LSFB-CONT and LSFB-ISOL

Fink, Jerome; Frénay, Benoît; Meurant, Laurence; Cleve, Anthony

Published in:
Proceedings of the 2021 International Joint Conference on Neural Networks (IJCNN 2021)

Publication date:
2021

Document Version
Early version, also known as pre-print

LSFB-CONT and LSFB-ISOL: Two New Datasets for Vision-Based Sign Language Recognition

Jérôme Fink
Fac. of Computer Science
NaDI & NaLTT institutes
University of Namur
Namur, Belgium
jerome.fink@unamur.be

Benoît Frénay
Fac. of Computer Science
NaDI institute
University of Namur
Namur, Belgium
benoit.frenay@unamur.be

Laurence Meurant
Fac. of Philosophy & Letters
NaLTT institute
University of Namur
Namur, Belgium
laurence.meurant@unamur.be

Anthony Cleve
Fac. of Computer Science
NaDI institute
University of Namur
Namur, Belgium
anthony.cleve@unamur.be

***Abstract*—While significant progress have been made in the field of Natural Language Processing (NLP), leading the commercially available products, Sign Language Recognition (SLR) is still in its infancy. The lack of large-scale sign language datasets makes it hard to leverage new Deep Learning methods. In this paper, we introduce LSFB-CONT, a large scale dataset suited for continuous SLR along with LSFB-ISOL, a subset of LSFB-CONT for isolated SLR. Baseline SLR experiments are conducted on LSFB-ISOL and the reached accuracy measures are compared with those obtained on previous datasets. The results suggest that state-of-the-art models for action recognition still lack sufficient internal representation power to capture the high level of variations of a sign language.**

***Index Terms*—Deep Learning, Dataset, Sign Language Recognition**

I. Introduction

In the last decade, significant progress has occurred in speech recognition, leading to the creation of commercially available products relying solely on speech to interact with their users. As a matter of fact, automatic Sign Language Recognition (SLR) has not followed the same trend. However, building robust methods for SLR would not only benefit the deaf community, but would also allow developing more natural ways to interact with software systems through gestures.

SLR is one of the most studied gesture recognition [1] problems and is considered as a challenging task, since sign languages (SLs) do not only rely on hands and arms configurations. Eye gaze, facial expressions and upper body movements also convey additional information to the interlocutor.

Early successes in gesture recognition were obtained in the 90's. They mostly relied on Hidden Markov Models (HMMs) to achieve sign classification [2] [3]. More recently, deep learning architectures, initially applied to image recognition, have been leveraged to dramatically improve the performance of both isolated and continuous SLR.

The first SLR datasets typically depicted a single signer in a strictly controlled video recording environment. These strict constrains are being relaxed, first by including more signers and, more recently, by introducing variations on the background and lightening conditions.

This research is supported by the *Fonds InBev-Baillet Latour*.

Some datasets try to provide additional information captured thanks to specialized devices, such as wristbands to track hand positions, or 3D cameras to capture depth information. As such advanced devices are not widely available, their use restricts the scope of applicability of SLR systems. Furthermore, videos gathered *in the wild* (e.g., online dictionaries and video shared on social media) will generally not provide such additional data. This might explain why most works in gesture or sign recognition prefer to focus on recognition based on raw videos without additional data sources or apparatus [1].

Despite the need of datasets and the encouraging progress already achieved in SLR, the number of *large* SL datasets still remains limited. Training a deep learning model typically requires hours of annotated SL video recordings. The training data are hard to acquire due to the time-consuming and costly annotation process, which consists in manually labeling each single sign occurrence in the recorded videos. Furthermore, in contrast with image recognition, only a few people are qualified to perform SL annotation tasks.

This paper introduces two new public datasets for continuous and isolated SLR, consisting of French Belgian Sign Language (LSFB) conversation videos. The new datasets provide conversations between both native and non-native signers, filmed with an RGB camera at 50 FPS. Signers were asked to speak freely without preset scripts nor vocabulary limitation. This resulted in a dataset depicting natural conversations, close to what could be expected in real life. Also, this SL dataset constitutes the largest available dataset for continuous SLR, both in terms of length and vocabulary size. The main contributions of this paper are:

- two new publicly available SL datasets: LSFB-CONT, a dataset for continuous SLR; and LSFB-ISOL, a dataset for isolated SLR extracted from LSFB-CONT;
- a benchmarking of LSFB-ISOL on three state-of-the art action recognition models;
- a comparison of the SLR results obtained on LSFB-ISOL with those obtained on two other SL datasets.

The remainder of this paper is structured as follows. Section II identifies the main SLR challenges and presents popular SL datasets. Section III introduces our two new SL datasets.

Section IV presents the experimental setup and our results. A qualitative analysis of them is presented in Section V before proposing future research perspectives in Section VII.

II. Related Work

This section presents the challenges of Sign Language Recognition (SLR), why this task is considered difficult and how SLR datasets should be designed to support the development of more robust algorithms. It also presents popular datasets for SLR.

A. Challenges of SLR

Sign Languages (SLs) are natural languages that emerged and are used within deaf communities around the world. Each sign, be it realized by one or two hands, can be described by a set of four manual components: handshape (or hand configuration), location, orientation and movement. Non-manual components such as facial expression, gaze direction, head and body position complete the manual parameters and convey grammatical and semantic information. In contrast with spoken languages, SLs allow one to transmit multiple information simultaneously, using several articulators in parallel [4].

In their survey on hand gesture recognition, Al-Shamayleh et al. [1] propose a classification of the issues that an SLR system should overcome to be useful in practice. Table I summarizes their taxonomy. The *system challenges* category contains all functional requirements that could be expected by the user, such as a short response time and a low cost of the system. Those challenges are not discussed in this paper as our goal is not the creation of a production-ready system for SLR. The two other categories are the *environment* and the *gesture challenges*, corresponding to problems raised by changes in the recording environment and problems related to the structure of a gesture and its variations, respectively.

Categories	**Challenge**
System Challenges	Response Time
	Cost Factor
Environment Challenges	Background
	Illumination
	Invariance
	Ethnic Groups
Gesture Challenges	Translation
	Scaling
	Rotation
	Segmentation
	Feature Selection
	Dynamic Gesture
	Size of Dataset

TABLE I: Taxonomy of challenges for vision-based SLR systems introduced by Al-Shamayleh et al. [1].

The environment and gesture challenges must be kept in mind when choosing or building a dataset to train and test an SLR system. Indeed, the quality of the dataset may have a significant impact on the system's ability to tackle those issues. A good dataset should contain videos involving a lot of signers, and include background and lightning variability to stress the robustness of a system. This need for variability is impossible to fulfill by small datasets. Therefore, a good SLR dataset should contain a (very) large number of videos.

Gesture challenges motivate the need of datasets depicting a large vocabulary of dynamic gestures. A dynamic gesture is a gesture whose semantics do not only rely on the hand and arm configurations, but also in the movements performed by the signers. In the context of sign languages, most common signs are dynamic gestures. The gesture challenges category also covers the robustness of a system in recognizing signs with different scales and rotations. However, this category does not highlight several other critical aspects of the SL gestures. One of these aspects is the *transition* between signs in a sentence. At the end of a sign, there is a moment where the signer repositions her hands to get ready for the next sign. This transition period is called a *movement epenthesis*. Those movements have no meaning but could be mistakenly interpreted as a gesture. The automatic detection of movement epenthesis constitutes a primordial step for sign language speech segmentation [5]. Despite those hand repositioning movements, signs are not executed from a clean deterministic starting position to a clean deterministic ending position. The final position of a sign has a huge influence on the starting position of the following sign. Furthermore, signers naturally vary the ending position of the current sign to anticipate the execution of the next sign. The influence of signs on the execution of their surrounding signs is called *co-articulation*.

In order to build SLR systems robust to movement epenthesis and co-articulation, it is helpful to rely on a dataset including a lot of variability in the way signs are chained. Datasets depicting several signers, but who perform the very same sentences may lack such kind of variability.

The majority of state-of-the-art SLR solutions primarily focus on the identification of *lexical signs* (LS). In modern SL corpus, each LS is associated to a unique fixed label called a *gloss* [6]. However, each manual component of a lexical sign can still vary and cause morphological changes (e.g., number, agreement and aspect variations). For instance, signers would rather execute the sign *walk* quickly instead of using sequentially the signs *walk* and *fast* [7].

In addition, signers often use less standardized signs called *depicting signs* (DS) to describe a situation instead of using standard lexical signs. These signs cannot be associated to a fixed label. They are highly dependent on the context of the sentences and on the cultural background of the signers. In French sign language conversations, about 75% of the signs are LS and 25% are DS [8]. An SLR system able to only recognize and translate lexical signs would never be able to fully address the sign language translation problem. However, isolated sign datasets (see Section II-B2) are typically restricted to lexical signs.

Continuous SL datasets (see Section II-B1) are more likely to depict DS, yet the way to annotate them is still discussed as they cannot be associated to a single word but rather represent a situation or a concept. Different annotation methods were proposed. The low level ones advice to annotate each hand configuration composing a sign and each facial expression [9].

The highest level ones consist of a timed translation of the signed utterance into a written language. However, the most commonly used approach in modern corpus is to use glosses for each LS and to discard or to mark with a single label all the signs that do not fall into the LS category [6].

B. Sign Language Datasets

The need for better SLR performance motivated the development of large sign language datasets. In this section, we consider datasets for both continuous and isolated sign language recognition. We restrict ourselves to RGB videos, without any equipment such as gloves or tracking points. Table II provides a comparison of the state-of-the-art datasets and of the two new datasets we introduce in Section III.

1) Continuous Signs Datasets : The **Phoenix Weather** dataset [10] is an early example of large-scale continuous sign language datasets for the German sign language. It is made of videos of signers translating the German weather forecast. The vocabulary contains 1080 glosses and 9 different signers are recorded. It constitutes one of the largest datasets for continuous SLR to date. One of its drawbacks is that SL speech in a translation context tends to not be representative of how signers would have structured their sentences in natural conversations. Signers use less DS than in a natural speech.

The **Greek Sign Language** (GSL) dataset [11] is a continuous sign language dataset containing 7 different signers. They perform 5 scripted scenarios representing classical interactions with public service employees (police officers, train station agents, etc.). Depth information is provided and the videos are recorded at 30FPS. As all the signers are performing the same speech, the dataset lacks variability in terms of sign co-articulation. In addition, the signers tend to perform the signs more slowly, as they follow a script. However, the dedicated vocabulary depicted in the dataset is interesting in order to develop a system assisting impaired people to communicate with public services. An isolated version of this dataset depicting isolated signs, extracted from the continuous dataset version, is also available.

2) Isolated Signs Datasets: The **American Sign Language Lexicon Video Dataset** (ASLLVD) is one of the first large datasets for isolated sign language recognition. It was initiated in 2008 [12] and then deeply refactored in 2012. It depicts 6 signers performing isolated signs starting from a neutral position. Each sign is performed only once by each signer. The videos are captured by 4 different cameras positioned around the signer. The main weakness of the ASLLVD dataset is that it provides few occurrences of each gloss.

The **DEVISIGN-L** dataset created in 2016 [13] contains isolated signs of the chinese sign language. The videos were captured by means of a Kinect in order to provide depth information for each video. The dataset involves 8 different signers and contains a vocabulary of 2000 glosses with 12 examples for each gloss. It suffers from the same general weakness as the ASLLVD.

The **Microsft American Sign Language** (MS-ASL) dataset was created in 2019 [14]. It has been built by scrapping ASL educational videos on YouTube. Hence, it depict signers performing isolated signs in various environments. The dataset contains a vocabulary of 1000 glosses and depicts 222 signers. The authors also propose four different split : MS-ASL-100, MS-ASL-200, MS-ASL-500 and MS-ASL-1000 class showing respectively 100, 200, 500 and 1000 different classes. This is the dataset containing videos with the most variety of different background, signers and lightning, making it an interesting candidate for training robust isolated SLR systems. Unfortunately, the authors only provide the YouTube url and the timestamps of each sign. As some videos have meanwhile been removed from YouTube or set as private, the MS-ASL dataset is slowly decreasing in size and quality.

III. Proposed Datasets

This section introduces the LSFB Corpus which provides the essential materials to build our two datasets. We present the content of each of our datasets, and we compare them with the pre-existing SLR datasets listed in Section II.

A. LSFB Corpus

Since 2012, researchers at the University of Namur [15] have been collecting and annotating LSFB (French Belgian Sign Language) conversations, with the aim to better understand this sign language. A high diversity of people is needed to document the language grammar and usage, taking into account the variations across various age ranges and regions of the country. Therefore, the videos gathered depict signers of diverse ages and genders from different locations in Belgium. This corpus also includes a mix of native and non-native speakers, that were asked to perform 19 tasks, encouraging them to engage into various types of discourse genres (e.g., narration, argumentation and explanations). All videos were recorded in a studio with a controlled setup and environment. Therefore, a limitation of the LSFB Corpus is that it lacks background and lightning variation. However, all large SL corpora currently under construction are being recorded in similar conditions. Thus, SLR models performing well on the LSFB Corpus could support the annotation of other corpora.

In total, 100 signers participated to the recording sessions (see Figure 1 for statistics). The corpus involves a lot of people with different styles and clothes. Meta-information about the signers are available such as their age, gender, handedness, and linguistic profile. This information could help to diagnose SLR systems and spot their weaknesses.

More than 90 hours of videos have been recorded using two 50-FPS cameras, placed in front of each signer (one camera per signer). Each camera has a resolution of 720x576 pixels. The annotation and translation of those videos are ongoing processes, which already required thousands of hours of work. At the time of writing, 25 hours of videos are fully annotated, among which 5 hours are also translated into French. Each annotation file contains one channel for each hand, and a unique gloss is assigned for each occurrence of lexical sign (LS). DS signs are annotated with a special label. For translated videos, an annotation channel has been used to

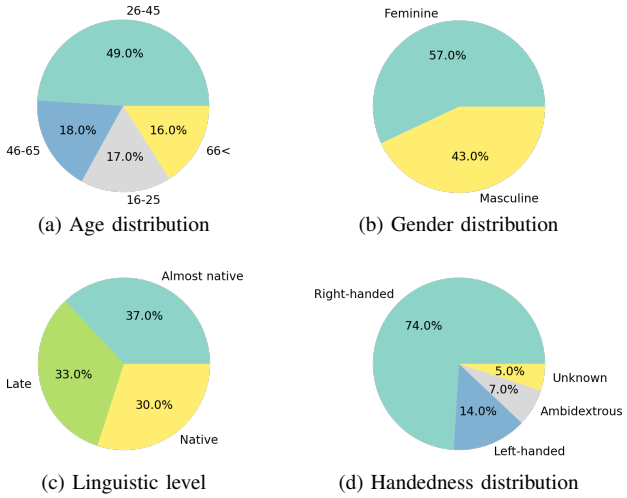
Continuous Sign Language Datasets

	Class	Signers	Videos	FPS	Background and lights	Camera Position
Phoenix Weather	1080	9	6841	30	Controlled	Controlled
GSL	310	7	40826	30	Controlled	Small variations
LSFB-CONT	**6883**	**100**	**85132**	50	Controlled	Controlled

Isolated Sign Language Datasets

	Class	Signers	Videos	FPS	Background and lights	Camera Position
ASLLVD	**3300**	6	9800	30	Controlled	Controlled
DEVISIGN-L	2000	8	24000	30	Controlled	Controlled
MS-ASL	1000	**222**	25513	Varying	Varying	Large variations
LSFB-ISOL	395	85	**47551**	50	Controlled	Controlled

TABLE II: Comparison of the datasets for isolated and continuous sign language recognition discussed in this work.



(a) Age distribution (b) Gender distribution

(c) Linguistic level (d) Handedness distribution

Fig. 1: Properties of recorded signers for the LSFB dataset. Age (a) shows a predominance of signers aged between 26 and 45 years old. The gender (b) and linguistic level (c) proportions are balanced. There is a majority of right-handed signers (d).

align each French translation sentence with its corresponding segment in the video. During the annotation process, special care was taken to annotate each sign and its variants with different labels. Figure 2 shows three variants of a LS extracted from the LSFB Corpus. When a LS is used as a proper noun, its annotation is prefixed with *NS* (Name Sign), meaning the sign refers to a person or a place. For instance, Belgian signers use the LS *star* to refer to the city of Bastogne due to a famous star-shaped monument, while foreigners will rather spell the city name. This precision in the annotations may be useful when building a continuous SLR system.

B. LSFB Continuous Dataset

The LSFB Corpus is available online[1] in a form convenient for linguistic research but hard to exploit programmatically. This motivates the creation of a curated dataset that would be easier to process automatically. This curated dataset contains all the videos from the LSFB Corpus, along with (1) their original XML annotation files cleaned up (2) a preprocessed version of those annotations stored in CSV format.

[1]www.corpus-lsfb.be

The whole LSFB Continuous Dataset (LSFB-CONT) currently depicts 6883 different lexical signs, making it the biggest public dataset for continuous SLR, both in terms of video length and vocabulary size.

C. LSFB Isolated Dataset

The LSFB-CONT dataset contains a large vocabulary compared with the other sign language datasets, as shown in Table II. However, the LSFB Corpus is heavily unbalanced, as the majority of signs are executed only a few times, and most common signs are executed hundreds of times. For this reason, we also built an isolated SLR dataset, called LSFB-ISOL, obtained from a subset of the LSFB-CONT dataset. LSFB-ISOL was produced by isolating all signs occurring 40 times or more in LSFB-CONT. Indeed, Joze et al. [16] have showed that below 40 examples, the performance of deep learning algorithms for video recognition drops significantly. This resulted in a new dataset, with a vocabulary of 395 glosses, and at least 40 occurrences per gloss.

D. Comparison With Other Datasets

Figure 3 compares the respective sign length distributions of the *GSL, MS-ASL, Phoenix Weather, ASLLVD* and *LSFB-ISOL* datasets. We can observe that signs extracted from continuous sign language datasets are made faster, and that LSFB signers are performing their signs faster than the signers of the other datasets. This difference could be a result of the recording setting of the signs, since speakers talking freely tend to perform signs faster than signers reading a script or translating a spoken language speech. The sign lengths standard deviation of the MS-ASL videos is larger than for other datasets. MS-ASL is made of educational videos, and some of the signers perform their signs either multiple times in a row or very slowly to help language learners. These repetitions are not isolated thus, a video of MS-ASL can contains multiple examples of a same sign.

IV. Baseline Methods

This section report results obtained on LSFB with classic models for action or gesture recognition. We describe the models retained and the preprocessing applied to our data for the experiments. Our results are reported along with results obtained on two other SL datasets to provide an order of magnitude of what is possible to achieve today for SLs.

Fig. 2: Three series of successive frames extracted from videos in the LSFB Corpus. The first row shows a lexical sign meaning *wait*; the two other rows depict variants of the same sign meaning *stop*. A typical example is the two-hand version of a one-hand sign indicated by the *-2h* suffix in the sign label (third row).

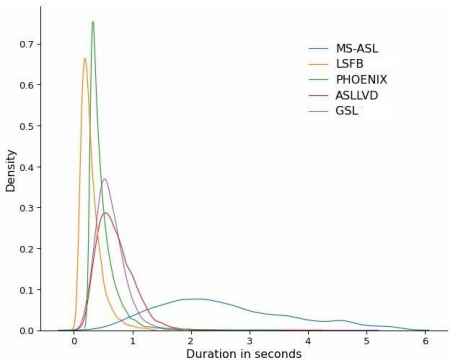


Fig. 3: Distribution of sign lengths in the SLR datasets.

A. Retained Methods

Many Deep Learning architecture were created for action recognition. Some of them also led to great results on simple gesture recognition. As SLR is a subfield of gesture recognition, we reuse those successful model to establish our baseline on our new datasets. The retained methods are the following:

CNN + RNN [17]: This method leverages successful models for image recognition by combining them with a sequence model. Each frame of the video is sent to a convolutional network to obtain an embedding. The sequence of embeddings is passed to a Recurrent Neural Network (RNN). The output of the RNN is then used to predict the correct label. In our implementation, the CNN embedding is provided by a VGG-16 network [18] pre-trained on ImageNet [19]. We removed the classification layer and the gradient information is computed only for the six last layers of the network. The RNN part of the model consists of a LSTM layer.

C3D net [20]: Introduced by Tran et al. [20] to test their 3D Convolution layer, this model aims to extend the capability of an image recognition architecture based on 2D convolutions, by allowing the convolution layers to process sequential data. Our implementation reproduces the architecture presented by Tran et al. and has been trained from scratch.

Inflated 3D Convolution Networks (I3D) [21]: This architecture aims to benefit from large pre-trained 2D convolution networks (e.g., Inception-V2 and VGG) by transforming their 2D convolution layers into 3D Convolutions. Training C3D networks is much faster as the model benefits from the pre-training on large image datasets, such as ImageNet. They are currently the state-of-the-art for various action recognition tasks such as the Charades Challenge [22]. We use Inception-V1 I3D introduced by Carreira et al. [21], pre-trained on the ImageNet [19] and Kinetics-600 datasets [23].

The above models are all implemented with PyTorch [24]. The source code for our baseline is publicly available[2].

B. Experimental Setup

The purpose of the experiment is to provide an idea of how popular video recognition methods perform on the LSFB dataset, but we replicate these experiments on two other SL datasets to compare the obtained results. We retained the MS-ASL-100 datasets and the GSL datasets. As the MS-ASL 100 dataset is an isolated SL dataset, the models are trained for isolated SLR. Thus, only the LSFB-ISOL is used for comparison purpose and we also use an isolated version of the GSL dataset made of signs isolated from their continuous videos.

Each dataset is made of videos showing only one sign. The preprocessing applied to each video is illustrated in Figure 4. First, each video is resized to 270 pixels, while keeping its aspect ratio. Then, the video is trimmed to 50 frames. Too short videos (< 50 frames) are looped, while too long videos (> 50 frames) are transformed in multiple videos with randomly selected first frame. Finally, patches of 224×224 pixels are randomly cut from the video and submitted to the networks.

The preprocessing is the same for each dataset and each model except for CNN + RNN models. As RNNs accept inputs of different lengths, too short videos did not have to be looped.

[2]https://github.com/Jefidev/Gesture-Recognition-Experiments

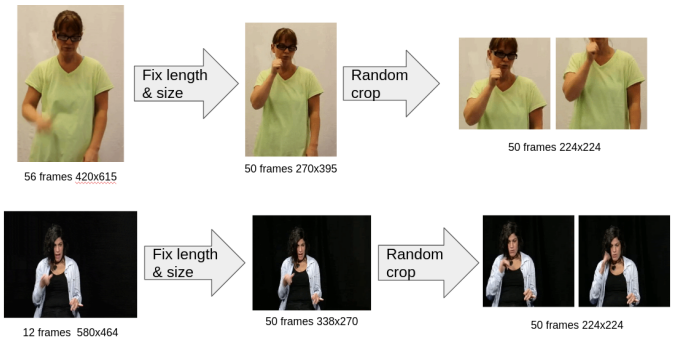


Fig. 4: Preprocessing of a video frame from the MS-ASL (top) and LSFB (bottom) datasets. Two random crops are shown.

C. Experimental Results

Table III reports the results obtained for each dataset. The CNN + RNN model is not able to learn useful features for SL classification. C3D also fails to perform well especially for MS-ASL. The I3D model is, by far, the best model for every SL dataset. The performance obtained on MS-ASL are close to the one of the other continuous datasets. The varying background and recording conditions makes it a challenging isolated SL dataset. Better results are achieved on LSFB compared to GSL despite the fact that they share common properties. This might be a consequence of the size of the LSFB dataset which is 15% bigger than GSL, but also because the preprocessing was calibrated for the LSFB dataset. Sometimes, The 224×224 crops cut important parts of the video for the GSL dataset.

	VGG + LSTM	C3D	I3D
MS-ASL 100	0.8%	1.3%	53%
GSL isol	6.1%	8.6%	36.5 %
LSFB-ISOL	3.6%	6.4%	51.5%

TABLE III: Percentage of Top-1 accuracy for isolated SLR.

V. Qualitative Discussion

This section provides a qualitative analysis of the best performing model on LSFB-ISOL dataset. The precise annotations of the LSFB datasets allows us to focus on special cases such as signs variants. A saliency map is computed for these special cases and our findings are presented.

A. Error Analysis in Sign Recognition

The I3D model provides the best performances for each SL datasets. Despite that, we identified a lot of recurrent mistakes occurring especially for predicting the less represented signs, discriminating sign variants and signs sharing common hand configurations. To better understand the root cause of the model mistakes, it is useful to look at the saliency maps [25] of the model. In our case, we computed a *vanilla gradient* saliency map for videos from each class of the LSFB-ISOL dataset. This allows us to highlight the region of the video playing the most important role in the prediction. Figure 5 shows examples of saliency map for LSFB videos. It also shows examples of common mistakes. For instance, signs sharing hand configurations like signs *LIEN* (link) and *RELATION* are often confused. Their difference lay in the sequential execution of the sign rather than in their hands configuration. In order to better distinguish those signs, the model should look at the first frames of the video. The sign *RELATION* starts with the hands apart joining to a final configuration, while the sign *LIEN* starts directly with the hands joined. The fact that they are confused with each other seems to indicate that I3D bases its prediction primarily on the hands configuration or it has troubles to capture the evolution of a sign over time.

Another frequent misclassification was identified by looking at the signs annotated as a variant of a common signs in the LSFB datasets. When faced with the two-hand variant of a single-hand sign I3D tends to always predict the label for the one-hand version. Figure 5 gives an example of such a confusion, with the signs *MAIS* (but) and *MAIS-2H* (its variant). When looking at the saliency map extracted for the video, we observe that the saliency is systematically higher on the right hand than on the left hand, indicating that the right hand plays a more important role in the prediction than the left hand. The signers tend to use their dominant hand to sign, and a majority of LSFB signers are right-handed. Hence, it is not surprising to see that the model learns to focus on the right hand as it plays an important role in nearly all videos it encountered during the training phase.

Other recurrent errors indicate that the model has troubles to distinguish small variations of fingers position or to determine if the hand palm faces up or down, leading to confusion between several signs. Also, we could not find any evidence supporting the fact that the model looks at facial expressions for its predictions. Yet, facial expression and eye gaze are often used in SL conversations to convey information. This means that either the representation capacity of the model is not strong enough to consider the facial expression, or that the facial expression does not provide any useful information in our isolated SLR setting. The resolution of the videos provided to the model could also be too low, thereby making details such as gaze direction impossible to exploit.

Despite those weaknesses, the model is able to perform well on several sign language datasets and it has correctly learned to track the hands of the signers even when occlusion occurs. We can then conclude that the Inflated Inception-V1 model is a good candidate for gesture recognition applications, although it still lacks representation capacity to completely solve the isolated SLR problem. Designing better preprocessing for SL could, also, help to improve the model accuracy.

B. Sign Representation Learned by the Model

Another way to diagnose the model behavior is to look at the feature map (or embedding) learned by the model. We decided to plot the embedding learned by the I3D model just before the prediction layer. By plotting the embedding of each video of the test set, we are able to visualize the representation of the signs used internally by the I3D models. Figure 6 shows

Fig. 5: Saliency maps computed on LSFB videos. *LIEN* and *RELATION* are instances of signs sharing common had configuration. *MAIS* and its variant *MAIS-2H* are also compared. I3D seems to have learn to focus its attention on the hands.

a UMap projection of the gathered I3D embedding. Video representations are grouped in distinct clusters. When looking in detail at those clusters, signs depicting similar concepts are close to each other in terms of embedding (e.g., the video recordings for *Think* and *Understanding*). This is due to the fact that signs representing close concepts tend to use the same kind of hands configuration and movements. This comfort us in the idea that the model uses relevant information in order to classify the various signs of the dataset.

VI. Threats to Validity

This section aims to identify all decisions that we made that could affect the quality of the results presented in this paper.

During the various experiments we conducted, the same preprocessing steps were applied to each dataset and for each SLR method considered. However, better performance could probably be reached if the preprocessing was fine-tuned for each method and each dataset. In addition, the way we handled short videos could be criticized. In our experiments we decided to loop such videos to make them reach the length expected by the SLR models. As we expect the model to take into account the relative movements of the signers hands, providing videos that, at some point, jump from the last frame to the first one could negatively impact the learning process.

Models are trained on datasets depicting solely Caucasian individuals. Systems built upon such models may potentially perform poorly when processing videos of showing signers belonging to other ethnic groups.

Fig. 6: UMap visualization of the embedding outputed by the I3D model. A larger resolution images is available at https://figshare.com/s/ba0485a95e74b63b750c

Finally, the results obtained on the MS-ASL dataset contain 20% less signs than what was initially reported by its authors, since some of the videos composing MS-ASL were removed from YouTube in the meantime.

VII. Conclusion and Future Work

This paper introduces two new datasets for sign language recognition. LSFB-CONT[3], for continuous Sign Language Recognition (SLR), depicts conversations between 100 signers. This new dataset includes more natural sentence constructions than previous datasets. Therefore, a model able to perform well on the LSFB Corpus is more likely to generalize better to real life sign language conversations. It is currently the largest dataset for continuous sign language recognition in terms of number of signers, vocabulary size and videos length. LSFB-ISOL, a subset of LSFB-CONT containing the most frequent signs is also introduced for isolated SLR. It contains a vocabulary of 395 classes with at least 40 examples for each class. LSFB-ISOL is the largest dataset for isolated SLR in terms of number of video clips. These two new datasets lack variability in terms of background, camera angle and lightning conditions. Yet, they are representative of the recording environment of other SL corpora in construction. We presented a review of the popular datasets for SLR. LSFB-ISOL is compared to other state-of-the-art SLR datasets. We then provide firsts results obtained on our datasets and a qualitative analysis of the best performing model was conducted. This analysis showed that the model was able to learn hand tracking of the signers even when occlusion occurs. We also observed that the model has trouble to focus on small hands configurations details, such as variations in fingers position. The model focuses primarily on the right hand to make predictions, and does not seem to exploit eye gaze or facial expression information to distinguish similar signs. In our future work, we plan to improve results obtained in this paper, by exploring other pre-processing methods and by developing new architectures more suited for the particular SLR problem. We also intend to integrate several models in a widely available system and to exploit user feedback to incrementally improve our results. We also aim to contribute to the challenging problem of continuous SLR.

Acknowledgments

The authors thank Adrien Bibal, Arnaud Bougaham and Paul Temple for their comments and discussions on this paper.

References

[1] A. S. Al-Shamayleh, R. Ahmad, M. A. M. Abushariah, K. A. Alam, and N. Jomhari, "A systematic literature review on vision based gesture recognition techniques," *Multimedia Tools and Applications*, vol. 77, no. 21, pp. 28 121–28 184, 2018.

[2] C. Vogler and D. Metaxas, "Parallel hidden markov models for american sign language recognition," in *Proceedings of the Seventh IEEE International Conference on Computer Vision*, 1999.

[3] S. Mitra and T. Acharya, "Gesture recognition: A survey," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 37, no. 3, pp. 311–324, 2007.

[4] W. C. Stokoe, "Sign language structure: An outline of the visual communication systems of the american deaf," *Journal of Deaf Studies and Deaf Education*, vol. 10, no. 1, pp. 3–37, Jan. 2005.

[5] R. Yang, S. Sarkar, and B. Loeding, "Handling movement epenthesis and hand segmentation ambiguities in continuous sign language recognition using nested dynamic programming," *IEEE transactions on pattern analysis and machine intelligence*, vol. 32, no. 3, pp. 462–477, 2009.

[6] T. Johnston, "Corpus linguistics and signed languages: no lemmata, no corpus," in *Proceedings of the Sixth International Language Representation and Evaluation Conference*, 2008, pp. 82–87.

[7] H. Cooper, B. Holt, and R. Bowden, "Sign language recognition," in *Visual Analysis of Humans*. Springer London, 2011, pp. 539–562.

[8] M.-A. Sallandre, A. Balvet, G. Besnard, and B. Garcia, "Étude exploratoire de la fréquence des catégories linguistiques dans quatre genres discursifs en LSF," in *Revue de linguistique et de didactique des langues*, 2019.

[9] L. Naert, C. Reverdy, C. Larboulette, and S. Gibet, "Per Channel Automatic Annotation of Sign Language Motion Capture Data," in *Workshop on the Representation and Processing of Sign Languages: Involving the Language Community, LREC 2018*, May 2018.

[10] J. Forster, C. Schmidt, T. Hoyoux, O. Koller, U. Zelle, J. Piater, and H. Ney, "Rwth-phoenix-weather: A large vocabulary sign language recognition and translation corpus." in *LREC*, 2012, pp. 3785–3789.

[11] N. Adaloglou, T. Chatzis, I. Papastratis, A. Stergioulas, G. T. Papadopoulos, V. Zacharopoulou, G. J. Xydopoulos, K. Atzakas, D. Papazachariou, and P. Daras, "A comprehensive study on sign language recognition methods," *preprint arXiv: 2007.12530v1*, 2020.

[12] V. Athitsos, C. Neidle, S. Sclaroff, J. Nash, A. Stefan, Q. Yuan, and A. Thangali, "The american sign language lexicon video dataset," in *CVPR Workshops*. IEEE, 2008, pp. 1–8.

[13] H. Wang, X. Chai, X. Hong, G. Zhao, and X. Chen, "Isolated sign language recognition with grassmann covariance matrices," *ACM Transactions on Accessible Computing*, vol. 8, no. 4, pp. 1–21, 2016.

[14] D. Bragg, O. Koller, M. Bellard, L. Berke, P. Boudreault, A. Braffort, N. Caselli, M. Huenerfauth, H. Kacorri, T. Verhoef *et al.*, "Sign language recognition, generation, and translation: An interdisciplinary perspective," in *The 21st International ACM SIGACCESS Conference on Computers and Accessibility*, 2019, pp. 16–31.

[15] L. Meurant, "Corpus lsfb. corpus informatisé en libre acces de vidéo et d'annotations de langue des signes de belgique francophone. namur: Laboratoire de langue des signes de belgique francophone (lsfb lab), frs-fnrs, université de namur," 2015.

[16] H. Joze and O. Koller, "Ms-asl: A large-scale data set and benchmark for understanding american sign language," *British Machine Vision Conference*, vol. 30, 2019.

[17] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *Proceedings of IEEE CVPR*, 2015, pp. 2625–2634.

[18] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *ICLR*, 2015.

[19] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," in *CVPR09*, 2009.

[20] D. Tran, L. D. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "C3D: generic features for video analysis," *1412.076v1*, 2014.

[21] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," *IEEE CVPR*, pp. 4724–4733, 2017.

[22] G. A. Sigurdsson, G. Varol, X. Wang, A. Farhadi, I. Laptev, and A. Gupta, "Hollywood in homes: Crowdsourcing data collection for activity understanding," *European Conference on Computer Vision*, pp. 510–526, 2016.

[23] J. Carreira, E. Noland, A. Banki-Horvath, C. Hillier, and A. Zisserman, "A short note about kinetics-600," 2018.

[24] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *NeurIPS*, 2019, pp. 8024–8035.

[25] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," in *ICLR Workshop*, 2014.

[3] lsfb.info.unamur.be