.

[38] S. Xie, C. Sun, J. Huang, Z. Tu, and K. Murphy, "Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 305–321.

[39] M. Zolfaghari, K. Singh, and T. Brox, "ECO: Efficient convolutional network for online video understanding," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 695–712.

[40] O. Koller, S. Zargaran, and H. Ney, "Re-sign: Re-aligned end-to-end sequence modelling with deep recurrent CNN-HMMs," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 4297–4305.

[41] N. C. Camgoz, S. Hadfield, O. Koller, and R. Bowden, "SubUNets: End-to-end hand shape and continuous sign language recognition," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 3075–3084.

[42] J. Huang, W. Zhou, Q. Zhang, H. Li, and W. Li, "Video-based sign language recognition without temporal segmentation," in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2018.

Authorized licensed use limited to: University of Science & Technology of China. Downloaded on June 03,2023 at 05:01:09 UTC from IEEE Xplore. Restrictions apply.
© 2022 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2022.3223260

[43] J. Pu, W. Zhou, and H. Li, "Dilated convolutional network with iterative optimization for continuous sign language recognition," in *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2018, pp. 885–891.

[44] Z. Yang, Z. Shi, X. Shen, and Y.-W. Tai, "SF-Net: Structured feature network for continuous sign language recognition," *arXiv preprint arXiv:1908.01341*, 2019.

[45] J. Pu, W. Zhou, H. Hu, and H. Li, "Boosting continuous sign language recognition via cross modality augmentation," in *Proceedings of the ACM International Conference on Multimedia (ACM MM)*, 2020, pp. 1497–1505.

[46] K. L. Cheng, Z. Yang, Q. Chen, and Y.-W. Tai, "Fully convolutional networks for continuous sign language recognition," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[47] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2006, pp. 369–376.

[48] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[49] Y. Min, Y. Zhang, X. Chai, and X. Chen, "An efficient PointLSTM for point clouds based gesture recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 5761–5770.

[50] K. Cho, B. Van Merriënboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder-decoder approaches," *arXiv preprint arXiv:1409.1259*, 2014.

[51] A. Graves and N. Jaitly, "Towards end-to-end speech recognition with recurrent neural networks," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2014, pp. 1764–1772.

[52] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[53] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 2014, pp. 3104–3112.

[54] X. Liu, H. Shi, H. Chen, Z. Yu, X. Li, and G. Zhao, "imigue: An identity-free video dataset for micro-gesture understanding and emotion analysis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 10 631–10 642.

[55] Y. Zhu and S. Jiang, "Attention-based densely connected LSTM for video captioning," in *Proceedings of the ACM International Conference on Multimedia (ACM MM)*, 2019, pp. 802–810.

[56] X. Shi, J. Cai, S. Joty, and J. Gu, "Watch it twice: Video captioning with a refocused video encoder," in *Proceedings of the ACM International Conference on Multimedia (ACM MM)*, 2019, pp. 818–826.

[57] S. Yang, L. Li, S. Wang, D. Meng, Q. Huang, and Q. Tian, "Structured stochastic recurrent network for linguistic video prediction," in *Proceedings of the ACM International Conference on Multimedia (ACM MM)*, 2019, pp. 21–29.

[58] Y. Xu, Y. Han, R. Hong, and Q. Tian, "Sequential video VLAD: Training the aggregation locally and temporally," *IEEE Transactions on Image Processing (TIP)*, vol. 27, pp. 4933–4944, 2018.

[59] F. Tao and C. Busso, "End-to-end audiovisual speech recognition system with multitask learning," *IEEE Transactions on Multimedia (TMM)*, 2020.

[60] J. Zhang, J. Du, and L. Dai, "Track, attend, and parse (tap): An end-to-end framework for online handwritten mathematical expression recognition," *IEEE Transactions on Multimedia (TMM)*, vol. 21, no. 1, pp. 221–233, 2018.

[61] D.-A. Huang, L. Fei-Fei, and J. C. Niebles, "Connectionist temporal modeling for weakly supervised action labeling," in *European Conference on Computer Vision (ECCV)*, 2016, pp. 137–153.

[62] X. Liu, H. Shi, X. Hong, H. Chen, D. Tao, and G. Zhao, "3D skeletal gesture recognition via hidden states exploration," *IEEE Transactions on Image Processing (TIP)*, vol. 29, pp. 4583–4597, 2020.

[63] D. Guo, W. Zhou, H. Li, and M. Wang, "Hierarchical LSTM for sign language translation," in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2018.

[64] H. Zhou, W. Zhou, and H. Li, "Dynamic pseudo label decoding for continuous sign language recognition," in *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, 2019, pp. 1282–1287.

[65] H. Zhou, W. Zhou, Y. Zhou, and H. Li, "Spatial-Temporal multi-cue network for continuous sign language recognition." in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2020, pp. 13 009–13 016.

[66] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 1–9.

[67] N. Adaloglou, T. Chatzis, I. Papastratis, A. Stergioulas, G. T. Papadopoulos, V. Zacharopoulou, G. J. Xydopoulos, K. Atzakas, D. Papazachariou, and P. Daras, "A comprehensive study on deep learning-based methods for sign language recognition," *IEEE Transactions on Multimedia (TMM)*, vol. 24, pp. 1750–1762, 2022.

[68] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: a method for automatic evaluation of machine translation," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2002, pp. 311–318.

[69] R. Vedantam, C. Lawrence Zitnick, and D. Parikh, "CIDEr: Consensus-based image description evaluation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 4566–4575.

[70] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Text summarization branches out: Proceedings of the ACL-04 workshop*, 2004, pp. 74–81.

[71] S. Banerjee and A. Lavie, "METEOR: An automatic metric for mt evaluation with improved correlation with human judgments," in *Proceedings of the ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 2005, pp. 65–72.

[72] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "ImageNet large scale visual recognition challenge," *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015.

[73] O. Koller, H. Ney, and R. Bowden, "Deep hand: How to train a CNN on 1 million hand images when your data is continuous and weakly labelled," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 3793–3802.

[74] N. C. Camgoz, O. Koller, S. Hadfield, and R. Bowden, "Sign language transformers: Joint end-to-end sign language recognition and translation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 10 023–10 033.

[75] O. Koller, O. Zargaran, H. Ney, and R. Bowden, "Deep sign: Hybrid CNN-HMM for continuous sign language recognition," in *Proceedings of the British Machine Vision Conference (BMVC)*, 2016.

[76] S. Wang, D. Guo, W.-g. Zhou, Z.-J. Zha, and M. Wang, "Connectionist temporal fusion for sign language translation," in *Proceedings of the ACM International Conference on Multimedia (ACM MM)*, 2018, pp. 1483–1491.

[77] Z. Niu and B. Mak, "Stochastic fine-grained labeling of multi-state sign glosses for continuous sign language recognition," in *European Conference on Computer Vision (ECCV)*, 2020, pp. 172–186.

[78] Y. Min, A. Hao, X. Chai, and X. Chen, "Visual alignment constraint for continuous sign language recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 11 542–11 551.

[79] S. Venugopalan, M. Rohrbach, J. Donahue, R. Mooney, T. Darrell, and K. Saenko, "Sequence to sequence-video to text," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 4534–4542.

[80] D. Guo, W. Zhou, A. Li, H. Li, and M. Wang, "Hierarchical recurrent deep fusion using adaptive clip summarization for sign language translation," *IEEE Transactions on Image Processing (TIP)*, vol. 29, pp. 1575–1590, 2019.

Hezhen Hu Hezhen Hu is currently pursuing the Ph.D. degree in information and communication engineering with the Department of Electronic Engineering and Information Science, from the University of Science and Technology of China (USTC). His research interests include sign language understanding, self-supervised pre-training, human-centric visual understanding, and computer vision.

Authorized licensed use limited to: University of Science & Technology of China. Downloaded on June 03,2023 at 05:01:09 UTC from IEEE Xplore. Restrictions apply.
© 2022 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2022.3223260

Junfu Pu Junfu Pu received the B.E. degree in Electronic Information Engineering and the Ph.D degree in Information and Communication Engineering from the University of Science and Technology of China (USTC) in 2015 and 2020, respectively. He is a senior researcher at ARC Lab (Applied Research Center), Tencent. His research interests include sign language recognition/translation/generation, multimedia understanding and applications, vision-language pretraining and search.

Wengang Zhou Wengang Zhou (S'20) received the B.E. degree in electronic information engineering from Wuhan University, China, in 2006, and the Ph.D. degree in electronic engineering and information science from the University of Science and Technology of China (USTC), China, in 2011. From September 2011 to September 2013, he worked as a postdoc researcher in Computer Science Department at the University of Texas at San Antonio. He is currently a Professor at the EEIS Department, USTC.

His research interests include multimedia information retrieval, computer vision, and computer game. In those fields, he has published over 100 papers in IEEE/ACM Transactions and CCF Tier-A International Conferences. He is the winner of National Science Funds of China (NSFC) for Excellent Young Scientists. He is the recepient of the Best Paper Award for ICIMCS 2012. He received the award for the Excellent Ph.D Supervisor of Chinese Society of Image and Graphics (CSIG) in 2021, and the award for the Excellent Ph.D Supervisor of Chinese Academy of Sciences (CAS) in 2022. He won the First Class Wu-Wenjun Award for Progress in Artificial Intelligence Technology in 2021. He served as the publication chair of IEEE ICME 2021 and won 2021 ICME Outstanding Service Award. He is currently an Associate Editor and a Lead Guest Editor of IEEE Transactions on Multimeida.

Houqiang Li Houqiang Li (S'12, F'21) received the B.S., M.Eng., and Ph.D. degrees in electronic engineering from the University of Science and Technology of China, Hefei, China, in 1992, 1997, and 2000, respectively, where he is currently a Professor with the Department of Electronic Engineering and Information Science.

His research interests include image/video coding, image/video analysis, computer vision, reinforcement learning, etc.. He has authored and co-authored over 200 papers in journals and conferences. He is the winner of National Science Funds (NSFC) for Distinguished Young Scientists, the Distinguished Professor of Changjiang Scholars Program of China, and the Leading Scientist of Ten Thousand Talent Program of China. He is the associate editor (AE) of IEEE TMM, and served as the AE of IEEE TCSVT from 2010 to 2013. He served as the General Co-Chair of ICME 2021 and the TPC Co-Chair of VCIP 2010. He received the second class award of China National Award for Technological Invention in 2019, the second class award of China National Award for Natural Sciences in 2015, and the first class prize of Science and Technology Award of Anhui Province in 2012. He received the award for the Excellent Ph.D Supervisor of Chinese Academy of Sciences (CAS) for four times from 2013 to 2016. He was the recipient of the Best Paper Award for VCIP 2012, the recipient of the Best Paper Award for ICIMCS 2012, and the recipient of the Best Paper Award for ACM MUM in 2011.

Authorized licensed use limited to: University of Science & Technology of China. Downloaded on June 03,2023 at 05:01:09 UTC from IEEE Xplore. Restrictions apply.
© 2022 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.