Proceedings of the 2nd International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/81/20241162

A New Sign Language Translation System Based on Expert Model

Jinbang Wu

Sichuan University of Science & Engineering, No. 1 Baita Road, Sanjiang New District, Yibin City, Sichuan Province, China

55327116@qq.com

Abstract. The linguistic diversity of sign languages across regions presents significant challenges for the development of accurate and culturally sensitive recognition systems. To address this, This paper propose a modular sign language recognition system that detects the regional origin of a sign language and uses specialized sub-models for translation, bypassing the limitations of generalized models. This approach allows for more precise translations by catering to regional variations, ensuring higher accuracy and preserving cultural nuances.This paper also explore the creation of a standardized, diverse dataset, integrating multiple sign languages, which serves as the foundation for the system's region detection process. The dataset was meticulously annotated, normalized, and augmented to support robust model training. By adopting sub-models and utilizing region-specific data, the system improves efficiency and scalability, offering a practical solution for real-world sign language translation. This work emphasizes the importance of preserving sign language diversity while enhancing the usability and adaptability of recognition systems.

Keyword: Sign Language Recognition, Region Detection, Sub-model Translation, Dataset Standardization, Cultural Diversity in Sign Language.

1. Introduction

Nowadays, there is a huge group of sign language users around the world. These people suffering a nature obstacle in communicating with other normal people. The challenge is how to make the various gestures and trajectory with complex changes and combinations understandable for non-sign language users, which highlighting the urgent need for effective sign language software. With the rapid advancement of artificial intelligence (AI) technologies, the quick learning ability and high degree of accuracy which are the mainly feature for the productions of AI technology that is perfectly suitable for solving the problems we meet in the sign language translation and sign language recognition field. For strengthening the communication between sign language communities and non-sign language communities, the software or applications of sign language translation and sign language recognition plays a significant role.

The development of sign language applications, faces a contradiction between the inherent diversity of sign languages and the convenient of communication brought by technology. On the one hand, we should take full advantage of applying advanced technologies on improving the convenience of humans' production and life faster, which is most article focus on, like implementing the computationally faster optimized algorithm or model. But there is another thing that we should concern either is, unlike spoken

DOI: 10.54254/2755-2721/81/20241162

languages which often have standardized forms, sign languages vary widely in grammar, lexicon, and usage across different regions and communities [1]. This diversity poses a significant challenge to creating a unified sign language standard around world [2]. Also, there is a prevailing belief among many within the Deaf community that sign languages should evolve organically instead of the imposing a rigid standard, to preserve their linguistic and cultural richness. What's more, some marginalized sign languages will face the risk of extinction due to the standardizing process, which is many people afraid to see [3]. Some deaf communities even view it as a form of hearing supremacy threatening traditional sign language communities [2]. Under this context, we must pay enough attention to this special situation when undertaking the relevant work. This paper seeking to digging out an approach which can strike a balance between protecting the diversity of sign languages and enhancing the convenient in practical application.

Currently, many efforts from relevant articles aimed at integrating technologies into sign language. They are generally concentrating on two technical paths, transforming sign language to normal language and generating normal language from sign language. The technical approaches of transform of sign language include sign language recognition and sign language translation. All striving to a more convenient and understandable communicating approach for sign language usage. The paper from Jin and Wu design a two-stream adaptive enhanced spatial temporal graph convolutional network, which practice sign language recognition based on isolated words[4]. Some contribution are put forward in improving sign language recognition algorithm of YOLO, such as the optimization design of YOLOv7-tiny from Han and the optimization design of YOLOv5s from Bao [5-6]. For video recognition of sign language, Min and Chen optimized adaptive keyframe selection for continuous sign language recognition [7]. Aiming at the difficulty of visual feature extraction in sign language, a continuous sign language recognition method based on multi-scale visual feature extraction and cross-modal alignment is proposed by Guo and Xue[8]. In the sign language translation field, a joint end-to-end sign language recognition and translation approach is propose by Necati[9]. Yin optimize the Sign Language Translation model with STMC-Transformer[10]. Sign Language Translation with Monolingual Data is improved by the method of Sign Back-Translation[11]. Liu and Zhou propose a Sign Language Translation system based on human posture research and hand recognition[12]. Virtual human sign Language translation based on offline and online speech recognition is put forward by Li in 2024.

However, most of these efforts are under an assumption that sign language is uniform, and neglect the significant diversity and regional variations that exist within sign languages. Admittedly these achievements lay a solid foundation on the application of model and optimization of algorithm in Machine Learning and Deep learning, but they also fail to consider the diversity of sign language, which may mislead the consequent researches to the direction contrary to the reality situation which overlooking the truly requirements for the people who in need.

This paper aims to critically review the recent researches and applications of AI technology applied in sign languages in different regions, discussing the need to preserve and recognize this diversity within technological advancements. To solve this problem, we propose a realistic technical route for identifying and classifying the diverse aspects of sign languages from different countries and regions, collecting kinds of sign languages and develop a dataset after preprocessing these data. We advocating for the creation of comprehensive attributions that reflect this variety of sign language. Such an approach will enable the development of more tailored and effective technological solutions, ensuring that advancements in AI and related technologies support rather than diminish the rich diversity of sign languages. By doing so, we aim to enhance accessibility and inclusiveness for sign language users, promoting better communication and understanding between different communities.

2. Dataset

Data collection is the cornerstone of developing a reliable sign language recognition system. A well-rounded and extensive dataset is essential for training models that are both accurate and generalizable across various linguistic contexts. In this research, we compiled a dataset totaling 50GB, encompassing over 3,000 sign language words and sentences from various countries. This diverse dataset significantly

DOI: 10.54254/2755-2721/81/20241162

enhances the system's ability to effectively detect and process different sign languages, which is critical for the region detection process—the next crucial step in our system.

However, existing sign language datasets available online often suffer from several limitations, including limited available resources, significant regional differences, and inconsistent data formats. These issues can impede the development of robust and adaptable models, particularly when it comes to detecting the regional origin of sign languages. To address these challenges, we undertook the task of collating and standardizing multiple sign language datasets, with the specific goal of creating a dataset that would optimize the performance of our region detection model.

The dataset was sourced from five different sign language datasets available from Kaggle and The National Center for Sign Language and Gesture Resources at Boston University[13]. The different sources of this dataset gathered from Kaggle including the World Level American Sign Language[14], Russian Sign Language Dataset[15], Indian Sign Language Dataset which funded by the Science and Engineering Research Board of India, and Arabic sign language dataset[16]. These datasets include videos, annotations, and metadata covering sign languages from four countries and regions. The raw data comprises over 3,000 individual sign language words and sentences, occupying a total of 50GB of storage space.

Table 1. The information of the different datasets.

Country	Form	Type	Gestures	Size(GB)
America	Video	Sentences	201	<1GB
America	Video	Words	1999	5GB
Russian	Video	Words	1000	16GB
India	Video	Sentences, Words	700	9GB
Arabia	Video	Words	8,467	2GB

To ensure the dataset was consistent, usable, and suitable for training models, we began by categorizing the data by region, labeling each dataset according to its country or region of origin. This step was crucial for the accuracy of the subsequent region detection task. Inconsistent or mislabeled data entries were identified and corrected to ensure that each data point was accurately classified.

2.1. Annotation Process

In addition to the existing annotations provided by the original datasets, we implemented our own detailed annotation process to enhance the dataset's utility for region detection. Using a combination of automated tools and manual verification, we labeled each gesture with comprehensive metadata. This metadata included hand shape, movement trajectory, facial expressions, and other linguistic features critical for accurate sign language recognition. The annotation process involved the use of Python's Pandas library for data manipulation and custom scripts to ensure that every gesture was consistently and accurately annotated. This additional layer of annotation provided a standardized and enriched set of labels, making the dataset more robust for training machine learning models.

2.2. Data Standardization and Normalization

The next step involved standardizing the data to create a consistent format across the entire dataset. We employed tools such as FFmpeg for video conversion and Python libraries including OpenCV and scikit-image for image processing and normalization tasks. All video files were converted to a uniform resolution of 640x480 pixels and a frame rate of 30 frames per second using FFmpeg, ensuring consistency in the visual data input. This resolution and frame rate were chosen to balance computational efficiency with visual clarity. Variations in lighting conditions across different videos were standardized using histogram equalization techniques in OpenCV, adjusting brightness and contrast levels to a consistent baseline. Additionally, videos were reoriented to ensure that all gestures were centered and uniformly aligned, correcting any tilt or skew in the original recordings. These normalization processes

DOI: 10.54254/2755-2721/81/20241162

were essential to ensure that the gesture data was consistent in size, shape, and movement across the entire dataset.

2.3. Data Augmentation for Robustness

Recognizing the importance of data diversity for model generalization, we further expanded the dataset using data augmentation techniques. Augmentation was necessary to increase the dataset's robustness, allowing the model to better generalize across different sign languages and regional variations. Techniques such as horizontal flipping of videos were used to simulate left-handed gestures, while rotating video frames accounted for variations in camera angles. Additionally, scaling videos simulated different distances between the signer and the camera, creating a more comprehensive training set. These augmentation techniques were implemented using the imgaug Python library, effectively increasing the dataset size by approximately 150%. This augmentation process not only enriched the dataset but also ensured that the model would be resilient to variations in signing style, orientation, and distance—factors that are crucial for accurate region detection.

2.4. Final Dataset Composition and Suitability for Training

After completing the standardization, annotation, and augmentation processes, the dataset was evaluated to ensure its suitability for training machine learning models. The final dataset comprises approximately 7,500 unique samples, distributed across five regional categories, each containing around 1,500 samples. The videos have an average duration of 3-5 seconds, providing a consistent and manageable input size for model training. The dataset's uniformity in resolution, frame rate, and annotation makes it ideally suited for training convolutional neural networks (CNNs).

For training the region detection model, we propose a CNN architecture tailored to this dataset. The model will consist of five convolutional layers, followed by two fully connected layers, designed to capture the nuanced differences between regional sign languages. The input size for the model will be 640x480 pixel frames with a 3-channel RGB input. We will use the Adam optimizer with an initial learning rate of 0.001, and the categorical cross-entropy loss function, which is appropriate for this multi-class classification task. The model will be trained for 50 epochs, with early stopping based on validation loss to prevent overfitting. A batch size of 32 will be used to balance computational efficiency with training stability.

By meticulously standardizing, annotating, and augmenting the dataset, we have created a robust resource that not only supports accurate region detection but also ensures that the subsequent translation processes are sensitive to the linguistic and cultural nuances of different sign languages. This dataset forms the backbone of our system, providing the necessary foundation for the development of a sophisticated and adaptable sign language recognition and translation model.

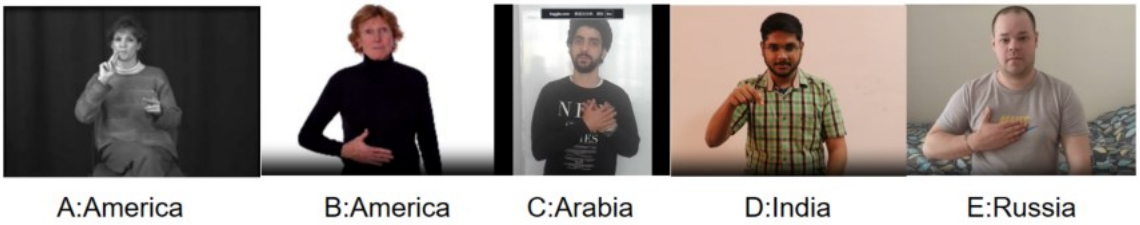

Figure 1. The different sources of this dataset.

DOI: 10.54254/2755-2721/81/20241162

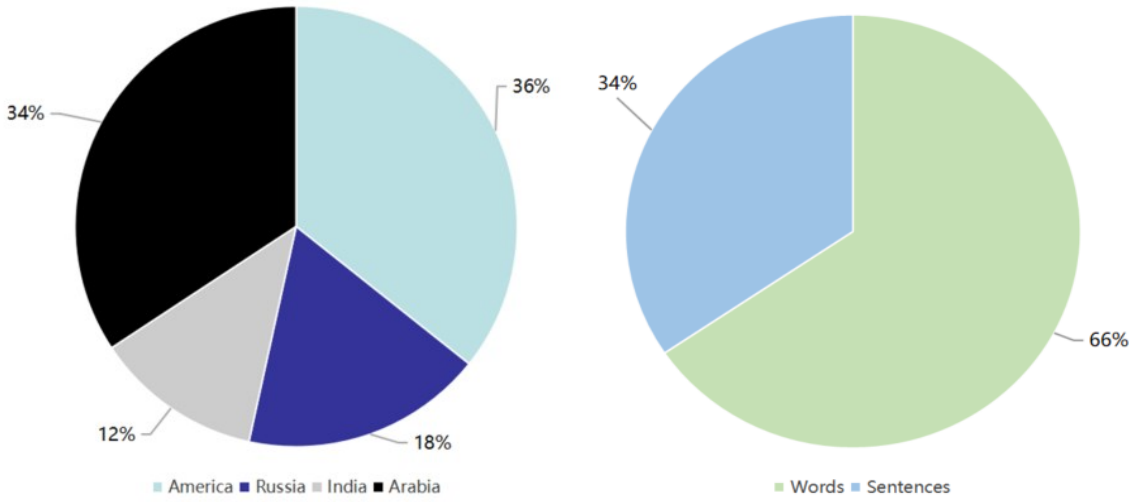

Figure 2. The proportion of some index in this dataset.

3. System

This article also proposes a design of a sign language recognition and translation system, presents a comprehensive system that consists of two major steps: region detection and exporting to sub-models for translation. The flow of the process details in the follow chart.

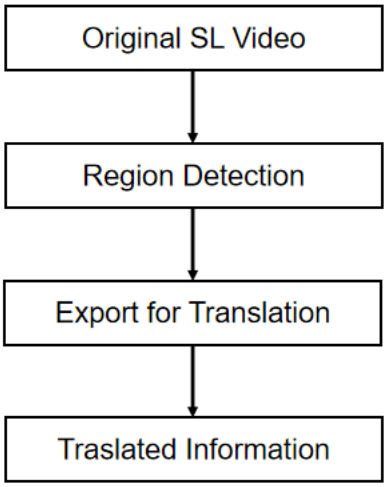

Figure 3. The flowchart of the proposed system.

The system is designed to efficiently recognize and translate sign language by first detecting the regional origin of the sign language and then exporting the identified region's data to the appropriate sub-model for translation.

The process begins with the input of the Original SL (Sign Language) Video, which serves as the initial data for the system. This video is then processed through the Region Detection stage, where the system identifies the geographical or cultural region from which the sign language originates. This step is crucial for ensuring that the nuances and specificities of different sign languages are accurately captured and interpreted.

After region detection, the system proceeds to the Export for Translation stage. Here, based on the identified region, the data is routed to a specialized sub-model that is trained to handle the sign language

DOI: 10.54254/2755-2721/81/20241162

of that particular region. This sub-model then translates the sign language gestures into a more universally understood format, such as text or spoken language, producing the Translated Information as the final output.

This multi-step process ensures that the system not only translates sign language accurately but also respects the diversity and uniqueness of sign languages from different regions. By incorporating region-specific sub-models, the system addresses the challenges posed by the linguistic and cultural variations inherent in sign language, making it a robust tool for global sign language translation.

3.1. Region Detection

Region detection is a crucial step where the system detects the regional origin of a given sign language sample using a classification model. Feature extraction involves identifying and extracting relevant features from the preprocessed data. These features help differentiate sign languages from different regions and are critical inputs for the classification model. Feature extraction methodologies are discussed in [17], highlighting their importance in sign language recognition. A classification model is trained using machine learning algorithms such as Convolutional Neural Networks (CNNs) to recognize the regional origin of sign language gestures. Training on a diverse dataset ensures the models' broad applicability. Validating the classification model with a separate dataset ensures its accuracy and reliability, which is crucial for real-world applications. Model validation techniques are discussed in [18], emphasizing their role in verifying model performance. TensorFlow or PyTorch for building and training the classification model. Libraries and tools for extracting key features from sign language gestures.

Presently, the most commonly used methods in sign language recognition focus primarily on translating sign language into actual spoken or written language. These methods typically target the semantic content of the signs, often overlooking the regional variations that exist within sign languages. However, recognizing these regional differences is essential for a comprehensive understanding and accurate translation, as highlighted by our approach.

3.2. Export to Sub-models for Translation

After identifying the regional origin of the sign language, the system exports the data to the corresponding sub-model for translation. Based on the detected region, the appropriate sub-model specialized in that particular sign language is selected for translation. The selected sub-model translates the sign language gestures into the desired output, such as text or spoken language, converting visual gestures into a more universally understood format.

Once the region is identified, the system dynamically selects the appropriate sub-model. These sub-models are designed specifically for the sign language of the detected region and are trained to handle the unique linguistic and grammatical structures of that language. The selection process leverages a region-based lookup table that maps detected regions to their corresponding sub-models, ensuring that the translation model is both regionally and linguistically appropriate. Each sub-model employs a tailored neural network architecture to achieve high translation accuracy. Typically, the architecture is based on a sequence-to-sequence model, which is particularly effective for translating sign language gestures into spoken or written language. The backbone of these sub-models is often a Recurrent Neural Network (RNN) or a Long Short-Term Memory (LSTM) network, which is well-suited for handling sequential data like sign language.

However, to enhance the model's ability to capture long-range dependencies and complex gestures, each sub-model is equipped with an attention mechanism. The attention mechanism allows the network to focus on different parts of the input sequence as needed, making the translation more accurate by emphasizing critical gestures or facial expressions at the appropriate time. This approach is particularly useful in translating sign languages that involve intricate spatial relationships and simultaneous multi-channel inputs (e.g., hand shape, movement, and facial expression). Xie and Ding demonstrates the benefits of using bidirectional LSTM in combination with an attention mechanism to enhance the accuracy of visual language interpretation, a concept that can be extended to the translation of sign

DOI: 10.54254/2755-2721/81/20241162

language gestures[19]. For more complex languages or larger datasets, as discussed in [20], the system can also utilize Transformer models. Transformers have been shown to outperform traditional RNNs and LSTMs in many translation tasks due to their parallel processing capabilities and their ability to capture global dependencies in the data. In cases where the sign language involves highly complex grammar or requires context-dependent interpretation, Transformer-based models are particularly advantageous.

Once the sub-model has been selected, the translation process begins. The neural network takes the input sequence of gestures, processes it through the LSTM or Transformer layers, and produces a sequence of tokens that correspond to words or phrases in the target language. The output sequence is then post-processed to ensure grammatical correctness and contextual relevance. In some cases, additional layers or models, such as a Conditional Random Field (CRF), may be used to refine the output sequence, ensuring that the translated language adheres to the correct syntax and semantics.

The final output is generated in the desired format—text, spoken language, or another form of communication. For text output, the sequence of tokens is converted into a coherent sentence or paragraph. For spoken output, a text-to-speech engine, such as Tacotron or WaveNet, may be employed to convert the text into natural-sounding speech. This flexibility in output generation allows the system to be applied in various contexts, from real-time communication to educational tools.

4. Discussion

The proposed system for sign language recognition and translation introduces several key innovations that differentiate it from existing systems. This section discusses these distinctions and highlights the advantages of the proposed approach. Existing sign language recognition systems often assume a standardized sign language, training on datasets that do not capture the full diversity of sign languages across different regions. Consequently, these systems may not perform well with regional variations or less common sign languages. Some systems focus on American Sign Language (ASL) or a few widely used sign languages, often neglecting regional diversity. The proposed system includes a distinct regional detection phase. This phase utilizes a classification model trained to identify the regional origin of a sign language sample before translation. This step ensures the translation process is accurate and contextually appropriate, addressing the linguistic and cultural nuances of different sign languages. Current systems generally do not incorporate a region-specific detection mechanism, leading to potential inaccuracies when dealing with diverse sign languages. After detecting the sign language's region, the system employs specialized sub-models tailored to each identified sign language. These sub-models are trained on datasets representative of their respective regions, ensuring higher accuracy in translation. Unlike generalized models used in many existing systems, our approach of using specialized sub-models, allows for more precise and contextually relevant translations, accommodating the unique grammatical structures and vocabularies of different sign languages.

By recognizing the regional origin of sign language gestures and using specialized sub-models for translation, the proposed system achieves higher accuracy. This ensures the system can handle a wide range of sign languages, including those that are marginalized or at risk of extinction. This approach promotes linguistic diversity and inclusivity, acknowledging and respecting the unique characteristics of various sign languages. It prevents the marginalization of less common sign languages, fostering a more inclusive environment for all users. The modular design of the proposed system, with distinct phases for data preprocessing, regional detection, and translation, allows for easy scalability and adaptability. New sign languages and regional variations can be incorporated by training additional sub-models and updating the classification model. This scalability ensures the system can adapt to evolving linguistic trends and the emergence of new sign languages. The focus on collecting a diverse and comprehensive dataset from various regions and sign language communities ensures that the system can generalize well across different sign languages. This dataset serves as a robust foundation for training accurate and inclusive models. For a robust data foundation, a comprehensive dataset is critical for developing effective sign language recognition systems that can cater to a wide range of users.

DOI: 10.54254/2755-2721/81/20241162

5. Conclusion

This paper addressed the limitations of existing sign language translation systems, particularly their difficulty in effectively managing the linguistic diversity inherent in sign languages across various regions. Through a detailed examination of these challenges, a modular system was developed that incorporates region detection and specialized sub-models for translation, enabling more precise and culturally sensitive recognition. Unlike generalized models, this approach facilitates tailored translations based on regional variations, significantly improving both accuracy and adaptability. A comprehensive, standardized dataset was also introduced to support both region detection and model training. This dataset was meticulously curated, featuring detailed annotation, normalization, and augmentation processes, ensuring its suitability for a wide range of sign languages. By utilizing this diverse and well-structured dataset, the system efficiently recognizes and translates sign languages from different regions while preserving cultural nuances. Overall, the proposed model offers a scalable, flexible, and practical solution for real-world applications, ensuring the preservation of sign language diversity and enhancing the overall performance and quality of translation systems.

References

[1] Kusters A, Lucas C. Emergence and evolutions: Introducing sign language sociolinguistics[J]. Journal of Sociolinguistics, 2022, 26(1): 84-98.

[2] ZHAO Xiaochi, YE Guichen. Standardization of Chinese Sign Language: Process and Review [J]. Journal of Guilin University of Aerospace Technology, 2019, 24(2): 304-310.

[3] Nonaka AM. The forgotten endangered languages: Lessons on the importance of remembering from Thailand's Ban Khor Sign Language[J]. Language in Society, 2004, 33(05).

[4] Bird JJ, Ekárt A, Faria DR. British Sign Language Recognition via Late Fusion of Computer Vision and Leap Motion with Transfer Learning to American Sign Language[J]. Sensors, 2020, 20(18): 5151.

[5] Min Y, Chen X. Adaptive Key Frame Selection for Continuous Sign Language Recognition [J/OL]. SCIENTIA SINICA Informationis, 2023. DOI:10.1360/SSI-2022-0467.

[6] Han Xiaobing, Hu Qisheng, Zhao Xiaofei, et al. Research on an Improved YOLOv7-tiny Algorithm for Sign Language Recognition [J]. Modern Electronic Technology, 2024, 47(1): 55-61.

[7] Bao Shuhan, Sun Mulun, Liu Shuqi, et al. Sign Language Recognition Algorithm Based on Improved YOLOv5s [J]. Journal of Jiaxing University: 1-12.

[8] Jin Yanliang, Wu Xiaowei. Sign Language Recognition Based on Dual-Stream Adaptive Spatio-Temporal Enhanced Graph Convolutional Network [J]. Journal of Applied Sciences, 2024, 42(2): 189-199.

[9] Guo Leming, Xue Wanli, Yuan Tiantian. Continuous Sign Language Recognition with Multi-Scale Visual Feature Extraction and Cross-Modal Alignment [J]. Computer Science and Exploration: 1-10.

[10] Li Xin. Research on Virtual Human Sign Language Translation Based on Offline and Online Speech Recognition [D]. Tianjin University of Technology, 2024.

[11] Liu Jixing, Zhou Xin, Zhang Shuaifeng, et al. Implementation of a Sign Language Translation System Based on Artificial Intelligence [J]. Science and Technology Innovation and Application, 2022, 12(23): 41-43+48.

[12] Tang Shengeng. Research on Sign Language Translation and Generation Technology Based on Deep Learning [D]. Hefei University of Technology, 2023.

[13] Dreuw P, Rybach D, Deselaers T, Zahedi M, Ney H. Speech recognition techniques for a sign language recognition system[C]//Interspeech 2007. Antwerp, Belgium, August 2007: 2513-2516.

[14] Li D, Rodriguez C, Yu X, et al. Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison[J]. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2020: 1459-1469.

DOI: 10.54254/2755-2721/81/20241162

[15] Kapitanov A, Karina K, Nagaev A, et al. Slovo: Russian Sign Language Dataset[J]. International Conference on Computer Vision Systems, Cham: Springer Nature Switzerland, 2023: 63-73.
[16] Balaha M M, El-Kady S, Balaha H M, et al. A vision-based deep learning approach for independent-users Arabic sign language interpretation[J]. Multimedia Tools and Applications, 2023, 82(5): 6807-6826.
[17] Barbhuiya A.A., Karsh R.K., Jain R. CNN-Based Feature Extraction and Classification for Sign Language [J]. Multimedia Tools and Applications, 2021, 80: 3051–3069.
[18] Al-Qurishi M, Khalid T, Souissi R. Deep Learning for Sign Language Recognition: Current Techniques, Benchmarks, and Open Issues [J]. IEEE Access, 2021, 9: 126917-126951.
[19] Xie T, Ding W, Zhang J, et al. Bi-LS-AttM: A Bidirectional LSTM and Attention Mechanism Model for Improving Image Captioning [J]. Applied Sciences, 2023, 13(13): 7916.
[20] Kamyab M, Liu G, Adjeisah M. Attention-Based CNN and Bi-LSTM Model Based on TF-IDF and GloVe Word Embedding for Sentiment Analysis [J]. Applied Sciences, 2021, 11(23): 11255.