Isolated Sign Language Recognition with Multi-Scale Spatial-Temporal Graph Convolutional Networks

Manuel Vázquez-Enríquez José L. Alba-Castro Laura Docío-Fernández
Eduardo Rodríguez-Banga
atlanTTic Research Center, Universidade de Vigo
36310 Vigo, Spain
mvazquez,jalba,ldocio,erbanga @gts.uvigo.es

Abstract

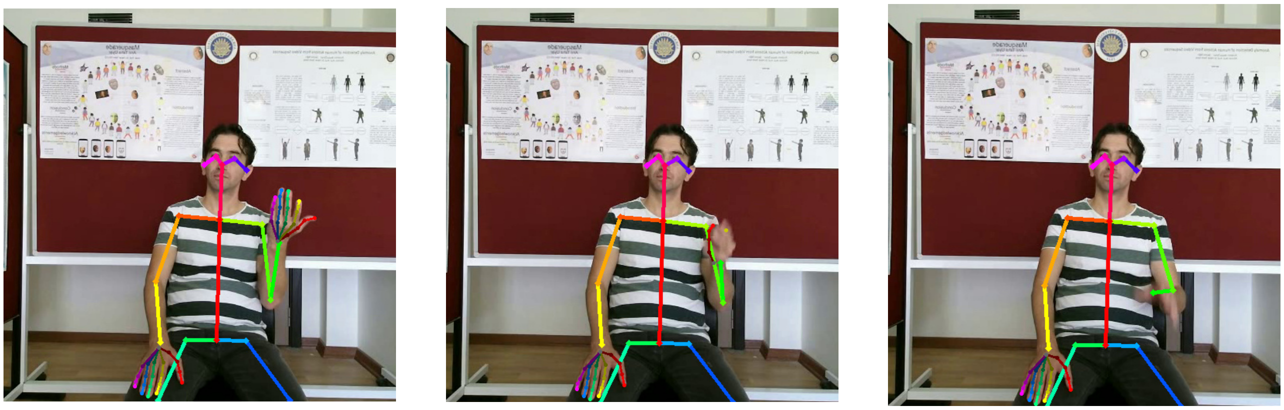
Isolated Sign Language Recognition (ISLR) fits nicely in the domain of problems that can be handled by graph-structured spatial-temporal algorithms. A recent multi-scale spatial-temporal graph convolution operator, MS-G3D, takes advantage of the semantic connectivity among non-neighbor nodes of the graph in a flexible temporal scale, which results in improved performance in classical Human Action Recognition datasets. In this work, we present a solution for ISLR using a skeleton graph that includes body and finger joints and makes use of this specific property of MS-G3D, which seems crucial to capture the internal relationship among semantically connected distant nodes in sign language dynamics. To complete the analysis, we compare the results with a 3D-CNN architecture, S3D, already used for SLR, and fuse it with MS-G3D. The performance achieved on the AUTSL dataset shows that MS-G3D alone stands out as a viable technique for ISLR. In fact, the improvement after fusing with a 3D-CNN approach, at least on this medium-scale dataset, appears marginal. The transfer learning capability of the trained models is also explored using pre-training with the larger WLASL dataset and post-training with the smaller LSE_UVIGO dataset. The classification performance based on the MS-G3D model over AUTSL does not benefit from pre-training with WLASL, but the performance on the more similarly acquired LSE_UVIGO dataset improves significantly from fine-tuning the MS-G3D AUTSL model.

1. Introduction

Sign Languages are the primary means of communication among millions of deaf people around the world. Any of them can be considered as a minority language coexisting with a dominant spoken language. Therefore, deaf people live with a double communication barrier: on the one hand, understanding other deaf people who use any other sign language; on the other hand, dealing with serious difficulties in communicating with their own family, friends, neighbor citizens and service providers. Sign Language Recognition (SLR) and Translation (SLT) have been historically underestimated by the research community for several technical and also, sadly, funding reasons. In contrast to spoken languages, which rely primarily on a single audio signal stream, sign languages are much more challenging because they rely on visual cues in a person-dependent 3D space, where several body parts interact in parallel at different rates and granularity, with multiple manual and non-manual components: arms, hands, fingers, torso, head, eyebrows, eyes, mouth, lips and even tongue [10].

The advent of Deep Learning techniques and the availability of large computational resources have helped to boost SLR and SLT research in the last decade [23]. In addition, many research groups have compiled datasets to capture and learn the visual cues in sign languages [37]. Nevertheless, a wide representation of cross-language factors, for learning SLR-specific hand configurations and short-term body kinematics, is still missing. These cross-language factors are: signing styles (speed and use of the 3D space while signing), the use of signing classifiers and their location, signs that differ just by the speed or repetitions or by a slight hand configuration change or context, the co-articulation effect in continuous signing, the high speed while finger-spelling, and the modifiers added by the facial expression, and head and body movement.

In the last few years, state-of-the-art (SOTA) methods for SLR are mainly based on a spatial-temporal texture feature extraction step performed with 2D or 3D CNNs backbones in a per-frame or group of frames basis [36, 1]. A subsequent step accounts for long-term temporal alignment, performed with Hidden Markov Models [26], Recurrent Neural Networks [5, 25] or, more recently, attention mechanisms and Transformer Networks [7, 39].



(a) (b) (c)

Figure 1: (a) Openpose keypoints estimation on a frame without blurring. (b) Openpose keypoints estimation on a frame with little blurring. (c) Openpose keypoints estimation on a frame with large blurring.

These methods, based on extracting features directly from raw RGB data, are quite data-hungry, since they need to learn to get rid of redundant and non-informative texture from background and clothes; but they have the advantage of not relying on previous image segmentation processes that could propagate errors and, also, they do not need any supplemental devices that could prevent a seamless deployment of the solutions. Obviously, the analysis of just a 2D projection in the camera plane comes with extra challenges like hands and fingers self-occlusion and poor estimation of the distance from the hands to the body, which is an important source of discrimination in sign languages. Although depth information from time-of-flight sensors is a useful complementary information [19], the trend nowadays is to simplify the systems from the hardware point of view and try to estimate depth from deep models trained over paired RGB+D information [35].

Besides, there is a very new research trend for SLR that leverages the great improvements achieved in automatic extraction of whole-body keypoints using, precisely, CNN-based deep learning architectures. Most of the applications that pushed forward the SOTA in seleton and hands keypoints detection are related to Human Action Recognition (HAR). The natural variety of scenarios for these applications have driven the research efforts towards robustness in keypoints detection regardless of background, partial occlusions and self-occlusions. These methods take full advantage of the properties of spatial-temporal Graph Convolutional Networks (GCN) [49, 40]. The straightforward application to SLR did not wait too long [13, 45] since it is easier to count on trained models for keypoints detection in the context of HAR (with many research groups and application scenarios) than in the context of SLR (with many fewer researchers, funding bodies and interested companies).

In any case, SLR can be considered a very special case of Human Action Recognition. The main differences are related to the fine granularity and speed of the actions performed with arms and hands while signing, in contrast to most of the typical actions of interest in HAR, and the role that facial expressions plays in SLR, which is almost negligible in most of HAR contexts. These cross-language factors still pose a great challenge to the reliable detection of keypoints and their long-term temporal dependencies. Moreover, the typical shutter speed of commercial off-the-shelf sensors for video acquisition at 25-30 fps is too slow for many dynamic signs and finger-spelling, so motion blur erase hand and finger movement details that are crucial to telling apart many confusing signs [43, 18, 8]. Figure 1 shows how blurring affects the keypoints estimation accuracy.

In this paper, we analyze the performance of SOTA techniques for SLR using raw RGB input sequences versus whole-body keypoints sequences for Isolated Sign Language Recognition. The analysis is performed on a recent ISLR dataset, AUTSL [42], while two other datasets are used for testing the gain with pre-training (WLASL [27]) and post transfer learning (LSE_UVIGO [15]). The main contributions of this paper are:

- Benchmarking SOTA approaches (RGB versus keypoints) over a common ISLR dataset (AUTSL) and applying to SLR, for the first time, the MS-G3D [32] GCN method.
- Evaluating the performance gain when combining multiple streams.
- Evaluating the advantages of pre-training with a larger dataset from another language and, in addition, the advantages of transferring the learned model to a smaller dataset in a different language, but acquired in similar conditions.

The rest of this paper is structured as follows: in Section 2, a brief review of the SOTA in RGB-based and Skeleton-based deep learning approaches for SLR is presented; Section 3 explains with more detail the two RGB- and Skeleton-based models used for the comparative analysis; Section 4 presents the dataset and the experimental protocol and results; Section 5 shows the transfer learning analysis for pre-training AUTSL and for post-training with AUTSL; finally, Section 6 draws some conclusions and future research lines of this work.

2. Related work

In early approaches to Action and Sign Language Recognition, the spatial-temporal representations were obtained through processing handcrafted features [11, 33, 3], resulting in systems with a very limited generalization capability. The advent of Deep Learning overtook these handcrafted features and quickly became the SOTA in tasks such as action, gesture and sign language recognition.

Recent works have approached these tasks from two different perspectives: raw RGB-based methods and skeleton-based methods. Some methods that combine both methodologies have also been proposed as they complement each other [6].

2.1. RGB-based approaches

Many works have leveraged the flexibility of standard CNNs for learning spatial filters that extract discriminative information for SLR [24, 12, 34, 51, 7, 39], and use recurrent networks like LSTM or BLSTM [24, 12, 34], Attention Layer [51], or Transformer [7, 39] for temporal information encoding. While these networks work quite well for dealing mid and long-term dependencies, some authors showed the advantage of temporally extended CNN for short-term dependencies. Thus, Tran *et al.* [44] proposed the first 3D CNN for action recognition with good results. Since then, many research works have demonstrated the advantages of this type of filters, becoming widely used in the context of HAR and applied to SLR [29, 46, 27]. Nowadays, the most widely used 3D CNN architectures are Inflated 3D CNN (I3D) [9], ResNeXt3D-101 [17] and separable 3D CNN (S3D) [48].

As described in [20], one drawback of the 3D CNN architectures resides on their final Temporal Global Average Pooling (TGAP) stage, which partially hinders long-term temporal dependencies. To overcome this effect, in the previous reference, the authors proposed the substitution of the TGAP block by a bidirectional transformer (BERT) [14], whose attention mechanism would deal with most of the temporal dependencies. The resulting performance depended on the particular 3D CNN architecture and the considered information streams (RGB, flow or both), in addition to the optional use of feature reduction blocks.

The TimeSformer model [4], an extension of the Vision-Transformer [16], seems a promising architecture for video classification in the near future. TimeSformer leaves CNNs out and focuses on spatial-temporal attention mechanisms to capture both short and long-term dependencies.

2.2. Skeleton-based approaches

These methods accept the hypothesis that all the information needed to decode the sign language message is conveyed in the time-sequence of a set of body keypoints. As mentioned in Section 1, the large number of HAR applications have pushed the SOTA in pose estimation methods. Once again, the incorporation of CNNs into the pose estimation framework [47] boosted the accuracy of body keypoints location. Soon, keypoint detection was extended to hand and face [8].

Temporal dynamics for HAR were quickly tackled with CNN or RNN strategies [31, 22, 49], although these models lacked a proper learning of the spatial-temporal interplay among keypoints in the skeleton. Yan *et al.* [49] proposed for the first time a spatial-temporal graph convolutional network (ST-GCN), and demonstrated the effectiveness of learning temporal skeleton dynamics with these networks. Multiple works flourished from this idea, such as AS-GCN[28], 2s-AGCN [40] and MS-G3D [32]. Having demonstrated its potential in HAR, GCN-based architectures have now also gained a strong presence in the SLR field [13, 45].

2.3. Transfer Learning

As mentioned before, sign language recognition and translation techniques still have to pave a long and tortuous road before reaching the maturity of current spoken language technologies. One of the main reasons is the scarcity of training resources to learn such a large variety of cross-language factors, that would make these languages fully understandable to a machine. Fortunately, it is possible to leverage a pre-trained deep learning model and use it in a similar task that lacks enough labeled training samples.

The benefits of transfer learning are highly correlated with the similarity (or distance) between the source and target domains. Therefore, depending on the similarity degree, we can encounter different scenarios [38, 50, 2]: i) when both domains or tasks are the same, significant improvements can be achieved by applying transfer learning on a small dataset of the target task to adjust the weights of the final layers, so that the training time and the generalization capability are improved; ii) in case of quite similar tasks, it is convenient to feed the pre-trained network with a larger amount of task-specific data of the target domain and, thereby, train additional layers; iii) in case of a clear mismatch between domains, the better option is to train the model from scratch provided that enough target-domain data are available.

Of course, the concept of similarity between different domains is very abstract. How to define the similarity between the HAR and the SLR domains? How to measure the similarity between two different sign languages for SLR purposes? In HAR and SLR the use of pre-trained models on ImageNet or Kinectics-400 datasets is quite widespread [7, 13], but not the use of models pre-trained on a specific sign language and, then, fine-tuned to another sign language using some additional data acquired in similar conditions. We discuss this case in detail in Section 5.

3. Methods

In this section, we present the SOTA methods we have selected to apply in ISLR. We used two different models: the S3D architecture as a raw RGB-based model and the MS-G3D architecture as a skeleton-based model.

3.1. RGB-based method: S3D

As mentioned in Section 2.1 3D CNN architectures are applied to a wide extent in the context of HAR. In this work, we have chosen the S3D architecture [48] as it has shown better results than other 3D CNN-based solutions in HAR tasks.

S3D is a separable 3D convolution version of I3D [9], which seeks to reduce the number of parameters to learn without impacting its performance. To this end, the 3D convolutions are factorized into spatial and temporal 3D convolutions. Since many of these 3D convolutions can be implemented as 2D+1D convolutions, the resulting model is more computationally efficient than traditional 3D CNN architectures. In fact, its authors suggest employing 2D+1D convolutions in the lowest layers of the network, where ordinary 3D convolutions would be much more computationally demanding. S3D has shown better accuracy than the original I3D model on different action classification benchmarks. Moreover, the introduction of a feature gating mechanism (S3D-G), which can be interpreted as an attention mechanism on the channel dimension, further improves its accuracy.

3.2. Skeleton-based method: MS-G3D

Among the SOTA skeleton-based methods we have selected the architecture proposed by Liu *et al*. named MS-G3D [32]. At a high level of abstraction, the architecture is made up of stacking blocks of spatial-temporal graph convolutional networks (ST-GCN), followed by a global average pooling layer and a softmax classifier. Focusing on the ST-GCN blocks, its authors propose a unified spatial-temporal graph convolution module called G3D, which unifies the GCN (spatial features) and TCN (temporal features) modules. The G3D modules are integrated within a multi-scale aggregation scheme, reducing redundant dependencies between close and remote neighborhoods within the graph, to obtain a better performance in long-range modeling.

One important characteristic of sign languages is the semantic connection between the configuration of both hands in bi-manual signs, and between one hand and other body parts in non-manual signs. The proposed MS-G3D approach introduces a flexible mechanism to understand the connected variations between nodes of any part of the graph on a predefined spatial and temporal scale by learning different levels of semantic information of the graph.

4. Experiments on AUTSL

In this section, we detail the experiments performed on the AUTSL dataset. First, we introduce the AUTSL dataset and the necessary data preparation. Then, we evaluate the performance using the S3D and MS-G3D models introduced in Section 3 individually. Finally, we describe the strategy followed to combine RGB and skeleton streams and report on its performance.

4.1. AUTSL dataset

AUTSL is a large-scale, multi-modal Turkish isolated sign language dataset [42] recorded using Microsoft Kinect v2, which contains RGB, depth and skeleton modalities. Its samples were recorded under different real-life scenarios with different backgrounds (indoor and outdoor settings) and different lighting conditions. Table 1 summarizes its main characteristics.

Property	**Count**
Number of signs	226
Number of signers	43
Total samples	36,302
Train samples	28,142
Average samples per sign	169.6
Number of different backgrounds	20
Validation samples	4,418
Test samples	3,742

Table 1: Main characteristics of the AUTSL dataset.

4.2. Data preparation

We pre-processed the raw RGB videos differently in order to get 5 streams which were the inputs to the models: RGB, joints, bones, joint-motion and bone-motion streams.

The input stream to the S3D model was the raw RGB video resized to a resolution of 256x256 (RGB stream). On the other hand, the inputs to the MS-G3D model were skeleton-based data. We analyzed 4 different streams obtained from the 67 keypoints (25 body/foot keypoints and 21 hand keypoints for each hand) estimated with Openpose. We discarded the keypoints corresponding to the lower part of the body, as they are not relevant in sign language, and we defined as bones the distance between two adjacent keypoints, as illustrated in Figure 2. Using the remaining 55 keypoints we obtained the following 4 streams:

- The joint stream: original joint position as input.
- The bone stream: distance between adjacent joints as input.
- Joint- and bone-motion streams: differences between adjacent frames.

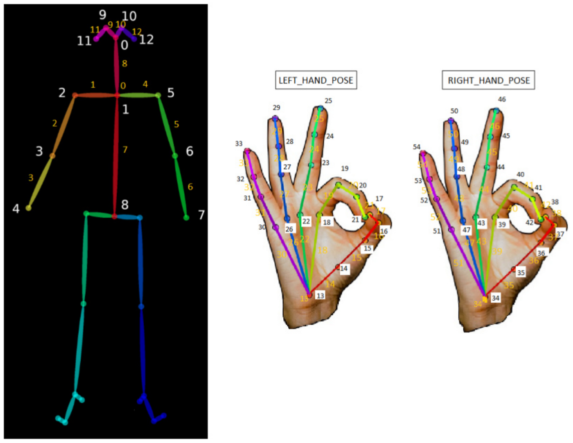


Figure 2: Skeleton and hand joints and bones.

4.3. Single stream evaluation

4.3.1 S3D implementation details

We used a modified version of the S3D implementation available at GitHub[1]. Following [9], we trained and tested the model using clips of 64 consecutive frames from the whole video. For shorter videos, we randomly repeat the first or last frames as many times as necessary. For data augmentation at training time, we randomly cropped a 224x224 patch and applied horizontal flipping. However, at test time, we used a 224x224 center crop.

We used an S3D model pre-trained on Kinetics-400 [21] which was necessarily fine-tuned on the AUTSL dataset to capture the spatial-temporal information (hand shapes and orientations, arm movements, facial expression, etc.) of Turkish signs. For fine-tuning, we employed the Adam optimizer (batch size: 8, weight decay: 10^{-7}, initial learning rate: 10^{-3}) and a learning rate scheduler which reduced the learning rate by a constant factor when the performance metric reached a plateau on the validation set (commonly known as ReduceLRonPlateau). It is worth noting that we employed as total loss the average of the classification and the temporal location losses.

4.3.2 MS-G3D implementation details

We used the implementation of MS-G3D available at GitHub[2]. All skeleton sequences were padded to T = 157 frames. For the disentangled aggregation scheme used for multi-scale learning, we set a number of scales of 8 in the G3D and GCN modules. We kept the architecture proposed by the author stacking up to 3 ST-GCN blocks with 96, 192 and 384 feature channels respectively.

To increase the generalization capability of the model, data-augmentation techniques were employed: random mirroring left-right, adding location and size noise, and randomly removing keypoints according to Openpose confidence score. The models were trained with SGD with Nesterov's accelerated gradient (momentum: 0.9, batch size: 64, weight decay: 0.0003, initial learning rate: 0.1) with a step LR decay factor of 0.1 at epochs 45 and 55. Four single-stream models were trained using these configuration parameters, one per feature stream: joints, bones, joint-motion and bone-motion.

4.3.3 Results

Table 2 shows the performance obtained with every input feature stream, evaluated on the validation set. The performance of a baseline model is also given as a reference for performance comparison [42]. This baseline uses a CNN + FPM + LSTM + Attention model, and its accuracy on validation set is reported into the CVPR 2021 ChaLearn LAP Large Scale Signer Independent Isolated SLR Challenge [41] (RGB TRACK)[3] within the development section.

Model	**Stream**	**Top1(%)**	**Top5(%)**
Baseline	RGB	42.58	-
S3D	RGB	90.27	97.98
MS-G3D	Joints	95.38	99.37
	Bones	94.50	99.07
	Joint-motion	92.92	99.16
	Bone-motion	90.22	98.60

Table 2: Top-1 and top-5 accuracy (%) achieved by each stream on AUTSL validation set.

[1] https://github.com/kylemin/S3D

[2] https://github.com/kenziyuliu/MS-G3D

[3] https://competitions.codalab.org/competitions/27901#results

Streams					Ensemble				
MS-G3D				S3D	Non-weighted		Weighted		
Joints	**Bones**	**J.Motion**	**B.Motion**	**RGB**	**Top1 (%)**	**Top5 (%)**	**Weights**	**Top1 (%)**	**Top5 (%)**
✓	✓				95.70	99.41	(0.64, 0.36)	95.78	99.43
		✓	✓		93.43	99.27	(0.66, 0.34)	93.75	99.32
✓	✓			✓	96.03	99.57	(0.42, 0.32, 0.26)	96.19	99.57
		✓	✓	✓	95.20	99.48	(0.48, 0.26, 0.26)	95.58	99.48
✓	✓	✓			95.74	99.57	(0.48, 0.28, 0.24)	95.95	99.57
✓	✓	✓	✓		95.47	99.56	(0.48, 0.28, 0.24, 0.0)	95.95	99.55
✓	✓	✓	✓	✓	96.15	99.64	(0.28, 0.18, 0.36, 0.0, 0.18)	96.51	99.64

Table 3: Top-1 and top-5 accuracy (%) and weigths achieved by each multi-stream ensemble on AUTSL validation set.

From Table 2, we observe that all the proposed models consistently outperform the baseline, improving the SLR performance by more than 47.64%. Comparing our approaches, we see that, in general, MS-G3D models using skeleton-based stream perform better than the S3D model using the RGB stream. In addition, the best performance of an MS-G3D model is obtained with the joints stream, and using motion streams results in worse performance.

4.4. Multi-stream ensemble strategy

In this section, we analyze and show the benefits of applying a multistream ensemble strategy on the output of the single stream models. Specifically, the outputs of the last linear layer, previous to the softmax layer, are combined following one of these two strategies:

- Unweighted sum of the score vectors.
- Weighted sum of the score vectors that maximize the Top1 accuracy over the validation set.

Table 3 shows the results obtained by the different explored strategies, including, in the case of the weighted ensemble, the optimal combination.

The simplest ensemble strategy that best combines complexity & performance is obtained with the scores from joint and bone streams. This combination strategy is the most common in most of the STGCN-based works.

In order to achieve the highest Top1 accuracy, regardless of the complexity of the final solution, the most accurate (both in the weighted and the unweighted sums) is obtained by combining all the models, with the exception of those trained with the bone-motion stream. In our experiments, these scores provided a negative influence when introduced in any combination.

It is worth noting the benefit for the S3D when adding MS-G3D (from 90,27% to 96,51%), but the accuracy increment is more marginal for MS-G3D when adding S3D (from 95,95% to 96,51%). This small performance gain opens the discussion on the opportunity to increase the complexity with a very demanding model in terms of the amount of training data and the computational burden in inference.

5. Effect of transfer learning

In this section, we show the results of analyzing the potential benefits of transfer learning in the specific task of ISLR. It is important to highlight that transfer learning was already built-in in the implementation of Section 4: S3D method used a pre-trained model on Kinetics-400 and MS-G3D used whole-body skeletons extracted with OpenPose, which was trained on the COCO dataset [30].

We now focus on the analysis of transfer learning from and to similar sign language recognition tasks. In the next subsections, we present the WLASL dataset, which can be used for pre-training the model of Section 4 (transfer from A to B) and another dataset, LSE_UVIGO, smaller than AUTSL, which can be used to test the influence of the similarity between domains when transferring knowledge from A or B to C. These datasets were collected for ISLR but in three different languages: A: ASL, B: TSL, C: LSE (Lengua de Signos Española - *Spanish Sign Language*)

5.1. WLASL dataset

Word-Level American Sign Language dataset (WLASL) [27] is a large-scale ASL dataset. The videos were directly extracted from public Internet resources: educational sign language websites and ASL tutorial videos on YouTube.

This database is publicly available[4] and distributed in 4 different subsets according to the number of included glosses it contains: WLASL100, WLASL300, WLASL1000 and WLASL2000.

In this block of experiments we used the WLASL2000 dataset with the characteristics shown in Table 4.

[4] https://github.com/dxli94/WLASL

Property	**Count**
Number of signs	2000
Number of signers	119
Total samples	21097
Train samples	14297
Validation samples	3920
Test samples	2880

Table 4: Characteristics of the WLASL2000 dataset.

5.2. LSE_UVIGO dataset

LSE_UVIGO[5] [15] is a multi-source Spanish Sign Language database collected in several scenarios for ISLR and CSLR purposes. Recordings were simultaneously gathered with a high-speed Nikon D3400 and a Kinect v2. Deaf people, SL interpreters and SL students participated in the recordings under lab controlled conditions. For the experiments on this work, only the Kinect v2 part, and the subset of ISLR were used, since they correspond to the most similar domain to AUTSL. The main difference between them, apart from the language, is that AUTSL portraits people at different distances, with varying backgrounds, indoor and outdoor, while the LSE_Lex40 (40 Isolated Signs) was acquired at the same distance and quite a uniform background. These differences are minimized when using only the skeleton-based approach.

Table 5 shows the content of the ISLR Kinect v2 part of LSE_UVIGO.

Subset	Glosses	Signers	Videos
LSE_Lex40 (train)	40	27	1128
LSE_Lex40 (test)	40	5	200

Table 5: Characteristics of the LSE_Lex40 dataset.

5.3. Transfer learning on AUTSL

Table 6 shows the Top 1 accuracy results on the AUTSL validation set when trained under 3 different settings: first, training the MS-G3D model with the training set of AUTSL from scratch; second, pre-training the MS-G3D model with WLASL2000 and fine-tuning with the AUTSL training set; third, the previous one but pre-trained with a smaller dataset, LSE_Lex40 (train+test). From these three tests it is clear that MS-G3D trained on AUTSL does not benefit from pre-training with a larger similar dataset and, of course, does not benefit from pre-training with a smaller dataset with almost equal acquisition conditions. The takeaway is that MS-G3D captures quite well the variety of sign dynamics from the AUTSL training set. The learning curves (not reproduced here) showed that the only gain when pretraining with WLASL is a faster convergence.

Experiment	**Stream**	**Top1%**
AUTSL trained from scratch	Joints Bones	95.33 94.45
AUTSL w/pre-trained weights on WLASL2000	Joints Bones	95.24 94.20
AUTSL w/pre-trained weights on LSE-Lex40	Joints Bones	95.21 94.23

Table 6: Top-1 accuracy (%) achieved by MS-G3D on AUTSL dataset trained from scratch or using pre-trained weights.

Experiment	**Stream**	**Top1%**
LSE-Lex40 trained from scrach	Joints Bones	85.91 90.95
LSE-Lex40 w/pre-trained weights on WLASL2000	Joints Bones	93.91 95.47
LSE-Lex40 w/pre-trained weights on AUTSL	Joints Bones	97.98 98.49

Table 7: Top-1 accuracy (%) achieved by MS-G3D on LSE_Lex40 dataset trained from scratch or using pre-trained weights.

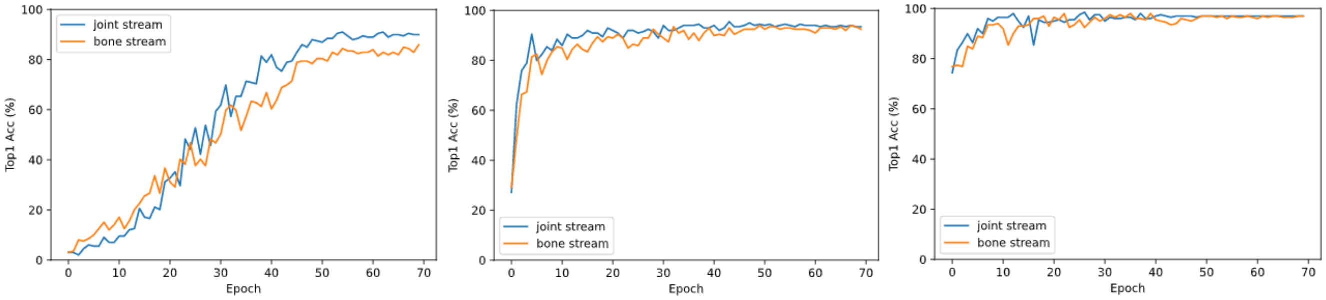
5.4. Transfer learning on LSE_Lex40.

Table 7 shows the Top 1 accuracy results on LSE_Lex40 test set when trained under three different settings: first, training the MS-G3D model with the training set of LSE_Lex40 from scratch; second, pre-training the MS-G3D model with WLASL2000 and fine-tuning with the LSE_Lex40 training set; third, the previos model but pre-trained with a very similar larger dataset, AUTSL (train+val). The takeaway is that a small dataset can benefit from pre-training MS-G3D on larger datasets and that it is more important the similarity of the domains (Kinect2 and similar acquisition settings) than the number of signs. As a side note, the 40 signs of LSE_Lex40 are different from the signs of AUTSL, meaning that MS-G3D pre-trained on AUTSL really captured fundamental dynamics, useful for other signs and configurations. Figure 3 shows the accuracy curve on the test set of LSE_Lex40 for the 3 pre-training scenarios, where is clear the benefit obtained with MS-G3D pre-training in similar domains but larger datasets.

6. Conclusions and future research lines

In this work, we have introduced, for the first time, the use of the skeleton-based MS-G3D architecture for ISLR.

[5] http://gtm.uvigo.es/content/descripcion-lselex40uvigo



(a)

(b)

(c)

Figure 3: Accuracy curves for LSE_Lex40 test set. (a) Trained from scratch. (b) Pre-trained weights on WLASL2000. (c) Pre-trained weights on AUTSL.

The intuition behind this selection is that it allows keeping more reliable semantic connection between hands and body parts in sign language dynamics. To test this idea, we have presented a comparative analysis with another SOTA method based on raw RGB input: the S3D. Both strategies, independent training and fusion over a common dataset of ISLR (AUTSL), show that MS-G3D outperforms S3D with resulting accuracy comparable to the best obtained on the AUTSL validation set of the 2021 ChaLearn LAP LSSII SLR Challenge (RGB TRACK)[6].

Finally, we have analyzed the benefits of transfer learning when training MS-G3D models and conclude that the ISLR task for midsize vocabulary does not benefit from a pre-trained model on a much larger dataset in a different language and slightly different acquisition settings. However, when trained on a small dataset, pre-training MS-G3D with a larger vocabulary improves the performance greatly, with a larger increment on similar domains, as in the case of AUTSL and LSE_UVIGO.

This work will be extended with a wider analysis on additional datasets to verify whether the conclusions regarding the opportunity of fusion with complex RGB-based architectures hold. We will also expect to improve MS-G3D results using keypoint detectors, which are more robust to motion blur, and compare the results to a model trained with the part of the LSE_UVIGO dataset acquired without motion blur.

7. Acknowledgements

This research has been funded by the Spanish Ministry of Science, Innovation and Universities, through the project RTI2018-101372-B-I00 Audiovisual analysis of verbal and nonverbal communication channels (Speech & Signs) and by the Xunta de Galicia and the European Regional Development Fund through the Consolidated Strategic Group AtlanTTic (2019-2022). Manuel Vázquez is also funded by the Spanish Ministry of Science and Innovation through the predoc grant PRE2019-088146.

References

[1] M. Al-Hammadi, G. Muhammad, W. Abdul, M. Alsulaiman, M. A. Bencherif, and M. A. Mekhtiche. Hand Gesture Recognition for Sign Language Using 3DCNN. *IEEE Access*, 8:79491–79509, 2020.

[2] Hossein Azizpour, A. Razavian, J. Sullivan, A. Maki, and S. Carlsson. Factors of transferability for a generic convnet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38:1790–1802, 2016.

[3] Purva C. Badhe and Vaishali Kulkarni. Indian sign language translator using gesture recognition algorithm. In *2015 IEEE International Conference on Computer Graphics, Vision and Information Security, CGVIS 2015*, pages 195–200. Institute of Electrical and Electronics Engineers Inc., apr 2016.

[4] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is Space-Time Attention All You Need for Video Understanding? *ArXiv*, abs/2102.05095, 2021.

[5] N. C. Camgoz, S. Hadfield, O. Koller, and R. Bowden. SubUNets: End-to-End Hand Shape and Continuous Sign Language Recognition. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 3075–3084, 2017.

[6] Necati Cihan Camgoz, Oscar Koller, Simon Hadfield, and Richard Bowden. Multi-channel Transformers for Multi-articulatory Sign Language Translation. *arXiv*, abs/2009.00299, 2020.

[7] Necati Cihan Camgöz, Oscar Koller, Simon Hadfield, and Richard Bowden. Sign Language Transformers: Joint End-to-end Sign Language Recognition and Translation. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 10023–10033. IEEE Computer Society, 2020.

[8] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[6] https://competitions.codalab.org/competitions/27901#results

[9] J. Carreira and A. Zisserman. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4724–4733, 2017.

[10] Helen Cooper, Brian Holt, and Richard Bowden. *Sign Language Recognition*, pages 539–562. Springer London, London, 2011.

[11] Helen Cooper, Eng-Jon Ong, Nicolas Pugeault, and Richard Bowden. Sign Language Recognition using Sub-Units. *Journal of Machine Learning Research*, 13(Jul):2205–2231, 2012.

[12] Runpeng Cui, Hu Liu, and Changshui Zhang. A Deep Neural Framework for Continuous Sign Language Recognition by Iterative Training. *IEEE Transactions on Multimedia*, 21(7):1880–1891, jul 2019.

[13] Cleison Correia de Amorim, David Macêdo, and Cleber Zanchettin. Spatial-Temporal Graph Convolutional Networks for Sign Language Recognition. *arXiv*, abs/1901.11164, 2019.

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[15] Laura Docío-Fernández, José Luis Alba-Castro, Soledad Torres-Guijarro, Eduardo Rodríguez-Banga, Manuel Rey-Area, Ania Pérez-Pérez, Sonia Rico-Alonso, and Carmen García-Mateo. LSE_UVIGO: A multi-source database for Spanish Sign Language recognition. In *Proceedings of the LREC2020 9th Workshop on the Representation and Processing of Sign Languages: Sign Language Resources in the Service of the Language Community, Technological Challenges and Application Perspectives*, pages 45–52, 2020.

[16] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*, 2021.

[17] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can Spatiotemporal 3D CNNs Retrace the History of 2D CNNs and ImageNet? In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 6546–6555. IEEE Computer Society, dec 2018.

[18] Sheng Jin, Lumin Xu, Jin Xu, Can Wang, Wentao Liu, Chen Qian, Wanli Ouyang, and Ping Luo. Whole-Body Human Pose Estimation in the Wild. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[19] Longlong Jing, Elahe Vahdani, Matt Huenerfauth, and Yingli Tian. Recognizing American Sign Language Manual Signs from RGB-D Videos. *ArXiv*, abs/1906.02851, 2019.

[20] M. Esat Kalfaoglu, Sinan Kalkan, and A. Aydin Alatan. Late Temporal Modeling in 3D CNN Architectures with BERT for Action Recognition. In Adrien Bartoli and Andrea Fusiello, editors, *Computer Vision - ECCV 2020 Workshops - Glasgow, UK, August 23-28, 2020, Proceedings, Part V*, volume 12539 of *Lecture Notes in Computer Science*, pages 731–747. Springer, 2020.

[21] Will Kay, João Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The Kinetics Human Action Video Dataset. *arXiv*, abs/1705.06950, 2017.

[22] Tae Soo Kim and Austin Reiter. Interpretable 3D Human Action Analysis with Temporal Convolutional Networks. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, volume 2017-July, pages 1623–1631. IEEE Computer Society, aug 2017.

[23] Oscar Koller. Quantitative survey of the state of the art in sign language recognition. *CoRR*, abs/2008.09918, 2020.

[24] O. Koller, N. C. Camgoz, H. Ney, and R. Bowden. Weakly Supervised Learning with Multi-Stream CNN-LSTM-HMMs to Discover Sequential Parallelism in Sign Language Videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(9):2306–2320, 2020.

[25] O. Koller, S. Zargaran, and H. Ney. Re-Sign: Re-Aligned End-to-End Sequence Modelling with Deep Recurrent CNN-HMMs. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3416–3424, 2017.

[26] Oscar Koller, Sepehr Zargaran, Hermann Ney, and Richard Bowden. Deep Sign: Enabling Robust Statistical Continuous Sign Language Recognition via Hybrid CNN-HMMs. *Int. J. Comput. Vis.*, 126(12):1311–1325, 2018.

[27] Dongxu Li, Cristian Rodriguez Opazo, Xin Yu, and Hongdong Li. Word-level Deep Sign Language Recognition from Video: A New Large-scale Dataset and Methods Comparison. In *Proceedings - 2020 IEEE Winter Conference on Applications of Computer Vision, WACV 2020*, pages 1448–1458. Institute of Electrical and Electronics Engineers Inc., mar 2020.

[28] Maosen Li, Siheng Chen, Xu Chen, Ya Zhang, Yanfeng Wang, and Qi Tian. Actional-Structural Graph Convolutional Networks for Skeleton-Based Action Recognition. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2019-June, pages 3590–3598. IEEE Computer Society, jun 2019.

[29] Zhi-jie Liang, Sheng-bin Liao, and Bing-zhang Hu. 3D Convolutional Neural Networks for Dynamic Sign Language Recognition. *The Computer Journal*, 61(11):1724–1736, nov 2018.

[30] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 740–755, Cham, 2014. Springer International Publishing.

[31] Hong Liu, Juanhui Tu, and Mengyuan Liu. Two-Stream 3D Convolutional Neural Network for Skeleton-Based Action Recognition. *arXiv*, abs/1705.08106, 2017.

[32] Ziyu Liu, Hongwen Zhang, Zhenghao Chen, Zhiyong Wang, and Wanli Ouyang. Disentangling and Unifying Graph Con-

volutions for Skeleton-Based Action Recognition. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 140–149. IEEE Computer Society, 2020.

[33] Bruce Xiaohan Nie, Caiming Xiong, and Song Chun Zhu. Joint Action Recognition and Pose Estimation from Video. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 07-12-June, pages 1293–1301. IEEE Computer Society, oct 2015.

[34] Ilias Papastratis, Kosmas Dimitropoulos, Dimitrios Konstantinidis, and Petros Daras. Continuous Sign Language Recognition through Cross-Modal Alignment of Video and Text Embeddings in a Joint-Latent Space. *IEEE Access*, 8:91170–91180, 2020.

[35] Maria Parelli, Katerina Papadimitriou, Gerasimos Potamianos, Georgios Pavlakos, and Petros Maragos. Exploiting 3d hand pose estimation in deep learning-based sign language recognition from rgb videos. In Adrien Bartoli and Andrea Fusiello, editors, *Computer Vision – ECCV 2020 Workshops*, pages 249–263, Cham, 2020. Springer International Publishing.

[36] L. Pigou, M. Van Herreweghe, and J. Dambre. Gesture and Sign Language Recognition with Temporal Residual Networks. In *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, pages 3086–3093, 2017.

[37] Razieh Rastgoo, Kourosh Kiani, and Sergio Escalera. Sign Language Recognition: A Deep Survey. *Expert Systems with Applications*, 164:113794, 2021.

[38] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson. Cnn features off-the-shelf: An astounding baseline for recognition. In *2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 512–519, 2014.

[39] Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. Continuous 3D Multi-Channel Sign Language Production via Progressive Transformers and Mixture Density Networks. *arXiv*, abs/2103.06982, 2021.

[40] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2019-June, pages 12018–12027. IEEE Computer Society, jun 2019.

[41] Ozge Mercanoglu Sincan, Julio C. S. Jacques Junior, Sergio Escalera, and Hacer Yalim Keles. Chalearn LAP large scale signer independent isolated sign language recognition challenge: Design, results and future research. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2021.

[42] O. M. Sincan and H. Y. Keles. AUTSL: A Large Scale Multi-Modal Turkish Sign Language Dataset and Baseline Methods. *IEEE Access*, 8:181340–181355, 2020.

[43] K. Sun, B. Xiao, D. Liu, and J. Wang. Deep High-Resolution Representation Learning for Human Pose Estimation. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5686–5696, 2019.

[44] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning Spatiotemporal Features with 3D Convolutional Networks. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 4489–4497, 2015.

[45] Anirudh Tunga, Sai Vidyaranya Nuthalapati, and Juan Wachs. Pose-Based Sign Language Recognition Using GCN and BERT. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops*, pages 31–40, January 2021.

[46] Hamid Vaezi Joze and Oscar Koller. MS-ASL: A Large-Scale Data Set and Benchmark for Understanding American Sign Language. In *The British Machine Vision Conference (BMVC)*, September 2019.

[47] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. In *CVPR*, 2016.

[48] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking Spatiotemporal Feature Learning: Speed-Accuracy Trade-offs in Video Classification. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision – ECCV 2018*, pages 318–335. Springer International Publishing, 2018.

[49] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition. *32nd AAAI Conference on Artificial Intelligence, AAAI 2018*, pages 7444–7452, jan 2018.

[50] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, page 3320–3328, Cambridge, MA, USA, 2014. MIT Press.

[51] H. Zhou, W. Zhou, and H. Li. Dynamic Pseudo Label Decoding for Continuous Sign Language Recognition. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1282–1287, 2019.