ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

Movements as Images: CNNs are Good Feature Extractors in Sign Language Recognition

Pierre Poitier *, Loïc Brangier, Ariel Basso Madjoukeng and Benoît Frenay

University of Namur - Computer Science Department
rue Grandgagnage 61, Namur - Belgium

Abstract. This work explores a simple approach to Isolated Sign Language Recognition (ISLR) by reframing the classification of pose sequences as a standard image classification task. While recent trends in Sign Language Processing (SLP) heavily favor complex temporal architectures like Transformers, we investigate the projection of spatio-temporal pose information into a static image representation. By mapping time and skeletal joints to spatial dimensions and coordinate values to color channels, we allow standard Convolutional Neural Networks (CNNs), like ResNets, to extract features effectively. Our experiments on challenging real-world ISLR datasets demonstrate that this method is not only computationally efficient, but also outperforms existing architectures like Pose-VIT and SPOTER in a simple classification setting.

1 Introduction

Deep learning has driven huge progress in computer vision and natural language processing. This success has naturally cascaded into the domain of Sign Language Processing (SLP), providing new tools to tackle the complexity and diversity of sign languages. Within SLP, we focus specifically on the task of Isolated Sign Language Recognition (ISLR), where the primary objective is to predict the class label associated with a video segment containing a single sign.

While recent works have achieved impressive results on ISLR datasets recorded in controlled environments, the transition to datasets derived from real-world sign language discourse remains a significant hurdle [1, 2]. In these "natural discourse" scenarios, the task becomes significantly more challenging. Signs are often articulated rapidly, and their visual appearance is heavily influenced by linguistic phenomena [3] such as prosody and coarticulation, where the formation of a sign is altered by the preceding and succeeding signs. To address these challenges, the community has largely adopted complex architectures designed to model temporal dependencies explicitly. Popular approaches include 3D-CNNs (e.g., I3D) for full-frame video processing [4], and hybrid models combining 2D-CNNs with recurrent units like LSTMs or GRUs [1]. More recently, the paradigm has shifted towards Transformer-based architectures [5], which are specifically designed for SLP tasks. However, in this work, we deviate from this complexity and propose a simpler approach. We consider the signer's entire movement sequence as a single, static image, rather than processing it as a temporal stream

*Support of the Walloon region with a Ph.D. grant from FRIA (F.R.S.-FNRS).

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

of frames. To the best of our knowledge, this approach has not been rigorously applied to challenging SLR datasets built from continuous discourse.

The remainder of this paper is organized as follows: Section 2 reviews the related work. Section 3 details our methodology for projecting pose sequences into static images while preserving critical spatio-temporal information. Section 4 evaluates the performance of off-the-shelf CNNs on this representation in a fully supervised setting and demonstrates that ResNet models act as powerful feature extractors in ISLR. The latter also provides a qualitative analysis of the features: how they can be interpreted by a human, and how models use them.

2 Related Work

This section outlines the current state of the art in Isolated Sign Language Recognition (ISLR). We first discuss prevalent deep learning architectures, followed by an overview of the specific characteristics of datasets used in this domain.

2.1 Deep Learning Architectures in SLR

The evolution of SLR models mirrors the broader computer vision landscape. Early deep learning approaches relied heavily on frame-wise feature extraction using 2D-CNNs, aggregated temporally via pooling or Recurrent Neural Networks (RNNs) such as LSTMs and GRUs [1, 5]. With the advent of more powerful computing capabilities, 3D-CNNs like I3D became popular for their ability to capture short-term temporal features directly from RGB video volumes [4, 1]. Still, they remain heavy and hard to apply in practice. In the past few years, attention mechanisms have become the standard. Transformer-based architectures, such as SPOTER [6], and Pose-VIT [7], treat pose landmarks as tokens, applying self-attention across both spatial and temporal dimensions. While highly effective, the performance of these models remains insufficient to consider the task of ISLR solved. More recent alternatives [1, 8] employ these architectures as sub-components within more complex frameworks, leveraging contrastive pretraining on pose data and utilizing word embeddings as target representations [8]. In this work, for the sake of simplicity, we focus exclusively on the fundamental classification task. We consider this relevant, as more robust feature extractors would drive improvements when integrated as a subcomponent in such architectures.

2.2 Dataset Characteristics and Modalities

In recent years, the landscape of ISLR datasets has expanded significantly, yet these resources exhibit considerable heterogeneity. A primary distinction lies in the recording environment. The vast majority of existing datasets are captured in controlled studio settings, characterized by uniform lighting, static backgrounds, and strict postural constraints. Furthermore, the articulation of signs in these datasets is often deliberate, with artificial pauses inserted before and after each gloss. In contrast, datasets derived from continuous discourses present

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

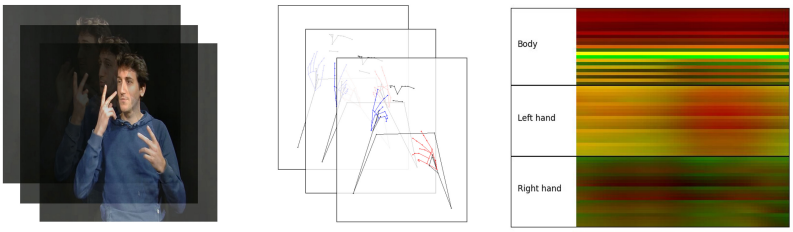


Fig. 1: Frames randomly selected from a video in the LSFB dataset. A skeletal pose is extracted from each frame. Each pose contains 65 landmarks: 21 per hand, and 23 for both the upper body and the head. Then, the pose sequence is transformed into a static image representation, with one row per landmark.

a distinct set of challenges [2]. These signs are performed by native signers spontaneously. Parallel to this environmental shift, there is a growing trend towards utilizing skeletal pose data rather than raw RGB frames [1, 7]. This modality offers substantial computational benefits and improves the signal-to-noise ratio by discarding irrelevant visual information, such as background or clothing color. Aligning with these advancements, our work focuses exclusively on skeletal pose data. To ensure relevance and robustness, we utilize the LSFB dataset [9], one of the largest ISLR datasets extracted directly from real-world continuous discourses.

3 Movement Representation as Images

The concept of encoding time-series motion data into static images is not entirely new in the field of Human Action Recognition [10]. In the specific context of sign languages, however, this approach remains under-explored. Previous work by Hamza et al. [11] attempted to visualize signs as images but simplified the problem by collapsing the temporal dimension, focusing only on spatial configurations. We assume that this loss of temporal granularity is detrimental to SLR, and we adopt a different representation that maintains the integrity of the temporal dimension within the spatial constraints of a 2D image grid. We collect isolated sign language videos extracted from real sign language discourses. They are then reduced to sequences of skeletal poses and finally transformed into a static image, as illustrated in Figure 1.

To leverage standard CNNs, we transform the pose sequence $P \in \mathbb{R}^{T\times V\times 3}$ (T frames, V landmarks) into an image $I \in \mathbb{R}^{V\times T\times 3}$. The 3D coordinates $P_{t,v}$ of the v^{th} landmark of the t^{th} frame of a sign are mapped to RGB values $I_{v,t}$ as

$$I_{v,t,c} = 255\frac{P_{t,v,c} - \min\limits_{t,v} P_{t,v,c}}{\max\limits_{t,v} P_{t,v,c} - \min\limits_{t,v} P_{t,v,c}} \quad \text{for each dimension / channel } c$$

that corresponds to a global min-max normalization following Laraba et al. [10].

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

4 Training CNNs for Sign Recognition

We selected the ResNet family as our backbone due to its ubiquity and proven robustness in image classification, especially for sign languages [12]. We experimented with ResNet-18, ResNet-50, and ResNet-152, pretrained on ImageNet and fine-tuned on the LSFB dataset [9]. The final linear layer is adapted to match the number of classes (500 or 2000) of the target subset. The training objective is a standard cross-entropy loss. We use the Adam optimizer (initial learning rate 10^{-3}) with a scheduling strategy: 20 epochs of linear warm-up followed by a reduce-on-plateau (patience 5, factor 0.1). We compare these architectures against specialized fully supervised SLR models, specifically SPOTER [6] and Pose-VIT [7]. All models were implemented in PyTorch, using the official repositories for baselines, and trained using CUDA 12.6 and a single GPU (NVIDIA RTX 5090).

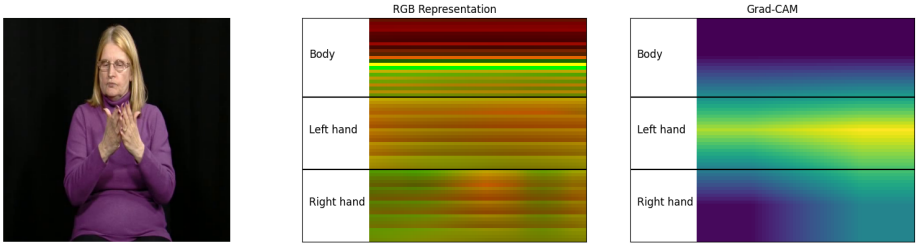
4.1 Quantitative Evaluation

Table 1: Classification balanced accuracy and accuracy on LSFB subsets.

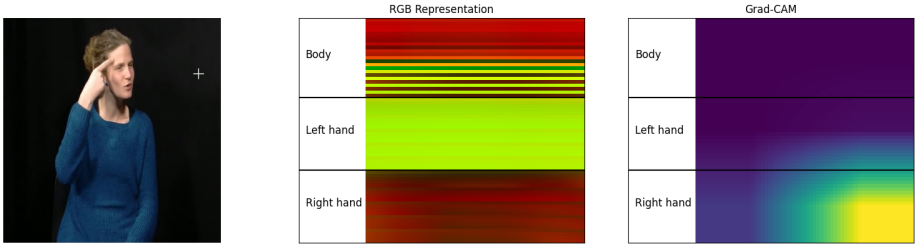
Architecture	**LSFB 500**		**LSFB 2000**	
	Bal. Acc.	**Acc.**	**Bal. Acc.**	**Acc.**
ResNet-18	$38.04 \pm 0.9\%$	$58.61 \pm 0.7\%$	$29.92 \pm 0.2\%$	$49.76 \pm 0.2\%$
ResNet-50	$\mathbf{40.78 \pm 0.8}\%$	$\mathbf{61.49 \pm 0.6}\%$	$32.85 \pm 0.3\%$	$52.93 \pm 0.3\%$
ResNet-152	$40.69 \pm 0.5\%$	$61.34 \pm 0.5\%$	$\mathbf{33.16 \pm 0.3}\%$	$\mathbf{53.27 \pm 0.3}\%$
SPOTER [6]	$32.11 \pm 0.6\%$	$51.98 \pm 0.7\%$	$25.96 \pm 0.5\%$	$44.45 \pm 0.6\%$
Pose-VIT [7]	$33.75 \pm 1.4\%$	$53.92 \pm 1.5\%$	$27.74 \pm 0.2\%$	$46.71 \pm 0.3\%$

The results in Table 1 indicate that standard ResNet architectures perform remarkably well compared to specialized architectures designed specifically for ISLR. Despite not having explicit mechanisms to handle temporal dependencies (such as attention heads for time steps), the movement-as-image representation allows the CNN to capture temporal dynamics as spatial patterns. The balanced accuracy of ResNet-50 is 7% higher than with SPOTER or Pose-VIT on both LSFB500 and LSFB2000 datasets, suggesting that for short isolated signs, explicitly modeling time might be less critical than extracting robust spatio-temporal features. These results are particularly encouraging given the difficulty of the LSFB dataset, where natural discourse and linguistic phenomena create high variance across the numerous classes. In additional experiments, pretraining, normalization and linear interpolation have negligible impact on the results. This eliminates factors other than model selection and it highlights the benefits of the ResNet architecture for ISLR. Those results suggest that ResNets are strong competitors and could be used as sub-components in more complex ISLR architectures. They are also popular models with existing computationally efficient implementations.

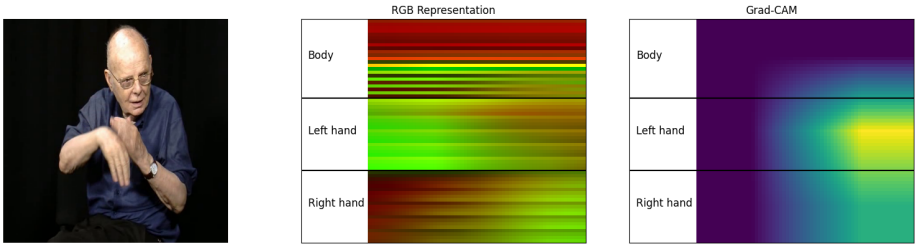
ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.



(a) Label: "LIVRE"



(b) Label: "PENSER"



(c) Label: "DANS"

Fig. 2: Visualization of three samples with different labels from the LSFB dataset. From left to right: an example RGB frame, the static image representation, and the corresponding ResNet50 Grad-CAM saliency map. The heatmaps highlight the regions most relevant for the classification of each sign.

4.2 Qualitative Analysis

When observing samples randomly drawn from different classes in the LSFB dataset, as illustrated in Figure 2, their static image representations differ greatly and provide rich information about the movements themselves. For example, Figure 2c shows two color fades: from light to dark for the left hand, and from dark to light for the right hand. This means that the left hand moves from bottom to top, and the right hand from top to bottom. We can also observe that the movement of the right hand is more pronounced than that of the left hand, in that it is more sweeping. In summary, static image representation is visually informative. This could reasonably explain its effectiveness with CNNs.

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

Saliency maps also provide interesting information. For example, Figure 2b shows that the model focuses on the end of the right hand's movement, which is relevant when examining the corresponding sign. While strictly qualitative, this type of analysis and interpretation are made much simpler by the static representation of movements. Just as a spectrogram provides information about the nature of a sound, an image of a movement provides substantial information about how it was performed, both to a human and to a CNN.

5 Conclusion

This study shows that ResNets are competitive feature extractors in Isolated Sign Language Recognition, considering movements as a static image following a simple transformation. The strong performance of this simple representation suggests that the difficulty in SLR may be partially alleviated by better data representation. It follows that this "movement-image" feature extraction could serve as a powerful backbone for heavier, more complex models. Our qualitative analysis also shows that this representation allows for rapid interpretation of a movement and opens the door to the use of saliency maps to analyze the importance of the features. Future work involves evaluating the impact of a static representation of movements on sign language processing as a whole, for example in continuous recognition or translation.

References

[1] Noha Sarhan and Simone Frintrop. Unraveling a Decade: A Comprehensive Survey on Isolated Sign Language Recognition. In *Proc. ICCV*, pages 3210–3219, 2023.

[2] Jérôme Fink et al. Trends and Challenges for Sign Language Recognition with Machine Learning. In *Proc. ESANN*, pages 561–570, 2023.

[3] Aurore Paligot. Abaissement et registres de langue: Étude de la variation des signes frontaux et des signes symétriques dans un corpus de langue des signes de belgique francophone. *Travaux du Cercle Belge de Linguistique*, 11, 2017.

[4] Dongxu Li et al. Word-level Deep Sign Language Recognition from Video: A New Large-scale Dataset and Methods Comparison. In *Proc. WACV*, pages 1459–1469, 2020.

[5] Shiwei Gan et al. Contrastive Learning for Sign Language Recognition and Translation. In *Proc. IJCAI*, pages 763–772, 2023.

[6] Matyáš Boháček and Marek Hrúz. Sign Pose-Based Transformer for Word-Level Sign Language Recognition. In *Proc. WACV*, pages 182–191, 2022.

[7] Jérôme Fink et al. Sign language-to-text dictionary with lightweight transformer models. In *Proc. IJCAI*, pages 5968–5976, 2023.

[8] Ryan Wong et al. Learnt contrastive concept embeddings for sign recognition. In *Proc. ICCV Workshops*, pages 1945–1954, 2023.

[9] Jérôme Fink et al. LSFB-CONT and LSFB-ISOL: Two New Datasets for Vision-Based Sign Language Recognition. In *Proc. IJCNN*, pages 1–8, 2021.

[10] Sohaib Laraba et al. 3D skeleton-based action recognition by representing motion capture sequences as 2D-RGB images. *CAVW*, 28(3-4):e1782, 2017.

[11] Hafiz Muhammad Hamza and Aamir Wali. Pakistan Sign Language Recognition: From Videos to Images. *Signal, Image and Video Process.*, 19(8):682, 2025.

[12] Ariel Basso Madjoukeng et al. Local-global data augmentation for contrastive learning in static sign language recognition. In *Proc. IDA*, pages 54–66, 2025.