Journal of Korea Robotics Society (2024) 19(4):328-334
https://doi.org/10.7746/jkros.2024.19.4.328 ISSN: 1975-6291 / eISSN: 2287-3961

수화 인식을 위한 다양한 액션 모델의 평가와 분석

Translating the Silent Voice: A Benchmark of Existing Action Models for Sign Language Recognition

사닷 타신[1] · 임 종 윤[2] · 브루스 맥도널드[3] · 안 호 석†
Sadat Taseen[1], Jong Yoon Lim[2], Bruce MacDonald[3], Ho Seok Ahn†

Abstract: Various different action recognition models have been used for Sign Language Recognition (SLR). Our research questions are 1) why a particular model was chosen, and 2) what the computational requirements, accuracy, and trade-offs of each architecture are. This paper experimentally studies the performance of different existing action recognition models against two SLR datasets: WLASL2000 and AUTSL. Our findings include one model named Channel-Separated Network (CSN) that achieves the highest accuracy in all benchmarks conducted with results comparable to the current state-of-the-art, SignBERT for WLASL100.

Keywords: Computer Vision, Sign Language Recognition

1. Introduction

Sign language is a visual language that conveys meaning through gestures, facial expressions, and body language. It is the first language of many deaf individuals and is used as a primary mode of communication in many deaf communities worldwide. However, automatic translation of sign language is a complex task requiring a deep understanding of natural language processing and computer vision. In its most basic form, Sign Language Recognition (SLR) is the task of classifying a gloss or an individual sign, such as the one in [Fig. 1] to their word meaning.

Over the years, there has been a massive interest in SLR within the computer vision community. Currently, most state-of-the-art models utilize the multi-stream model architecture, as shown in [Fig. 2], to extract features from different modalities. These modalities might include RGB, optical flow, depth, skeleton and sometimes, different crops of the signer, e.g., the hands and the face.

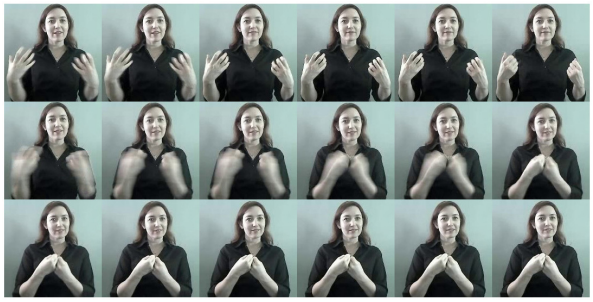
Aside from the modalities used, the only main difference between current state-of-the-art models is their choice of their backbones, i.e., the part of the model that works on the raw modality inputs. For the MSNN model in [Fig. 2], this is ST-GCN[1] for skeleton input and I3D[2] for all the other modalities.

Backbones have often been chosen more so for their popularity than their actual performance for SLR task as seen in [3], [4] and [5]. As a result, there is an opportunity to optimize such networks by selecting backbones that work best for SLR.

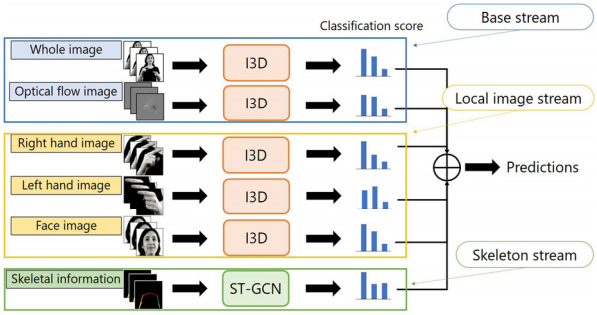
Our contributions are as follows:

· We run extensive experiments on existing action recognition models for RGB on three SLR datasets: WLASL100, AUTSL, and WLASL2000. We choose RGB as it is trivial

Received : Jul. 31. 2024; Revised : Aug. 16. 2024; Accepted : Oct. 2. 2024
※ This work was supported by the Industrial Fundamental Technology Development Program (20023495, Development of behavior-oriented HRI AI technology for long-term interaction between service robots and users) funded By the Ministry of Trade, Industry & Energy (MOTIE, Korea), the Te Pūnaha Hihiko: Vision Mātauranga Capability Fund (Te Ara Auaha o Muriwhenua Ngāti Turi: The Journey of Muriwhenua Māori Deaf Innovation, UOAX2124) funded by the Ministry of Business, Innovation & Employment (MBIE, New Zealand), and the Science for Technological Innovation (UOAX2123, Developing a Reo Turi (Māori Deaf Language) Interpreter for Ngāti Turi, the Māori Deaf Community). Ho Seok Ahn* is the corresponding author.
1. Student, CARES, University of Auckland, New Zealand (stas444@auckland.ac.nz)
2. Senior Research Engineer, CARES, University of Auckland, New Zealand (jy.lim@auckland.ac.nz)
3. Professor, CARES, University of Auckland, New Zealand (b.macdonald@auckland.ac.nz)
† Professor, Corresponding author: CARES, University of Auckland, New Zealand (hs.ahn@auckland.ac.nz)



[Fig. 1] American gloss for 'accident'[6]



[Fig. 2] Multi-Stream Neural Network (MSNN)[3]

to adapt the models to other input types.

· To better understand the current limitations of models on RGB, we select the best model and qualitatively analyze the glosses that it predicts wrong.

In Section 2 of this paper, we review existing sign language models and dataset and in Section 3, we discuss our experiments and their results. We conclude this paper in Section 4.

2. Related Works

2.1 Sign Language Recognition Research

Recent approaches for sign language recognition have entirely focused on using machine learning techniques to learn sign language patterns directly from the data. The main challenge with these approaches is in extending the existing computer vision model's capabilities to being able to interpret video since static images are insufficient for many of the glosses used. Many action recognition models, including 3DCNN[7], I3D, R(2+1)d[8], and newer transformer-based approaches, i.e., TimeSFormer[9] have been proposed over the years for the broader action recognition task, which have not yet been benchmarked against each other on sign language recognition datasets.

2.2 Existing Datasets

Since sign language differs for each country, sign language recognition datasets developed over recent years have different signs for the same words. The goal of these datasets is to provide a benchmark for different models so that they can be used for the harder Continuous Sign Language Recognition (CSLR) task, which involves translating sign language to sentences.

While some datasets e.g., ASSLLVD, DEVISIGN and WLASL boast high vocabulary sizes, they fall short on the number of samples per class as seen in [Table 1]. On the other hand, although datasets like CSL and AUTSL have much smaller vocabulary sizes, they have large numbers of samples per class, which generally allows the models trained on them to attain higher accuracies than the same models trained on their counterparts as seen in [Table 2]. With very few samples per class, the already difficult task of SLR becomes not only a problem of action recognition but also a problem of few-shot learning.

2.3 Models for Sign Language Recognition

All current state-of-the-art models exploit multimodal inputs to make their models output better predictions, as shown in [Table 2]. These multimodal inputs may include but are not limited to RGB, depth, optical flow, depth flow, HHA depth, skeleton, cropped hands, and cropped face. The concept behind this is to extract features from different modalities and combine them so that the classifier head has more information to work with. There is an accuracy-time trade-off with these models, as multimodal models tend to be more computationally intensive

[Table 1] Overview of existing large-scale SLR datasets, where a: the number of samples per class given is the mean number of samples per class, b: this dataset is not publicly available for use

Dataset	Sign Language Recognition Datasets				
	Language	*# of Signs*	*# of Samples*	*Samples per Class*[a]	*Ref.*
ASLLVD	American	2742	9794	3.6	[10]
DEVISIGN[b]	Chinese	2000	24000	12	[11]
MSASL	American	1000	25513	25.5	[12]
CSL[b]	Chinese	500	125000	250	[13]
WLASL	American	2000	21083	10.5	[6]
AUTSL	Turkish	226	38336	169.6	[14]

[Table 2] Current SLR model benchmarks on various datasets, where a: the number of samples per class given is the mean number of training samples per class

Dataset		Sign Language Recognition Model				
Name	*Samples per Class*[a]	*Year*	*Model*	*Modality*	*Accuracy (%)*	*Ref.*
WLASL100	17.8	2021	SignBERT	RGB, Pose, HandCrop	**83.30**	Hu et al. [4]
		2021	MSNN	RGB, Flow, Pose, HandCrop, FaceCrop	81.38	Maruyama et al. [3]
		2022	I3D	RGB	67.06	Kimsong et al. [15]
		2020	I3D Pose	RGB, Pose	65.89	Li et al. [16]
		2022	SPOTER	RGB, Pose	63.18	Bohacek & Hruz [17]
WLASL300	14.8	2021	SignBERT	RGB, Pose, HandCrop	**75.27**	Hu et al. [4]
		2021	MSNN	RGB, Flow, Pose, HandCrop, FaceCrop	73.43	Maruyama et al. [3]
		2022	SPOTER	RGB	68.75	Bohacek & Hruz [17]
		2020	I3D Pose	RGB, Pose	56.14	Li et al. [16]
WLASL2000	9.1	2021	SAM-SLR-v2	RGB, Flow, Depth, Depth Flow, Depth HHA, Pose	**59.39**	Jiang et al. [5]
		2021	SignBERT	RGB, Pose, HandCrop	52.08	Hu et al. [4]
		2021	MSNN	RGB, Flow, Pose, HandCrop, FaceCrop	47.26	Maruyama et al. [3]
AUTSL	111	2021	SAM-SLR-v2	RGB, Flow, Depth, Depth Flow, Depth HHA, Pose	**98.53**	Jiang et al. [5]
		2020	CNN+FPM+LSTM+Attention	RGB, Depth	95.95	Ozge & Hacer [14]
		2021	VTN-PF	RGB, Pose, Flow	92.92	Mathieu et al. [18]
MSASL	21.5	2021	SignBERT	RGB, Pose, HandCrop	**89.96**	Hu et al. [4]
		2021	MSNN	RGB, Flow, Pose, HandCrop, FaceCrop	84.22	Maruyama et al. [3]

than their unimodal counterparts, making them slow for real-time predictions.

3. Experiments

In this section, we run experiments on the different action recognition models to empirically evaluate their effectiveness in SLR.

First, we train the models on two different datasets: WLASL and AUTSL, using only the RGB modality. We select WLASL for its large vocabulary size and use two subsets: one with 100 signs and another with full 2000 signs to test the performance of the models with different number of classes. In contrast, we choose AUTSL for its high number of training samples per sign and to test the robustness of the different models under dynamic backgrounds and settings, as shown in [Fig. 3].

We train each model three times, using different seeds on the training and validation splits of the dataset. For each run, we evaluate using the top-1 and top-5 accuracies for each model. The final evaluation uses the mean top-1 and top-5 accuracies from the three runs.

Due to time constraints, we use the stock model with the

[Fig. 3] Sample frames from AUTSL (left)[6] and WLASL (right)[14]

default data augments and hyperparameters available from MMAction2[19]. Each model is trained for 150 epochs or until the validation top-1 accuracy converges to a set value. For the evaluation phase, all models are tested on the WLASL100 subset, but due to limited time, we only test the best-performing models on the larger datasets.

3.1 Quantitative Results

CSN outperforms all the other models for WLASL100 with a top-1 accuracy of 79.20% followed closely by TPN and TimeSformer as seen in [Table 3].

Since WLASL100 has a relative low number of samples per class, we previously speculated that the transformer-based model, TimeSFormer did not have enough data to have adequate performance. However, even after evaluating against the large AUTSL dataset, we still find CSN and TPN outperforming TimeSFormer by about 17% as seen in [Table 4].

In contrast to other action recognition datasets where transformers achieve state-of-the-art results, SLR datasets often contain low number of training samples per class, which benefits CNN designs due to their built-in inductive biases that transformer-based vision models lack.

Since both CSN and TPN performed at a similar level for AUTSL, as a 'tiebreaker' between CSN and TPN, we evaluate the two against the WLASL2000 dataset, which is a dataset to test models under many classes. This time, we find CSN achieving a large lead of 14.37% from TPN as seen in [Table 5].

Despite its high vocabulary size, the WLASL2000 dataset falls behind in the number of samples per class, with a mean of only 9.1 training samples per class. We infer from the results that TPN struggles from a low number of training samples and not the high number of training classes. This is evident from the results of WLASL100 where it achieves a lower accuracy than CSN and

[Table 3] Model Performance against WLASL100, where *: the results shown are the accuracies from evaluating the model against the test data

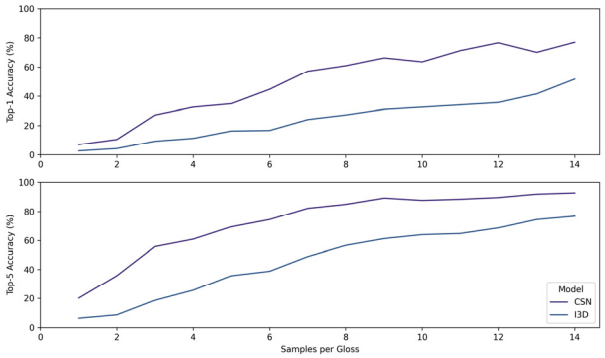
Model		Seed 0		Seed 1		Seed 2		Mean	
Name	*Ref.*	*Top-1* (%)*	*Top-5* (%)*	*Top-1* (%)*	*Top-5* (%)*	*Top-1* (%)*	*Top-5* (%)*	*Top-1* (%)*	*Top-5* (%)*
C3D	[20]	50.77	79.45	48.45	79.46	49.22	79.84	49.48	79.58
TSN	[21]	44.57	78.29	47.67	78.68	45.61	77.78	45.95	78.25
TSM	[22]	55.81	86.82	55.81	87.6	55.81	86.05	55.81	86.82
TANet	[23]	58.91	82.95	57.36	87.21	60.08	84.1	58.78	84.75
CSN	[24]	77.91	93.02	79.84	93.02	79.84	93.02	**79.20**	**93.02**
TimeSformer	[9]	62.4	84.88	61.63	84.11	61.24	84.11	*61.76*	*84.37*
R(2+1)d	[8]	23.26	12.02	29.46	68.6	30.23	66.67	27.65	49.10
I3D	[2]	60.77	60.08	60.85	60.08	57.93	57.75	59.85	59.30
TPN	[25]	67.44	88.37	62.79	90.7	58.14	88.37	62.79	89.15
TIN	[26]	56.98	85.27	53.1	82.95	37.21	72.48	49.10	80.23

[Table 4] Model Performance against AUTSL

Model		Seed 0		Seed 1		Seed 2		Mean	
Name	*Ref.*	*Top-1* (%)*	*Top-5* (%)*	*Top-1* (%)*	*Top-5* (%)*	*Top-1* (%)*	*Top-5* (%)*	*Top-1* (%)*	*Top-5* (%)*
CSN	[24]	90.25	98.72	93.32	99.36	89.01	98.72	**90.86**	**98.93**
TimeSFormer	[9]	76.58	93.45	70.22	90.99	72.84	92.68	73.21	92.37
TPN	[25]	89.82	98.93	89.23	98.34	89.54	98.56	89.53	98.61

[Table 5] Model Performance against WLASL2000

Model		Seed 0		Seed 1		Seed 2		Mean	
Name	*Ref.*	*Top-1* (%)*	*Top-5* (%)*	*Top-1* (%)*	*Top-5* (%)*	*Top-1* (%)*	*Top-5* (%)*	*Top-1* (%)*	*Top-5* (%)*
CSN	[24]	46.02	77.25	44.04	77.28	38.87	71.48	42.98	75.34
TPN	[25]	30.64	65.72	28.45	62.28	26.75	59.22	28.61	62.41



[Fig. 4] Samples per class against accuracy for I3D and CSN

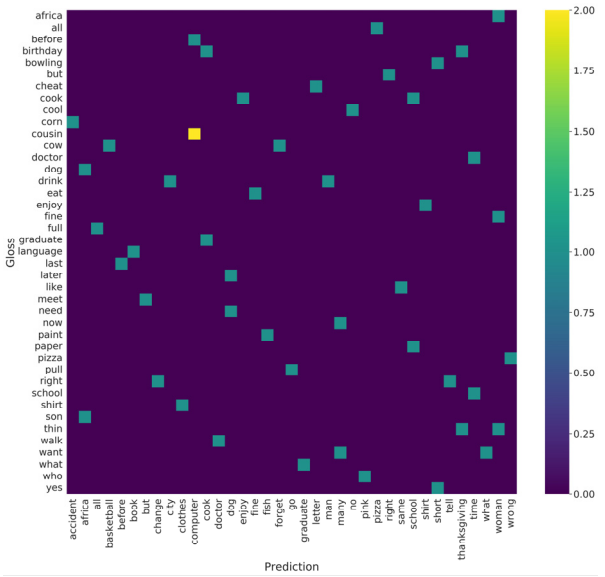
from the results of AUTSL where it achieves an accuracy about the same level as CSN while working with 126 more classes than the WLASL100.

Compared to other CNN-based models, CSN has a unique advantage in relation to the lower number of samples per class, prevalent in all sign language datasets. The channel separation in CSN factorizes the usual 3D convolution into channel and spatial-temporal components. This acts as a form of regularization as the channel interactions and spatial-temporal interactions are separated to different parts of the network[24]. As a result, this makes CSNs inherently less prone to overfit to training data, allowing for a higher testing accuracy with lower number of samples, compared to its CNN counterparts. This is seen in [Fig. 4] where we compare the performance of CSN with I3D at different number of training samples per class.

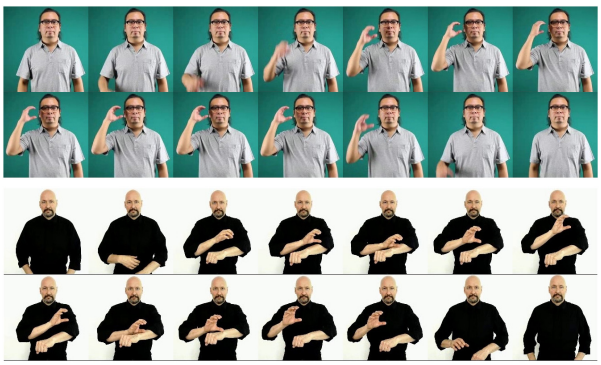
3.2 Qualitative Results

Since CSN attains an accuracy of 79.20%, which is only 4.1% behind the current state-of-the-art, SignBERT for WLASL100, we use this model to gain valuable insights in understanding the glosses it predicts wrong. Among these glosses, the gloss ‘cousin’ confuses the model the most into predicting the gloss, ‘computer’ as seen in [Fig. 5].

Both glosses are similar as they both have one hand making a ‘C’ shape. The only difference is that for ‘computer’, another arm needs to be underneath the main arm as seen in [Fig. 6] while for ‘cousin’, the other arm can be resting. We speculate that this is due to a lack of skeletal understanding of the model as it misses the other arm’s position every time and predicts the ‘computer’ gloss instead during training. This is why models like SignBERT rely on the pose modality to being able to distinguish between glosses like these.



[Fig. 5] Heatmap showing the wrong predictions of CSN on WLASL100



[Fig. 6] Ground truth ‘cousin’ (top) against predicted ‘computer’ (bottom)[14]

In [Fig. 7], both glosses have the same arm motion, but the starting position is at the chin for ‘never’ while it being the nose for ‘who’. For glosses, ‘heavy’ and ‘lightweight’ in [Fig. 8], we find an interesting insight in understanding sign language. Both glosses have the same hand movement with the only distinguishing factor being their face. For ‘lightweight’, the facial emotion is neutral but for ‘heavy’, the facial emotion is of a gruntled one. It is highly likely that the model misses this difference due to this feature being lost with convolutions since the face is a very small part of the whole frame. Many multi-modal models usually include a cropped face (and sometimes, also hands) as separate inputs such that these features don’t get lost with convolutions.

It should also be noted that facial expressions don’t always have any meaning for most glosses. The gloss, ‘accident’ for

[Fig. 7] Glosses ‘never’ (top) and ‘who’ (bottom)[14]

[Fig. 8] Glosses ‘lightweight’ (top) and ‘heavy’ (bottom)[14]

example in [Fig. 1], can have a positive or a negative connotation given the context. From a gloss standpoint, however, there can be no context. This raises an interesting issue with the current datasets as there is a need for a variety of facial emotions such that glosses where facial features are important, the model can have a clear indication of it.

4. Conclusions

In this paper, we run extensive experiments on various existing action recognition models to benchmark them against SLR datasets. Our findings include the best performing model, CSN, which achieves a high accuracy of 79.20% using RGB, that is only 4.1% behind the current state-of-the-art, SignBERT for WLASL100. We discuss the inherent drawbacks of using only one modality, the importance of facial and hand features, and how CSN has an inherent advantage over other models due to its built-in regularization for datasets with low number of training samples, which is common for sign language datasets.

To the best of our knowledge, CSN has not been used previously for SLR even though it has inherent advantages over other models.

For future work, we recommend exploring combinations of multiple modalities to get more comprehensive insights for sign language understanding. We hope this research serves as a baseline for future researchers and moving forward, we hope to see CSN being used as a strong backbone for multi-stream models to further improve the accuracy of word level sign language recognition.

References

[1] S. Yan, Y. Xiong, and D. Lin, “Spatial temporal graph convolutional networks for skeleton-based action recognition,” *AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, Apr., 2018, DOI: 10.1609/aaai.v32i1.12328.

[2] J. Carreira and A. Zisserman, “Quo vadis, action recognition? a new model and the kinetics dataset,” *arXiv:1705.07750,* 2017, DOI: 10.48550/arXiv.1705.07750.

[3] M. Maruyama, S. Ghose, K. Inoue, P. P. Roy, M. Iwamura, and M. Yoshioka, “Word-level sign language recognition with multi-stream neural networks focusing on local regions,” *arXiv:2106.15989*, 2021, DOI: 10.48550/arXiv.2106.15989.

[4] H. Hu, W. Zhao, W. Zhou, Y. Wang, and H. Li, “Signbert: pre-training of hand-model-aware representation for sign language recognition,” *arXiv:2110.05382,* 2021, DOI: 10.48550/arXiv.2110.05382.

[5] S. Jiang, B. Sun, L. Wang, Y. Bai, K. Li, and Y. Fu, “Skeleton aware multi-modal sign language recognition,” *arXiv:2103.08833*, 2021, DOI: 10.48550/arXiv.2103.08833.

[6] 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,” *arXiv:1910.11006*, 2020, DOI: 10.48550/arXiv.1910.11006.

[7] L. Zhang, G. Zhu, P. Shen, J. Song, S. A. Shah, and M. Bennamoun, “Learning spatiotemporal features using 3dcnn and convolutional lstm for gesture recognition,” *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, Venice, Italy, pp. 3120-3128, 2017, DOI: 10.1109/ICCVW.2017.369.

[8] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, “A closer look at spatiotemporal convolutions for action recognition,” *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, pp. 6450-6459, 2018, DOI: 10.1109/CVPR.2018.00675.

[9] G. Bertasius, H. Wang, and L. Torresani, “Is space-time attention all you need for video understanding?,” *arXiv:2102.05095,* 2021, DOI: 10.48550/arXiv.2102.05095.

[10] C. Neidle, A. Opoku, C. Ballard, K. M. Dafnis, E. Chroni, and D. Metaxas, “Resources for Computer-Based Sign Recognition from Video, and the Criticality of Consistency of Gloss Labeling across Multiple Large ASL Video Corpora,” *LREC2022 10th Workshop on the Representation and Processing of Sign Languages: Multilingual Sign Language Resources*, Marseille, France, pp. 165-172, 2022, [Online], https://hdl.handle.net/2144/45985.

[11] L. R. Cerna, E. E. Cardenas, D. G. Miranda, D. Menotti, and G. Camara-Chavez, “A multimodal LIBRAS-UFOP Brazilian sign language dataset of minimal pairs using a microsoft Kinect sensor,” *Expert Systems with Applications*, vol. 167, Apr., 2021,

DOI: 10.1016/j.eswa.2020.114179.

[12] H. R. Vaezi Joze and O. Koller, "Ms-asl: A large-scale data set and benchmark for understanding American sign language," *arXiv:1812.01053*, 2018, DOI: 10.48550/arXiv.1812.01053.

[13] L.-G. Zhang, X. Chen, C. Wang, Y. Chen, and W. Gao, "Recognition of sign language subwords based on boosted hidden Markov models," *7th international conference on Multimodal interfaces*, pp. 282-287, 2005, DOI: 10.1145/1088463.1088511.

[14] O. M. Sincan and H. Y. Keles, "Autsl: A large scale multi-modal turkish sign language dataset and baseline methods," *IEEE Access*, vol. 8, pp. 181340-181355, 2020, DOI: 10.1109/ACCESS.2020.3028072.

[15] E. Liu, J. Y. Lim, B. MacDonald, and H. S. Ahn, "Weighted Multi-modal Sign Language Recognition," *2023 Australasian Conference on Robotics and Automation (ACRA 2023)*, Islamabad, Pakistan, 2023, [Online], https://www.araa.asn.au/conference/acra-2023.

[16] 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," *arXiv:1910.11006,* 2020, DOI: 10.48550/arXiv.1910.11006.

[17] M. Boháček and M. Hrúz, "Sign pose-based transformer for word-level sign language recognition," *2022 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)*, Waikoloa, HI, USA, pp. 182-191, 2022, DOI: 10.1109/WACVW54805.2022.00024.

[18] M. De Coster, M. Van Herreweghe, and J. Dambre, "Isolated sign recognition from rgb video using pose flow and self-attention." *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW),* Nashville, TN, USA, pp. 3441-3450, 2021, DOI: 10.1109/ACCESS.2019.2946870.

[19] *OpenMMLab's Next Generation Video Understanding Toolbox and Benchmark,* [Online], https://github.com/open-mmlab/mmaction2. Licensed under Apache-2.0, Accessed: 13 Oct., 2023.

[20] Y. Jing, J. Hao and P. Li, "Learning Spatiotemporal Features of CSI for Indoor Localization With Dual-Stream 3D Convolutional Neural Networks," *IEEE Access*, vol. 7, pp. 147571-147585, 2019, DOI: 10.1109/ACCESS.2019.2946870.

[21] 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," *Computer Vision –ECCV 2016*, pp. 20-36, 2016, DOI: 10.1007/978-3-319-46484-8_2.

[22] J. Lin, C. Gan, and S. Han, "Tsm: Temporal shift module for efficient video understanding." *arXiv:1811.08383*, 2019, DOI: 10.48550/arXiv.1811.08383.

[23] Z. Liu, L. Wang, W. Wu, C. Qian, and T. Lu, "Tam: Temporal adaptive module for video recognition," *2021 IEEE/CVF International Conference on Computer Vision (ICCV),* Montreal, QC, Canada, pp. 13708-13718, 2021, DOI: 10.1109/ICCV48922.2021.01345.

[24] D. Tran, H. Wang, M. Feiszli, and L. Torresani, "Video classification with channel-separated convolutional networks," *2019 IEEE/CVF International Conference on Computer Vision (ICCV),* Seoul, Republic of Korea, pp. 5552-5561, 2019, DOI: 10.1109/ICCV.2019.00565.

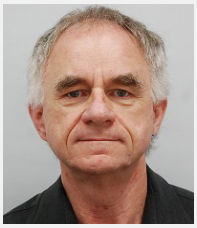
[25] C. Yang, Y. Xu, J. Shi, B. Dai, and B. Zhou, "Temporal pyramid network for action recognition," *arXiv:2004.03548*, 2020, DOI: 10.48550/arXiv.2004.03548.

[26] H. Shao, S. Qian, and Y. Liu, "Temporal interlacing network," *AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, pp. 11966-11973, 2020, DOI: 10.1609/aaai.v34i07.6872.

Sadat Taseen

2020~현재 University of Auckland Electrical, Computer and Software Engineering(학사)

관심분야: Computer Vision, Robotics



Bruce MacDonald

1978 University of Canterbury Electrical Engineering(학사)

1983 University of Canterbury Electrical Engineering(박사)

1995~현재 The University of Auckland, 교수

관심분야: 농업로봇, 로보틱스

임 종 윤

2007 연세대학교 컴퓨터공학과(학사)

2021 University of Auckland Software Engineering(석사)

2007~2021 삼성전자 연구원

2015~현재 The University of Auckland 연구원

관심분야: Computer Vision, NLP, Robotics

안 호 석

2005 성균관대학교 정보통신공학부(학사)

2010 서울대학교 전기컴퓨터공학부(박사)

2010-2012 한국생산기술연구원 선임연구원

2012-2013 일본 전자통신기초기술연구원 선임연구원

2013~현재 The University of Auckland, 부교수

관심분야: 인공지능, HRI, 소셜로봇