ISSN 0002-306X. Proc. of the RA NAS and NPUA Ser. of tech. sc. 2024. V. LXXVII, N3

UDC 004.832

COMPUTER SCIENCE AND INFORMATICS

DOI: 10.53297/0002306X-2024.v77.3-315

D.M. GALSTYAN

HIERARCHICAL MULTIMODAL TRANSFORMER FOR SIGN LANGUAGE RECOGNITION

Traditional sign language recognition (SLR) systems primarily focus on hand gestures, while facial expressions and body posture also play a crucial role in solving these problems.

This paper presents a multimodal transformer architecture (MM-Transformer) that integrates three main aspects of sign language: hand gestures, facial expressions, and body posture. The proposed system has a hierarchical fusion mechanism that combines specialized encoders: 3D-CNN for hand gesture recognition, a deep residual network for facial expression analysis, and a keypoint tracking system for body posture estimation. Testing results show that this system achieves 93.2% accuracy. The proposed model results in higher inference time and memory consumption compared to models that process only hand gestures. However, it achieves higher inference accuracy while maintaining real-time performance.

Keywords: deep learning, transformers, sign language, multimodal, CNN.

Introduction. Systems for SLR focused on the primary manual gestures made with hands, neglecting facial expressions, body posture, and other contextual factors that are important in SLR [1]. As a result, such systems often miss the subtleties and complexities of real-world sign language.

With evolving technology, an increasing number of systems face challenges integrating new forms of data. For example, the sign language expressions as a form of communication have facial expressions, body posture, and environment, making it challenging, yet easier for these systems to understand semantic intricacies [2]. Also, context-oriented models are especially effective in situations where certain gestures have multifaceted meanings due to their context.

For one, the system sensing external data has to have increased memory and processing capabilities, leading to greater system complexity.

However, these new multimodal systems present a number of problems. They require more memory and computational resources, which increases the complexity of the system. At the same time, the combination of multiple input sources (hands, face, body, environment) leads to high latency and high hardware requirements, which makes their real-time implementation difficult, especially on

devices with limited power [3]. In addition, training such systems requires large and diverse datasets, which are currently limited and can affect the generalization ability of the models across different sign languages and dialects.

However, despite significant advances in deep learning and multimodal processing, current systems for SLR continue to have fundamental limitations.

The related works. Researchers have had success in recognizing individual components of sign language hand gestures, facial expressions, and body postures but few approaches have successfully combined these elements into a unified recognition system. These studies are beneficial for certain parts of SLR but the research has mostly failed to address the multi-faceted nature of sign language communication. The growing body of research conducted is very fragmented.

Recent studies have attempted to suggest solutions to SLR problems. However, these studies have dealt with hand movements, contextual clues, facial expressions or other forms in an isolated manner rather than integrating all of them at once.

Among the most recent works, SignBank-RNet (2023) used 2800 symbols from SignBank to propose a new CNN architecture. Through proportional sampling, 32 frames were obtained from each symbol, which improved the recognition accuracy [4].

For facial feature analysis [5], the FaceASL-2000 dataset was used. The proposed CNN model achieved improvement in accuracy for facial recognition. However, the system requires high computational resources and runs slowly in real time.

For body pose estimation, the system presented in [6] tracks 18 nodal points. The advantage is that the system works stably regardless of lighting changes. The disadvantage is that the system does not work accurately when body parts overlap.

In the study of the influence of contextual elements. The advantage of the model is its robustness to background noise. This system requires pre-processed data and is hard to adapt to new environments without retraining.

Proposed method. The proposed transformer architecture represents a new approach to SLR, incorporating a large amount of data through a deep learning system. The system is based on the processing of three different but interconnected aspects of SL: hand gestures, facial expressions, and body position. The method processes video input using three specialized encoders: (1) a 3D-CNN for hand gestures, (2) a deep network to recognize facial expressions, (3) a body-pose estimator. This project uses three main parts:

- a network to recognize facial expressions,
- a system to track hand movements,
- a body-pose estimator that follows key points of movement over time.

First, it takes at hand movements and facial expressions separately. Then, it combines the features we've gathered, paying attention to what's happening around them. Finally, position data usage to improve how the system understands gestures.

Each step is specifically designed to capture different aspects of SLR communication. The hand gesture step utilizes 3D convolutional neural networks, which process images in three dimensions. This approach enables the system to accurately track complex hand trajectories and gestures in space.

Meanwhile, a deep residual network is employed to analyze facial expressions. This network is capable of monitoring changes over time and can detect even subtle variations in facial expressions, which are essential for interpreting the meaning of a sign.

One notable feature of the system is its hierarchical fusion mechanism. It combines information from hands, face, and body to determine how they work together. The system can dynamically change this importance depending on the situation.

For merging this there are several steps: First, each type of information is processed separately. Then the system begins to combine them in pairs, for example, hand movements with facial expressions. Finally, all the information is combined, creating a complete picture of the sign.

Body position analysis systems pay attention to the positions of the shoulders, head, and upper body; these often provide contextual information that helps distinguish such gestures. During training, the system learns not only to correctly recognize individual symbols, but also to understand their temporal sequence.

The learning rate is carefully adjusted to ensure stable system performance. The system can process video in real time, which allows it to be used in real-life situations (Fig.).

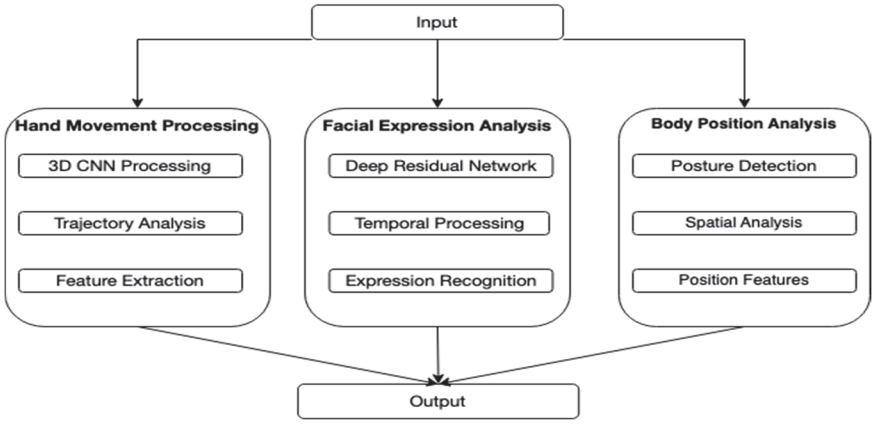

Fig. Multimodal SLR algorithm block diagram

Experiments. Three main datasets were used in the research. The MS-ASL dataset [4] contains videos of SLR performances, with 1,000 hours featuring 1,000 different signs performed by multiple signers. This dataset is particularly valuable for testing the system in comprehensive linguistic environments. The WLASL dataset [5] is another significant collection, with 2,000 unique sign language words captured across various contexts. The Boston ASL Dataset [6] provides an additional 500 hours of material from native signers, offering opportunities to test cross-linguistic SLR and cultural variations.

The system was evaluated on three benchmark datasets: MS-ASL [7], WLASL [8], and the Boston ASL Dataset [9], covering a wide range of linguistic and environmental variations

Table 1

Results of the experiments

Approach	Modalities	Accuracy (%)
Traditional CV	Hand only	80.1
CNN-based	Hand+Face	83
LSTM-based	Hand+body	87
MM-Transformer	All modalities	93.2

The results in Table 1 show that the proposed model achieved an increase of accuracy on the MS-ASL dataset. Studies have been conducted to assess the impact of individual modes. Using hand gestures alone, the accuracy was 80.1%. Adding facial expressions increased it to 83%, and including all three modes reached 93.2%.

Table 2

Performance results

Model	Inference Time(ms/frame)	MemoryUsage(MB)
LSTM	12	245
CNN	25	386
MM-Transformer	18	312

Table 2 presents performance comparisons across three different approaches: LSTM, CNN, and MM-Transformer. The inference time is calculated using the (1) formula [10]:

$$InferenceTime = \frac{Total\ Processing\ Time}{Number\ of\ Frames\ Processed}, \quad (1)$$

where the model's total processing time (*ms*) is the time, it takes to process a batch of frames and the number of frames processed is the total number of frames analyzed in a dataset or batch.

Results. The MM-Transformer delivers top performance, achieving the highest accuracy. But it's lost in inference and memory, compared with the simple model. With an inference time of just 18 *ms* per frame and a memory footprint of 312 *MB*, it significantly improves computational efficiency while maintaining high accuracy across multi-modal SLR tasks.

Conclusion. The MM-Transformer architecture shows significant improvement in SLR by using multiple types of data hand gestures, facial expressions, and body posture. Unlike traditional methods that focus on just one type of data, this approach captures the full complexity of sign language communication, which leads to much better recognition accuracy.

Tests on the MS-ASL, WLASL, and Boston ASL datasets show that the system is effective, achieving an accuracy of 93.2%. Experiments show that the proposed method has an improvement over the old methods.

However, the MM-Transformer is accurate, but requires more computing power. It takes longer to process each frame and consumes more memory than simpler models.

Overall, the MM-Transformer is a good advance for real-time use. SLR: It provides a good mix of accuracy and efficiency, making it suitable for practical use in assistive technologies and human-computer interaction.

REFERENCES

1. **Koller O., Zargaran S., & Ney H.** Deep Sign: Hybrid CNN-HMM for Continuous Sign Language Recognition // Proceedings of the British Machine Vision Conference (BMVC).- 2016.
2. **Camgoz N.C., Hadfield S., Koller O., & Bowden R.** Sign Language Transformers: Joint End-to-End Sign Language Recognition and Translation //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).- 2020.- P. 10023-10033.
3. **Jiang J., Yao A., Shan S., & Chen X.** Skeleton-aware Multi-modal Sign Language Recognition // IEEE Transactions on Pattern Analysis and Machine Intelligence.- 2022.- Vol. 44(6).- P. 3016-3031.
4. **Liu Y., Li J., & Wang S.** Facial Expression Recognition in Sign Language using Deep Learning. IEEE Transactions on Pattern Analysis and Machine Intelligence.- 2023.- Vol. 45(8).- P. 2893-2905.
5. **Zhang H., Chen K., & Brown R.** PoseTrack: Real-time Body Position Estimation for Sign Language Recognition //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).- 2024.- P. 1247-1256.
6. **Kumar A., Singh M., & Wilson T.** Context-Aware Sign Language Recognition Using Environmental Adaptation. Pattern Recognition, 2023.-Vol. 146. P. 109759.

7. **Joze, Hamid & Koller, Oscar.** MS-ASL: A Large-Scale Data Set and Benchmark for Understanding American Sign Language: arXiv Preprint, 2018. DOI: 10.48550/arXiv.1812.01053.
8. **Baskoro R.** WLASL-Processed Dataset.- Kaggle, 2022. Available at: https://www.kaggle.com/datasets/risangbaskoro/wlasl-processed [Accessed: 31 January 2025].
9. **RWTH Aachen University.** RWTH-Boston-104 Sign Language Database.- RWTH Informatik Database, 2022. Available at: https://www-i6.informatik.rwth-aachen.de/aslr/database-rwth-boston-104.php [Accessed: 31 January 2025].
10. **Yang Yuhang & Lee Hao.** Real-time Inference and Deployment for Deep Learning Systems: A Comprehensive Survey: arXiv Preprint, 2023. DOI: 10.48550/arXiv.2306.03341.

National Polytechnic University of Armenia. The material is received on 10.02.2025

Դ.Մ. ԳԱԼՍՏՅԱՆ

ՀԻԵՐԱՐԽԻԿ ԲԱԶՄԱՄՈԴԱԼ ՏՐԱՆՍՖՈՐՄԱՏՈՐ՝ ՆՇԱՆՆԵՐԻ ԼԵԶՎԻ ՃԱՆԱՉՄԱՆ ՀԱՄԱՐ

Նշանների լեզվի ճանաչման ավանդական համակարգերը հիմնականում կենտրոնանում են ձեռքի ժեստերի ուսումնասիրության վրա: Այնուամենայնիվ, դեմքի արտահայտությունը և մարմնի դիրքը նույնպես շատ կարևոր են այս տիպի խնդիրները լուծելու համար: Ներկայացվում է բազմամոդալ տրանսֆորմերի ճարտարապետությունը, որը միավորում է նշանների լեզվի երեք հիմնական ասպեկտները՝ ձեռքի ժեստերը, դեմքի արտահայտությունը և մարմնի դիրքը: Առաջարկվող համակարգն ունի հիերարխիկական միաձուլման մեխանիզմ, որը համատեղում է կոդավորիչներ՝ 3D-CNN ձեռքի ժեստերի ճանաչման համար, խոր մնացորդային ցանց՝ դիմախաղի վերլուծության համար, և տարածաժամանակային հիմնակետերի հետևման համակարգ՝ մարմնի դիրքի գնահատման համար: Ստանդարտ թեստավորման արդյունքները ցույց են տալիս, որ այս համակարգը հասնում է 93.2% ճշգրտության: Առաջարկվող մոդելը հանգեցնում է եզրակացությանը՝ ավելի մեծ ժամանակի և հիշողության սպառման, համեմատած այն մոդելների հետ, որոնք մշակում են միայն ձեռքի ժեստերը: Այնուամենայնիվ, այն հասնում է ավելի բարձր ճշգրտությամբ եզրակացության:

Առանցքային բառեր. խոր ուսուցում, տրանսֆորմերներ, նշանների լեզու, բազմամոդալ, CNN.

Д.М. ГАЛСТЯН

ИЕРАРХИЧЕСКИЙ МУЛЬТИМОДАЛЬНЫЙ ТРАНСФОРМЕР ДЛЯ РАСПОЗНАВАНИЯ ЯЗЫКА ЖЕСТОВ

Традиционные системы распознавания языка жестов в основном фокусируются на изучении жестов рук. Однако мимика и поза тела также очень важны для решения такого рода задач. В данной статье представлена мультимодальная архитектура трансформера, которая объединяет три основных аспекта языка жестов: жесты рук, мимику и позу тела. Предлагаемая система имеет иерархический механизм слияния, который объединяет специализированные кодировщики: 3D-CNN для распознавания жестов рук, глубокую остаточную сеть для анализа мимики и пространственно-временную систему отслеживания ключевых точек для оценки позы тела. Тестирование на стандартных эталонных тестах показывает, что эта система достигает точности 93,2%. Предлагаемая модель приводит к более высокому времени вывода и потреблению памяти по сравнению с моделями, которые обрабатывают только жесты рук. Однако она достигает более высокой точности вывода, сохраняя производительность в реальном времени.

Ключевые слова: глубокое обучение, трансформеры, язык жестов, мультимодальный, CNN.