Benchmarking Data Augmentation for Contrastive Learning in Static Sign Language Recognition

Ariel Basso Madjoukeng[1] Jérôme Fink[1]
Pierre Poitier [*1] Edith Bélise Kenmogne[2] Benoît Frénay[1] [†]

1- University of Namur - NaDI - PReCISE - HuMaLearn
Rue Grangagnage, 21, 5000 Namur - Belgium

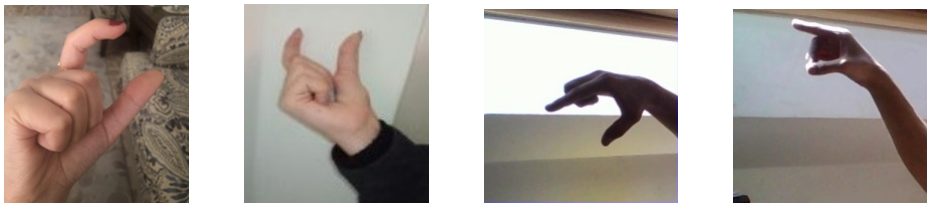
2- University of Dschang - Faculty of Sciences
Foto, 96, Dschang, Cameroon

Abstract. Sign language (SL) is a communication method used by deaf people. Static sign language recognition (SLR) is a challenging task aimed at identifying signs in images, for which acquisition of annotated data is time-consuming. To leverage unannotated data, practitioners have turned to unsupervised methods. Contrastive representation learning proved to be effective in capturing important features from unannotated data. It is known that the performance of the contrastive model depends on the data augmentation technique used during training. For various applications, a set of effective data augmentation has been identified, but it is not yet the case for SL. This paper identifies the most effective augmentation for static SLR. The results show a difference in accuracy of up to 30% between appearance-based augmentations combined with translations and augmentations based on rotations, erasing, or vertical flips.

1 Introduction

To achieve high-quality results, current deep learning methods require a large amount of annotated data. In the case of sign languages (SLs), the acquisition process requires skilled labor, making it slow and costly, limiting the amount of annotated data [1]. SL is far from being the only field with limited access to annotated data, and this motivated the creation of several unsupervised learning schemes. A popular approach is contrastive algorithms [2, 3]. Such algorithms create artificial labels by applying augmentations to each instance and learning common representations between those augmented instances. The resulting contrastive model could, then, be fine-tuned on a limited amount of annotated data. The performance that can be expected from the fine-tuned model is correlated with the data augmentation used during pre-training [4, 5]. Therefore, selecting the correct set of augmentation is critical. Currently, no work has evaluated the impact of augmentation on SL recognition. Figure 1 shows signs from the American Sign Language (ASL) and Arabic Sign Language (ArASL) that can be confused if a simple rotation is applied. This highlights the importance of finding the most relevant data augmentation in the case of SL (e.g. avoid rotations).

[*]Ph.D. grant from FRIA (F.R.S.-FNRS)
[†]Supported by SPWR under grant n°2010235 - ARIAC by DIGITALWALLONIA4.AI.



(a) *Dal* [6] (b) *Noon* [6] (c) Letter P [7] (d) Letter Q [7]

Fig. 1: Two signs (a) and (b) from ArASL [6] and two signs (c) and (d) from ASL [7] showcasing signs that can be easily confused if a rotation is applied.

Our experiments show that choosing one augmentation over another can lead to performance variations of up to 30% during the fine-tuning of the model. Beyond SLR, these results confirm how important the choice of a suitable augmentation can be to leverage contrastive learning. This paper evaluates the performance obtained with several augmentation techniques on SimCLR [2] and MoCo v2 [3], two common contrastive algorithms. Section 2 introduces contrastive learning, focussing on SimCLR and MoCo v2 used in the experiments. Section 3 presents a brief review of SLR literature, and finally Section 4 presents the experiments conducted and the results obtained.

2 Contrastive Representation Learning

Contrastive learning leverages self-supervised labels to learn useful representations of data, that can be reused to solve several downstream tasks. This method is also used to align representations between multimodal inputs [8]. Its versatility and simplicity lead to the creation of several algorithms, such as the popular SimCLR [2] and MoCo v2 [3] that involve several key steps. First, the **data augmentation** module generates two distinct views of each input image with data augmentation, forming positive pairs. Next, these augmented images are processed by an **encoder** to obtain their representations. For our experiments, a ResNet-50 pre-trained on ImageNet is used as the encoder [9, 10]. The obtained representations are then passed through a **projection function** consisting of a multilayer perceptron that maps them to lower-dimensional vectors. Finally, a **contrastive loss** function is applied to minimize the distance between the embeddings of the positive pairs while increasing the one for the negative pairs (i.e., all the other augmented images, except for the positive pairs). The main difference between the SimCLR and MoCo v2 algorithms is that SimCLR uses only the images from the current batch at each iteration, while MoCo v2 maintains a queue that stores elements from both the current and previous batches and uses two encoders, augmenting the number of negative pairs considered.

3 Related Work

Contrastive learning enables to learn useful representation efficiently from unannotated data [2]. Zhao et al. [11] conducted a study on self-supervised learning methods and concluded that contrastive methods are efficient in several tasks such as hand motion estimation. Experiments also show that performances obtained with contrastive learning depend on the augmentation used therefor. Several works [12, 13, 14] have studied the impact of data augmentation for their domain, but their results remain limited in terms of transferability. For example, in facial recognition, the application of Generative Adversarial Networks and horizontal flip yields the best results [14]. In other fields such as plant anomaly detection and human activity recognition, rotations perform best [12, 15].

To overcome the data scarcity issue in static SLR, the usage of pre-trained models was studied [9, 10], but SLR datasets are often small and need to be augmented during training to obtain satisfactory results. The application of few-shot learning [16] or contrastive learning approach [17] was also studied to assess the effectiveness of those methods in the case of SLR. In particular, constrastive learning has the potential to leverage the non-annotated parts of SLR datasets, whose size may be significantly larger than their annotated counterparts. However, none of the above work provide experimental evidence about the best set of augmentation to use in the case of SLR. Thus, this research aims to fill this gap by identifying the most effective augmentation(s) for SLR. We focus on the case of contrastive learning because (i) it uses augmentation at its very core (see Section 2) and (ii) it is a promising trend in SLR [18, 17].

4 Data Augmentation for Contrastive Learning in SLR

This section evaluates the relevance of several augmentations for SLR and gives more insight to the practitioner on the augmentation to use. The most common augmentation for images [2] are considered: rotation, translation, vertical and horizontal flip, Gaussian blur, erasing, and color distortion.

4.1 SLR Datasets

To determine the most effective data augmentation, three SLR datasets of different sizes were used: (i) the small *Bahasa Isyarat Indonesia* (BISINDO) [19] dataset with 312 images from 26 classes, (ii) the medium *Arabic Sign Alphabet* (ArASL) [6] dataset with 7,856 images from 31 classes and (iii) the large *American Sign Language* (ASL) [7] dataset with 87,000 images from 26 classes. These SLR datasets were chosen to be diverse in terms of size and variety.

4.2 Training Process and Hyper-Parameters

To evaluate the impact of each data augmentation, we divided each dataset into three sets: a test set made of the 30% of the data. The 70% remaining data are used to pre-train and fine-tune the model with 80% of the training set used

for the pre-training 20% for fine-tuning. For each augmentation listed in the beginning of Section 4, several variants are created (e.g., rotations at different angles, coloration with various effects, etc.) and the SimCLR and MoCo v2 algorithms were trained using those augmentations.

The experiments are run on PyTorch version 2.4.1+cu118, using a 24GB VRAM GPU. All images are resized to 255x255 pixels. The models are trained on 70 epochs, with batch sizes of 128. The contrastive models are trained with the SGD optimizer and a learning rate of 0.001. The momentum of the encoder for MoCo v2 has a value of 0.99 and the temperature is set to 0.1, as recommended.

For each augmentation, 10 variants of the same augmentation with different parameters are created and the SimCLR and MoCo v2 models are trained ten times. After the contrastive training, they are fine-tuned on the remaining data. The performance reported for each augmentation is the average of the performances obtained over the ten training runs on that augmentation.

4.3 Experimental Results for Contrastive Learning in SLR

Table 1 presents the results obtained on each dataset. From this table, we can observe that for datasets like ASL and ArASL, a performance difference of over 30% is observed depending on the choice of augmentation. Moreover, the SimCLR algorithm generally yields better results than MoCo on these datasets, highlighting its effectiveness on static SLR. Additionally, despite the small size of the BISINDO dataset, a performance difference of 9% is also observed. Augmentations such as rotation, erasing, and vertical flips do not help the SLR models. This can be explained by the fact that the sign orientation plays a crucial role in its meaning, and erasing it might damage important parts of the image. Augmentations related to the appearance of the image significantly improve the SLR model accuracy, as they not only preserve the geometric structure of the image but also help the model recognize the same sign with various shades and colors. Translation also works well because it teaches the model to recognize the same sign regardless of its position.

5 Conclusion

This work evaluates the impact of image augmentations for static SLR using contrastive representation learning. It is shown that, for static SLR, the most effective augmentation technique includes Gaussian blur, color distortion, and translation. Those augmentations enhance the mode robustness to variations in lighting conditions and the positioning of signs. In contrast, rotations, vertical flips, and erasing are damageable to the performance. This can be explained by the fact that the orientation of the hand plays a major role in the meaning of a sign. Additionally, erasing does not yield good performance, most likely because removing parts of a sign image can completely distort its meaning, the sign is located in a small proportion of the image. This work highlights the importance of choosing the proper augmentation for a given task. In future work, we will expand this research by considering dynamic signs and video-based SLR.

Table 1: Accuracy after the fine-tuning of a contrastive backbone trained using various image augmentation methods. For each dataset, the best results are highlighted in light blue and bold, while the worst results are in yellow.

Datasets	Augmentations	SimCLR	MoCo v2
ASL	Rotation (R_t)	39.22%	36.07%
	Vertical Flip (V_f)	34.58%	33.20%
	Horizontal Flip (H_f)	49.54%	47.06%
	Erasing (E_r)	33.72%	32.52%
	Color Distortion (C_d)	69.32%	63.32%
	Gaussian Blur (G_b)	64.72%	61.73%
	Translation (T_r)	65.26%	64.01%
	C_d and G_b	72.05%	68.78%
	C_d and V_f	50.17%	42.25%
	C_d and H_f	54.31%	55.61%
	T_r and C_d	75.49%	69.07%
	R_t and E_r	28.03%	30.7%
	All Augmentation	57.23%	51.23%
	T_r, C_d and G_b	**79.07 %**	**72.07%**
ArASL	Rotation (R_t)	24.26%	24.71%
	Vertical Flip (V_f)	24.53%	24.98%
	Horizontal Flip (H_f)	29.24%	25.04%
	Erasing (E_r)	23.72%	21.36%
	Color Distortion (C_d)	40.36%	36.53%
	Gaussian Blur (G_b)	36.97%	32.07%
	Translation (T_r)	42.02%	41.92%
	C_d and G_b	45.29%	46.11%
	T_r and V_f	39.15%	33.17%
	T_r and H_f	42.03%	39.54%
	R_t and E_r	20.30%	20.07%
	T_r and C_d	46.23%	45.81%
	All Augmentation	41.74%	39.89%
	T_r, G_b and C_d	**54.03%**	**49.83%**
BISINDO	Rotation (R_t)	3.42%	3.63%
	Vertical Flip (V_f)	4.58%	5.02%
	Horizontal Flip (H_f)	6.24%	6.84%
	Erasing (E_r)	3.62%	3.06%
	Color Distortion (C_d)	9.52%	8.08%
	Gaussian Blur (G_b)	7.95%	7.65%
	Translation (T_r)	9.32%	9.01%
	C_d and G_b	11.07%	**11.52%**
	C_d and V_f	7.78%	7.53%
	C_d and H_f	9.96%	9.60%
	R_t and E_r	2.89%	2.17%
	T_r and C_d	11.32%	11.23%
	All Augmentation	6.74%	7.28%
	T_r, C_d and G_b	**12.09%**	11.11%

References

[1] Mathieu De Coster, Dimitar Shterionov, Mieke Van Herreweghe, and Joni Dambre. Machine translation from signed to spoken languages: State of the art and challenges. *Universal Access in the Information Society*, pages 1305–1331, 2024.

[2] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607, 2020.

[3] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

[4] Jiyang Zheng, Yu Yao, Bo Han, Dadong Wang, and Tongliang Liu. Enhancing contrastive learning for ordinal regression via ordinal content preserved data augmentation. In *12th International Conference on Learning Representations*, 2024.

[5] Haolin Pan, Yong Guo, Qinyi Deng, Haomin Yang, Jian Chen, and Yiqun Chen. Improving fine-tuning of self-supervised models with contrastive initialization. *Neural Networks*, pages 198–207, 2023.

[6] Muhammad Al-Barham. RGB arabic alphabets sign language dataset, 2023.

[7] Akash Nagaraj. Kaggle asl alphabet, 2018.

[8] Alec Radford, Jong Wook Kim, Chris Hallacy, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763, 2021.

[9] Bader Alsharif, Ali Salem Altaher, Ahmed Altaher, et al. Deep learning technology to recognize american sign language alphabet. *Sensors*, 23(18):7970, 2023.

[10] Baraa Wasfi Salim and SR Zeebaree. Kurdish sign language recognition based on transfer learning. *International Journal of Intelligent Systems and Applications in Engineering*, pages 232–245, 2023.

[11] Zehui Zhao, Laith Alzubaidi, Jinglan Zhang, Ye Duan, and Yuantong Gu. A comparison review of transfer learning and self-supervised learning: Definitions, applications, advantages and limitations. *Expert Systems with Applications*, page 122807, 2023.

[12] Kenichi Kobayashi, Junpei Tsuji, and Masato Noto. Evaluation of data augmentation for image-based plant-disease detection. In *IEEE international conference on systems, man, and cybernetics (SMC)*, pages 2206–2211. IEEE, 2018.

[13] Phillip Chlap, Hang Min, Nym Vandenberg, et al. A review of medical image data augmentation techniques for deep learning applications. *Journal of Medical Imaging and Radiation Oncology*, pages 545–563, 2021.

[14] Simone Porcu, Alessandro Floris, and Luigi Atzori. Evaluation of data augmentation techniques for facial expression recognition systems. *Electronics*, page 1892, 2020.

[15] Chi Ian Tang, Ignacio Perez-Pozuelo, Dimitris Spathis, and Cecilia Mascolo. Exploring contrastive learning in human activity recognition for healthcare. *arXiv preprint arXiv:2011.11542*, 2020.

[16] Irma Permata Sari. Closer look at image classification for indonesian sign language with few-shot learning using matching network approach. *International Journal on Informatics Visualization*, pages 638–643, 2023.

[17] Shiwei Gan, Yafeng Yin, Zhiwei Jiang, Kang Xia, Lei Xie, and Sanglu Lu. Contrastive learning for sign language recognition and translation. In *IJCAI*, pages 763–772, 2023.

[18] Zifan Jiang, Gerard Sant, Amit Moryossef, et al. SignCLIP: Connecting text and sign language by contrastive learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9171–9193, 2024.

[19] Achmad Noer. Bahasa isyarat indonesia (bisindo) alphabets. Kaggle Dataset.