https://doi.org/10.59849/aidd.2024.71

Human Action Recognition Model Reuse for Word-Level Sign Language Recognition

Gulchin Abdullayeva[1], Nigar Alishzade[2]

[1]MSERA Institute of Control Systems, Baku, Azerbaijan,
E-mail: gulchinabdullayeva1947@gmail.com
[2]Azerbaijan State Oil and Industry University, Baku, Azerbaijan,
E-mail: nigar.alishzada@ufaz.az

Abstract. Sign languages are visual, non-verbal communication systems that use sequences of manual and non-manual gestures to convey meaning. Word-level sign language recognition (SLR) aims to automatically identify individual lexical items within sign language vocabularies. While deep learning has achieved notable advancements in SLR, it typically requires large-scale, domain-specific data. Many sign languages, however, lack extensive labeled datasets, making it challenging to train models from scratch. To address this issue, we propose reusing models originally developed for human action recognition, which have been trained on large and widely available datasets. By leveraging transfer learning, we adapt these pretrained 3D convolutional neural networks to the target sign language data. In this study, we focus on an isolated Azerbaijani Sign Language dataset, which is comparatively small. The model's prior knowledge of human actions provides a strong initialization, allowing us to fine-tune the network with limited sign language examples. Our approach demonstrates that reusing human action recognition models can effectively bridge data scarcity, enhance training efficiency, and improve performance in word-level SLR tasks.

Keywords: Sign Language Recognition, Human Action Recognition, Word-level SLR

1. Introduction

Sign language recognition (SLR) serves as an essential bridge that can facilitate communication between sign language users and non-signers, promoting inclusivity and accessibility. Within SLR, word-level recognition plays a critical role in accurately identifying individual lexical items from visual cues. However, this task often requires substantial volumes of annotated training data, which are not always available for many sign languages. To overcome this limitation, we explore reusing models originally trained for human action recognition—an area where extensive datasets are readily available. This approach aligns with the principles of transfer learning, whereby knowledge gained from a data-rich source domain is adapted to a target domain with limited data [1].

By leveraging the expertise encapsulated in a pre-trained deep learning model, originally developed for tasks with more abundant data, we can fine-tune the network on a smaller dataset—such as an Azerbaijani Sign Language corpus—and significantly improve recognition performance. Our research focuses on the importance of pre-trained models in visual-based SLR, emphasizing the efficacy of 3D convolutional neural networks (3D CNNs) and their integration with recurrent layers like LSTM for capturing dynamic, temporally evolving signs. This paradigm allows us to transfer representational knowledge gleaned from large-scale human action recognition datasets and apply it to the nuanced, word-level classification of signs in languages for which resources remain scarce.

3D CNNs are inherently suited for capturing the spatiotemporal nature of sign languages. Initially proven effective in human action recognition tasks [2], [3], [4], these architectures have been embraced by the Sign Language Processing community for their ability to model both the spatial configuration of hands and the temporal progression of gestures. By aligning this technology with transfer learning, we establish a roadmap that allows researchers to tackle the scarcity of annotated data in lesser-resourced sign languages. This approach ensures that communities with limited linguistic resources can still

benefit from state-of-the-art SLR solutions, fostering greater inclusion in digital communication platforms.

Through pre-training on human action recognition data and subsequently fine-tuning on an Azerbaijani Sign Language dataset, we enable our model to effectively capture the underlying spatiotemporal patterns crucial for accurate sign identification. The adaptability of 3D CNN architectures, combined with transfer learning principles, directly addresses the unique challenges posed by limited datasets. Studies such as [5] confirm the value of these frameworks for extracting complex temporal cues, positioning them as an ideal solution for word-level SLR tasks. In doing so, we not only enhance recognition performance but also contribute to bridging the gap between well-studied sign languages and those that remain underrepresented in contemporary research.

2. Related Work

We primarily focus on CNN-oriented computer vision techniques because SLR involves human action recognition. Sign languages consist of complex sets of human gestures and body articulations, so the majority of the techniques were adapted from this superclass. The study by Khurana et al. [6] offers a thorough overview of the current literature on sign language recognition, employing deep learning methods and transfer learning techniques.

In [7], the authors proposed Spatial-Temporal Graph Convolutional Networks for skeleton-based action recognition, which move beyond the limitations of previous methods by automatically learning both the spatial and temporal patterns from data.

The authors of [8] proposed a two-stream CNN framework for action recognition, incorporating both spatial and temporal information from RGB and optical flow images.

In [9], the authors used 2D CNNs for training and bootstrapped the weights into 3D. They devised a 2D-Inflated operation and a parallel 3D ConvNet architecture for converting pre-trained 2D ConvNets into 3D ConvNets, which avoids video data pre-training.

In the field of sign language recognition, studies have used transfer learning and convolutional neural networks with pre-trained models like InceptionV3 and ResNet-50 to improve accuracy [10]. These studies have shown promising results in recognizing various sign languages, including American Sign Language, British Sign Language, and Indian Sign Language [11, 12]. Furthermore, the utilization of multi-modal data sources, such as infrared, contour, and skeleton information, has been proven to enhance the performance of sign language recognition systems [1, 2].

The results of the experiments [13] give clear empirical evidence that transfer learning can be effectively applied to isolated SLR. The accuracy performances of the networks applying transfer learning increased substantially by up to 21% as compared to the baseline models that were not pre-trained on the MS-ASL dataset.

In [1] the authors proposed a Skeleton Aware Multi-modal Sign Language Recognition framework, which takes advantage of multi-modal information to improve recognition rate. Specifically, the framework incorporates a Sign Language Graph Convolution Network to model the dynamics of sign language gestures, and a Separable Spatial-Temporal Convolution Network to exploit skeleton features.

In [14], the authors introduced multi-scale spatiotemporal graph convolutional networks for isolated sign language recognition. These networks leverage the spatiotemporal characteristics of sign language gestures by incorporating graph convolutional networks at multiple scales.

In [15], the authors proposed a (2+1)D-SLR network based on (2+1)D convolution that can achieve higher accuracy with a faster speed. Because (2+1)D-SLR can learn spatio-temporal features from the raw sign RGB frames.

The authors of [16] proposed a Temporal Interaction Module to capture both spatial and temporal information in sign language videos. They suggest not only can it obtain higher accuracy, but also inference speed and parameters of the network can meet practical application scenarios, because the TIM-SLR network is only composed of 2D convolution and temporal interaction module (TIM).

In [17], the authors provided comparative evaluations of different deep learning architectures for word-level sign language recognition, including 3D CNNs.

In [18] and [19] authors provide different approaches for Azerbaijani Sign Language fingerspelling alphabet and also employ transfer learning.

Addressing the dynamic nature of signing, including temporal dependencies as well as spatiotemporal characteristics, has been an area of focus in various works aiming to enhance the accuracy and efficiency of sign language recognition systems [20].

In conclusion, sign language recognition research has made significant progress by utilizing convolutional neural networks, transfer learning, and multi-modal information. In our study, we aim to build upon these advancements. By leveraging transfer learning on 3D convolutional neural network architecture, and multi-modal data streaming we aim to improve the accuracy of isolated sign language recognition for our specific dataset.

3. Methodology

In this study, we adapt a pre-trained human action recognition model to the task of word-level sign language recognition, underscoring the importance of leveraging existing knowledge obtained from large-scale datasets. Specifically, we employ a 3D Convolutional Neural Network (3D CNN) architecture that was originally trained on a substantial, well-curated action recognition dataset (e.g., Kinetics). By repurposing these weights, our approach aims to transfer the model's robust spatiotemporal feature representations into a domain with more limited resources—the Azerbaijani Sign Language dataset.

For fine-tuning, we collected and prepared an isolated Azerbaijani Sign Language dataset from frontal RGB video recordings. The dataset construction process began with sentence-level video clips, which were segmented into individual words. To ensure a sufficient number of samples per class, we selected only those words that had at least 20 labeled instances, resulting in a focused vocabulary of the top 100 words. The resulting corpus was divided into training and testing subsets at a 9:1 ratio, ensuring a balanced and representative evaluation.

Preprocessing steps were applied to the RGB video data to enhance input quality and improve training stability. Center-zoom operations ensured that the signer's hands and facial expressions remained consistently within the frame, while normal

ization techniques helped maintain uniformity across samples. These preprocessing measures were critical in refining the input data, enabling more efficient fine-tuning of the 3D CNN backbone.

Our results indicate that utilizing the prior knowledge encapsulated in a pre-trained 3D CNN significantly enhances performance on the low-resource Azerbaijani Sign Language recognition task.

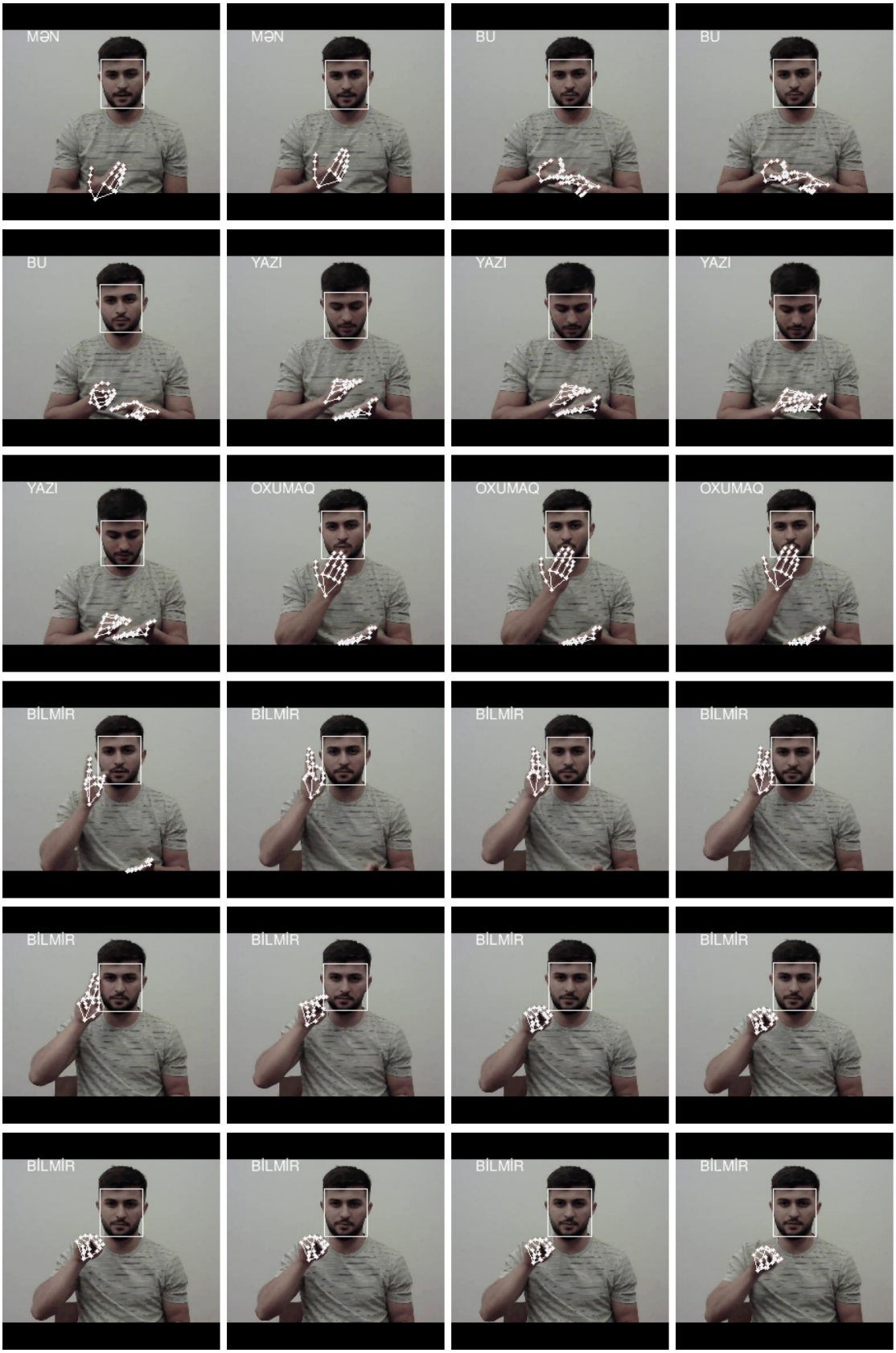

Fig. 1. The results of real-time inference for a sentence with 5 words

Compared to conventional models trained from scratch, transfer learning from an action recognition source domain facilitates a richer, more transferable representation of spatiotemporal patterns, ultimately leading to improved accuracy and robustness in recognizing individual signs.

Before I3D, we utilized a model with late diffusion that involved generating feature maps for each frame using the pretrained MobileNet model. These sequential feature maps were then inputted into the LSTM model to address the final prediction problem. However, this approach was limited in capturing frame interactions as it only relied on late feature maps from the model, resulting in some loss of information about connections and interactions between frames. The core issue lay in capturing temporal features essential for identifying dynamic patterns. The I3D model tackled this limitation by incorporating continuous fusion techniques. This allowed for the extraction of both spatial and temporal features, capturing the dynamic patterns present in the target dataset more effectively.

During the training phase, we followed a two-step process. Initially, we trained the top layer of the model to fit the new initialized parameters to the previously trained parameters. This process helps prevent the top untrained parameters from interfering with the previously trained parameters. Then, we unfreeze the rest of the backbone layers and fine-tune the entire model using the target dataset.

The number of epochs during experiments typically varied between 50-100 depending on the convergence of the training loss and validation accuracy. During the first step of the training phase, we froze the backbone layers and only trained the top layer with a lower learning rate to focus on fine-tuning the new parameters. In the second step, we unfreezed the backbone layers and used a higher learning rate to fine-tune the entire model. Fig. 1 shows the result for word-level recognition within one sentence.

4. Conclusion

Our findings underscore the potential of repurposing pre-trained 3D CNNs—originally developed for human action recognition—to significantly advance word-level sign language recognition, with a particular emphasis on Azerbaijani Sign Language. By leveraging transfer learning techniques and fine-tuning a model initialized on a large-scale action recognition dataset, we achieved marked improvements in accuracy, demonstrating that knowledge acquired in a data-rich domain can successfully translate to a resource-constrained setting.

One of the key advantages observed was the enhanced training efficiency. Transfer learning not only accelerated the convergence process, requiring fewer epochs, but also provided a stronger initial baseline performance on the validation dataset. This improvement in starting accuracy suggests that the pre-trained network's latent knowledge of spatiotemporal patterns effectively generalizes to the sign language domain. Additionally, the model's real-time prediction capability—achieving rates of approximately 50 to 100 Hz—positions it as a practical solution for real-world applications where responsiveness is critical.

Looking ahead, we plan to broaden our dataset, incorporating a more extensive vocabulary of Azerbaijani signs and potentially exploring other underrepresented sign languages to further validate the generalizability of our approach. Another direction involves investigating attention-based architectures and comparing their performance against our current backbone. Attention mechanisms, known for their ability to highlight salient features within sequences, may offer additional gains in accuracy and robustness, particularly in complex signing scenarios where subtle differences in hand shape or movement trajectory are crucial.

In conclusion, our research presents a promising methodology for word-level sign language recognition that capitalizes on the strengths of pre-trained 3D CNNs from human action recognition tasks. This approach not only delivers strong initial results—achieving an F1 score of 84.85%—but also lays a foundation for future advancements in low-resource sign language processing. By continuing to refine our models, expand our datasets, and incorporate cutting-edge neural architectures, we aim to further close the gap between resource-rich and resource-scarce sign language domains, ultimately promoting more inclusive and accessible communication technologies.

References

1. Charuka, K., Wickramanayake, S., Ambegoda, T.D., Madhushan, P., Wijesooriya, D. (2024). Sign Language Recognition for Low Resource Languages Using Few Shot Learning. In: Luo, B., Cheng, L., Wu, ZG., Li, H., Li, C. (eds) Neural Information Processing. ICONIP 2023. Communications in Computer and Information Science, vol 1964. Springer, Singapore. https://doi.org/10.1007/978-981-99-8141-0_16
2. Chen, Y., Zuo, R., Wei, F., Wu, Y., Liu, S., & Mak, B. (2022, November 2). Two-Stream Network for Sign Language Recognition and Translation. Cornell University. https://doi.org/https://doi.org/10.48550/arxiv.2211.01367
3. Ji, Shuiwang & Xu, Wei & Yang, Ming & Yu, Kai. (2010). 3D Convolutional Neural Networks for Human Action Recognition. Pattern Analysis and Machine Intelligence, IEEE Transactions on. 35. 495-502. 10.1109/TPAMI.2012.59.
4. J. Arunnehru, G. Chamundeeswari, S. Prasanna Bharathi, Human Action Recognition using 3D Convolutional Neural Networks with 3D Motion Cuboids in Surveillance Videos Procedia Computer Science, Volume 133, 2018, Pages 471-477, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2018.07.059
5. Y. Xu, Y. Feng, Z. Xie, M. Xie and W. Luo, "Action Recognition Using High Temporal Resolution 3D Neural Network Based on Dilated Convolution," in *IEEE Access*, vol. 8, pp. 165365-165372, 2020, doi: 10.1109/ACCESS.2020.3022407.
6. S. Khurana, R. Sreemathy, M. Turuk and J. Jagdale, "Comparative Study and Performance Analysis of Deep Neural Networks for Sign Language Recognition using Transfer Learning," *2023 Third International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT)*, Bhilai, India, 2023, pp. 1-8, doi: 10.1109/ICAECT57570.2023.10118114.
7. Yan, S., Xiong, Y., & Lin, D. (2018, April 27). Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition. Proceedings of the AAAI Conference on Artificial Intelligence. https://doi.org/10.1609/aaai.v32i1.12328
8. Simonyan, K., & Zisserman, A. (2014). Two-stream convolutional networks for action recognition in videos. Advances in neural information processing systems, 27.
9. Y. Huang, Y. Guo and C. Gao, "Efficient Parallel Inflated 3D Convolution Architecture for Action Recognition," in *IEEE Access*, vol. 8, pp. 45753-45765, 2020, doi: 10.1109/ACCESS.2020.2978223.
10. Novopoltsev, M., Verkhovtsev, L., Murtazin, R., Milevich, D. \& Zemtsova, I. Fine-tuning of sign language recognition models: a technical report. (2023)
11. Sakshi Sharma, Sukhwinder Singh, ISL recognition system using integrated mobile-net and transfer learning method, Expert Systems with Applications, Volume 221, 2023, 119772, ISSN 0957-4174, https://doi.org/10.1016/j.eswa.2023.119772
12. H. Hameed *et al.*, "Privacy-Preserving British Sign Language Recognition Using Deep Learning," *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, Glasgow, Scotland, United Kingdom, 2022, pp. 4316-4319, doi: 10.1109/EMBC48229.2022.9871491.
13. Töngi, R. (2021, February 25). Application of Transfer Learning to Sign Language Recognition using an Inflated 3D Deep Convolutional Neural Network. arXiv.org. https://arxiv.org/abs/2103.05111
14. M. Vázquez-Enríquez, J. L. Alba-Castro, L. Docío-Fernández and E. Rodríguez-Banga, "Isolated Sign Language Recognition with Multi-Scale Spatial-Temporal Graph Convolutional Networks," 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Nashville, TN, USA, 2021, pp. 3457-3466, doi: 10.1109/CVPRW53098.2021.00385.
15. Wang, Fei & Du, Yuxuan & Wang, Guorui & Zeng, Zhen & Zhao, Lihong. (2022). (2+1) D-SLR: an efficient network for video sign language recognition. Neural Computing and Applications. 34. 10.1007/s00521-021-06467-9.
16. Fei Wang, Libo Zhang, Hao Yan, and Shuai Han. 2023. TIM-SLR: a lightweight network for video isolated sign language recognition. Neural Comput. Appl. 35, 30 (Oct 2023), 22265–22280. https://doi.org/10.1007/s00521-023-08873-7.

17. D. Li, C. R. Opazo, X. Yu and H. Li, "Word-level Deep Sign Language Recognition from Video: A New Large-scale Dataset and Methods Comparison," 2020 IEEE Winter Conference on Applications of Computer Vision (WACV), Snowmass, CO, USA, 2020, pp. 1448-1458, doi: 10.1109/WACV45572.2020.9093512.
18. G. Abdullayeva et al., Transfer learning for Azerbaijani Sign Language Recognition. Journal of Informatics and Control Problems. https://doi.org/10.54381/icp.2022.2.08.
19. J. Hasanov et al., Development of a hybrid word recognition system and dataset for the Azerbaijani Sign Language dactyl alphabet, Speech Communication, Volume 153, 2023, 102960, ISSN 0167-6393, https://doi.org/10.1016/j.specom.2023.102960.
20. A. Mino, M. Popa and A. Briassouli, "The Effect of Spatial and Temporal Occlusion on Word Level Sign Language Recognition," *2022 IEEE International Conference on Image Processing (ICIP)*, Bordeaux, France, 2022, pp. 2686-2690, doi: 10.1109/ICIP46576.2022.9897770.