Annotation of LSF subtitled videos without a pre-existing dictionary

Julie Lascar, Michèle Gouiffès, Annelies Braffort, Claire Danet

Annotation of LSF subtitled videos without a pre-existing dictionary

Julie Lascar, Michèle Gouiffès, Annelies Braffort, Claire Danet
Université Paris-Saclay, CNRS, LISN
Campus Universitaire Bat 507, rue du Belvédère, 91405 Orsay, France
julie.lascar, michele.gouiffes,annelies.braffort@lisn.upsaclay.fr, claire.danet@gmail.com

Abstract

This paper proposes a method for the automatic annotation of lexical units in LSF videos, using a subtitled corpus without annotation. This method, based on machine learning and involving linguists for added precision and reliability, comprises several stages. The first consists of building a bilingual lexicon (including potential variants of a given lexical unit) in a weakly supervised manner. The resulting lexicon is then refined and cleaned by LSF experts. This data serves next to train a supervised classifier for automatic annotation of lexical units on the Mediapi-RGB corpus. Our Pytorch implementation is publicly available.

Keywords: French Sign Language, bilingual lexicon, sign spotting, automatic annotation

1. Introduction

Sign languages (SL) are natural languages used in Deaf communities. Their visuo-gestural nature allows information to be conveyed simultaneously using multiple articulators (hands, arms, body and facial components). SL content, where iconicity plays a central role, is spatially organised. The analysis of SL videos for annotation, recognition or translation requires the design of appropriate computer vision and natural language processing methods. A large amount of data is also needed, for instance videos with annotations, translations or subtitles. However, this kind of data is still scarce, particularly for French Sign Language (LSF).

Our study aims to devise a method for annotating videos with lexical signs with utmost precision while simultaneously reducing the manual annotation time required by experts.

After a short review on the related works (section 2), the paper describes a three-stages approach for automatic annotating LSF videos subtitled in French. The first stage (section 3) consists in a weakly supervised segmentation of specific signs in the videos, without use of any isolated example. The quality of the outputs is next assessed by LSF experts (section 4). Then, a supervised classifier (section 5) is trained using the previous annotations. In Section 6, experiments are conducted to investigate the impact of expert analysis on supervised classification and the scalability of our model.

2. Related works

The automatic annotation of lexical units in a SL video consists in determining the presence of such units and their temporal localization. We are therefore interested in *sign-spotting* approaches, which highly rely on *video encoding* methods. Regarding *LSF*, there is unfortunately a scarcity of data for effective automatic processing.

Sign spotting in continuous videos. Sign spotting consists in localizing a sign temporally in a continuous video given a query. This is generally done by computing similarities between an example of the query sign and the video, and finding local maxima. While first works (Yang et al., 2009; Buehler et al., 2009) relied on similarities computed from hand-crafted features and involved limited dictionaries, more recent methods use learned classifiers, as in Jiang et al. (2021) where a transformer architecture is used. When available, subtitles can be used for a weak supervision as in Momeni et al. (2020), where multiple instance learning is leveraged. In Albanie et al. (2020), multiple modalities are used in the sign spotting, such as “mouthing”. These approaches rely on a dictionary of isolated signs, which is not available for all SL.

In Momeni et al. (2022), several methods are proposed to increase the density of annotations on continuous signing data. For instance, they localize unknown signs (not present in a lexicon), by selecting keywords in subtitles and finding the corresponding signs within continuous signing data. Our work enriches this technique to precisely locate the beginning and end of a sign.

Video encoding. The choice of the video encoding has a large impact on sign spotting performances. Most SL recognition models are inspired from the action recognition domain. First of all, a large number of works use pose-based representations to encode videos (Belissen et al., 2020b; Ouakrim et al., 2023); it has advantages for SL, in particular invariance with respect to the setting and the appearance of the signer, to keep only the gesture information and, to a certain extent,

Proceedings of the 11th Workshop on the Representation and Processing of Sign Languages, pages 204–212
25 May 2024.

facial expressions. However, recent studies have obtained some very interesting results, by using pre-trained models designed for action recognition in videos based on RGB images. Examples of such models include I3D (Carreira and Zisserman, 2017) and the more recent Video Swin Transformer (Liu et al., 2022). Fine-tuning these models specifically for sign recognition tasks yields impressive results, as demonstrated in tasks like fingerspelling recognition (Prajwal et al., 2022), sign spotting (Momeni et al., 2022) or translation (Li et al., 2020).

LSF resources. These various methods require the use of large quantities of data. However, many SL are under-resourced, such as for LSF (Kopf et al., 2022). Note however the 8h dialogues dataset DictaSign (Belissen et al., 2020) which is partly annotated by linguists and useful to recognize signs in context whether they are lexical (Ouakrim et al., 2023) or non-lexical (Belissen et al., 2020a,b). Recently, the corpus Mediapi-RGB has been released (Bull et al., 2024) comprises 86 hours of videos in LSF produced by deaf journalists or presenters from the bilingual online media Média'Pi![1], with French subtitles produced by Deaf translators (Ouakrim et al., 2024). During a post-production phase, the videos are subtitled by professional translators. These translations are manually aligned with the corresponding SL video content. Mediapi-RGB is therefore, by construction, a perfectly aligned bilingual corpus. Our annotation system is built upon this dataset.

3. Step 1: Weakly supervised annotation

The lack of a freely available bilingual LSF/French dictionary led us to build our own one, by using the bilingual Mediapi-RGB corpus.

3.1. French vocabulary choice

The first step implies to draw up a list of French words for which the corresponding signs are searched in the videos. The initial list was established from the subtitles by selecting lexical terms belonging to defined categories: days of the week, months, cities, countries, sports, vocabulary linked to current events (mask, unemployment, yellow waistcoats, film, etc.). These words were selected because they appear frequently in the dataset and are supposed to have stable meaning depending on the context.

For each word of this list, all video clips representing its LSF equivalent have to be segmented automatically and precisely, i.e. the full sign has to be detected, with less transitions as possible. The main difficulty is the lack of isolated examples of the signs to be detected, since the videos are subtitled but not annotated. The method used for this task is outlined in the next section (3.2).

[1] https://www.media-pi.fr/

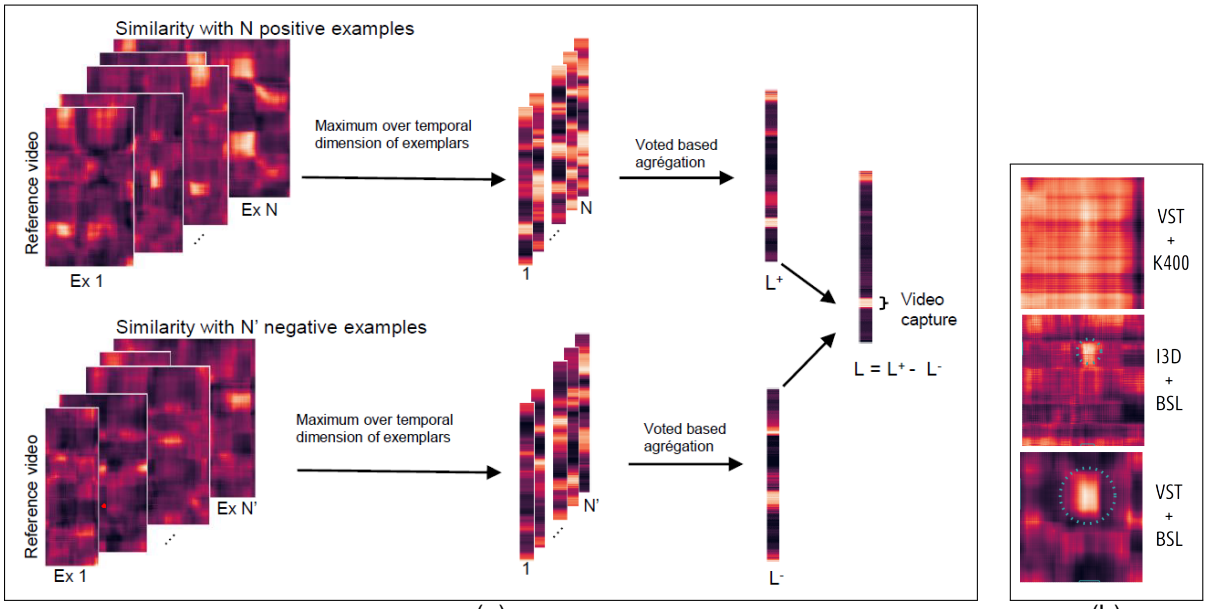
3.2. Sign Spotting method

The technique described in Momeni et al. (2022) is used with different settings in order to fit to our dataset and our own objectives. Let us describe its principle on an example shown in Figure 1(a).

In this example, the objective is to capture the visual representation of “rugby” in a reference video. A similarity matrix (with values ranging in $[0, 1]$) is computed between this reference video and N other positive examples, which are videos with subtitles containing the word "rugby". For each of the N matrices, the maximum similarity value of each row is kept, leading to N vectors that are next aggregated using a voted scheme (threshold set at 0.6). This results in a vector L^+, which shows areas of significantly high similarity between the reference video and the positive examples. In these areas of high similarity, it is very likely to find the sign corresponding to “rugby”, but it may contains other frames belonging to transitions, or even signs that often appear in the same context. To avoid capturing these frames, the process is repeated using N' negative examples, i.e. videos for which the subtitle does not contain the word “rugby”. It yields a vector L^- which is useful to locate these non-positive frames. Finally, the vector $L = L^+ - L^-$ improves the localization of the visual representation(s) corresponding to the word “rubgy”. Unlike Momeni et al. (2022), our video clips have various sizes: a video clip is made for any consecutive sequence of at least 3 frames for which L is above a fixed threshold (set at 0.5). The maximum number of positive videos N is set to 100, and N' is set to $3 \times N$. Since the effectiveness of this method heavily relies on the way videos are encoded, the next section discusses the choice of video encoding.

3.3. Videos encoding

To optimise the performance of this encoding step, three methods are compared. Figure 1(b) shows three similarity matrices computed between two videos which are supposed to contain the sign corresponding to the word “village”. The same couple of videos is used in each case, with the following subtitles on the vertical axis “A l’arrière, c’est-à-dire dans les *villages*, comme tous les hommes sont partis combattre,” (“In the rear, i.e. in the villages, as all the men had left to fight,”), and on the horizontal axis: “126 villes ou *villages* ont été placés en état de catastrophe naturelle.” (“126 towns and villages have been declared natural disasters.”).



(a) (b)

Figure 1: (a) Detecting unknown signs by spotting through exemplars. (b) In this example, we calculated the frame-by-frame cosine similarity between two videos with the word “village” in the subtitle, which had previously been encoded with 3 different backbones. The lightest area is the one with the greatest similarity, allowing us to locate the image sequences that visually represent the word “village”.

Each matrix relies on a different video encoder: at the top, a Video Swin Transformer (VST) trained for action recognition with Kinetic 400 (Liu et al., 2022); in the middle, an I3D model trained for sign recognition with BSL data (Renz et al., 2021); at the bottom, a Video Swin Transformer also trained on sign recognition with BSL data (Prajwal et al., 2022; Bull, 2023). The latter is selected for our study since it clearly provides the most discriminant similarity.

3.4. Refinement of the method

After this stage, various errors may occur and need to be thoroughly investigated and eliminated, as automatically as possible.

Dealing with the variability of form or meaning. Some signs may vary in form depending on the signer. For example, some signs representing the months can be made with one hand or with both hands, depending on the signer. Others can also be completely different in form. In these cases, the method fails in finding similarities. To overcome this problem, the videos are automatically clustered by signer[2] when the number of positives examples is high enough (superior to 20), before applying the similarity search.

In addition, the method could fail due to the polysemous nature of the chosen word, leading to distinct interpretations depending on the context. To address this issue, when necessary, we categorized the videos according to the word’s specific meaning in the context of each sentence before applying the previous method. To that aim, a Bert language model[3] (Devlin et al., 2019) is used.

Clustering video clips. For each query word, a classification is performed on the detected videos in order to discover potential variants. The videos are clustered into classes of similar form. A K-means algorithm is used to that aim and the number of clusters is determined using the Silhouette method (Rousseeuw, 1987). It selects the optimal number of clusters by simultaneously maximizing the distance between clusters and the density of points within each cluster.

Figure 2 shows an example of clustering result for words “Italy” (on the left) and “November” (on the right). For “Italy”, the larger group contains the videos that actually correspond to the sign “Italy”, while the smaller group contains detection errors. For “November” (right), the two groups correspond to two real variants: the two-handed variants on the left, and the one-handed variants on the right.

As the detection errors are automatically grouped during the clustering stage, the use of negative examples to prevent the detection of non-positive frames (section 3.2) may not always be necessary. Nevertheless, we have also employed negative examples for other purposes, as explained below.

[2]Beforehand, each video is labeled with the signer identity using the face recognition library Deepface.

[3]Specifically the bert-base-multilingual-cased version from Hugging Face.

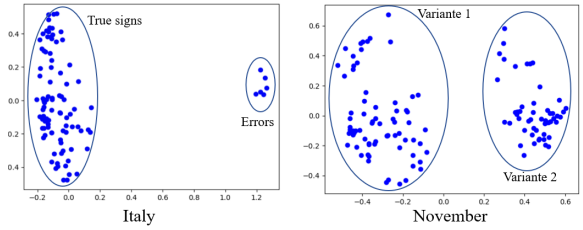


Figure 2: Visualisation of the sequences obtained for “Italy” and “November”. After reducing the size of the data using PCA, the sequences obtained for “Italy” and “November” were projected onto the two main axes.

Separating frequently associated signs. For certain words in our list, the model gathers video clips featuring the desired sign alongside another sign. For instance, in Mediapi-RGB, the word “Tokyo” is often associated with “Jeux Olympiques” (Olympic games). To capture only the desired sign, we employed the similarity search of section 3.2 by modifying the set of negative videos. As shown in Figure 3, instead of defining the set of negative videos as those for which the subtitles do not contain the word “Tokyo” (Classic method), the negative samples are defined such as they do not contain Tokyo but contain ‘Jeux Olympiques” or its equivalent (Custom method).

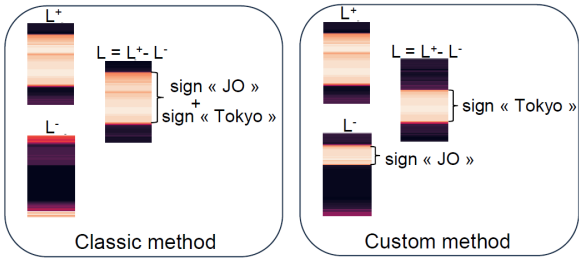


Figure 3: Splitting frequently associated signs using negative examples. In the Custom method (on the right), negative examples are used to locate the sign corresponding to “Jeux Olympiques” in a video for which the subtitle is “Tokyo 2021 Olympic Games are getting closer”. The final vector L is useful to precisely segment “Tokyo”.

This method is used when precision of the signs segmentation is poor. It proves to be very effective on our data.

After the first stage, a large number of videos clips are extracted from continuous videos. They have various sizes depending on the context and the signer. Each clip is associated with a label based on the word but taking into account any variations in form e.g. juillet_0, juillet_1, juillet_2 (July). As a result, our bilingual dictionary consists of a list of labels to which are associated LSF video clips containing lexical units.

It is worth noticing that the method is able to discover signs that are not currently available in existing online LSF dictionaries.

In order to assess the quality of the resulting lexicon, a first version of 36 labels is built and audited by LSF experts.

4. Step 2: Expert reviewing

The evaluation phase was carried out by two LSF experts. More specifically, the aim was to assess the quality of the segmentation of each clip for each label. To do this, three quality levels are defined:

1: when the sign is correctly segmented, that is when it is fully present and there is no frame belonging to the transition parts before or after the sign;

2: when it is acceptably segmented, that is some frames belonging to the transitions are present, or a few frames seem to be missing at the beginning or end of the sign;

3: otherwise. These are cases where we are able to identify the partial presence of the sign, i.e. it is truncated or accompanied by another sign, possibly not complete. Thus, these occurrences should not be kept for future use.

The choice between categories 1 and 2 is sometimes empirical, typically for signs that include a preparation or retraction phase, which can be blended into the transitions between signs.

Even when the occurrence is perfectly segmented, there may be variations in the shape of the sign, despite the solutions presented in the previous section. We felt it was important to identify the different types of variation so that we could decide whether or not to create separate classes. Three types of variations have been singled out:

- *Lexical*, where there are several signs associated with a given word, for example for certain months such as July.
- *Morphological*, such as the addition of a forward or backward movement with the signs expressing the days of the week, to specify that it is the day of the next or previous week.
- *Internal*, with changes in one of the parameters of the sign (handshape, location, orientation, contact), the number of repetitions or the posture of the dominated arm.

In the first two cases, the form or meaning is different, so separate classes are needed. In the third case, the variations are due to articulatory constraints or individual variants that do not require separate classes.

At the beginning of the process, we had 36 labels with a number of occurrences ranging from 5 to 213. This represented a total of more than 3,000 clips that were manually evaluated by the LSF experts. At the end of the process, we ended up with 53 labels (44 of which had more than 5 LSF examples). Indeed, some of the clusters had to be split because, for example, they contained variants with different meanings (e.g. Wednesday, next Wednesday, previous Wednesday). In addition, because we retained only the occurrences with a 1 or 2 quality level, the total number of video clips has been halved. Therefore, the number of occurrences for each label is lower (from 3 to 202), but the occurrences are more representative. The experiments of section 6 examine how expert enhancement affects classification performance.

5. Step 3: Supervised classification

Since we have a French-LSF lexicon (with or without refinement by experts), it becomes possible to design a supervised classification, which will be useful for annotating any continuous LSF video.

5.1. Preparing data

For this step, we select from a French-LSF lexicon the labels that have at least 5 LSF examples. Complete videos containing any of these labeled instances are retained. Each frame within these complete videos is assigned to a class label (coded as an integer). However, due to potential missed annotations, some signs may not have been annotated, leading to a partial ground-truth. For example, in a video with the subtitle “Hello, we are Tuesday, April 3rd,” we only captured the sign corresponding to “Tuesday”. The annotation for this video is in the form $[00...0066600...00]$, where 6 is the identifier for “Tuesday”. This annotation is incomplete since the sign corresponding to “April” is not annotated (nor the sign for “Hello”). Therefore, we trained models with data that is partially annotated, making model optimization challenging and quantitative evaluation approximate.

5.2. Model Architecture

The system architecture is illustrated in Figure 4:

- The first model extracts video features using a Video Swin Transformer trained on BSL data (the same model as used in Section 3).
- The second model is a lightweight straightforward MLP classifier. It takes the features as input and produces sequences of integers as output. Each integer in the output sequence identifies the class corresponding to each frame.

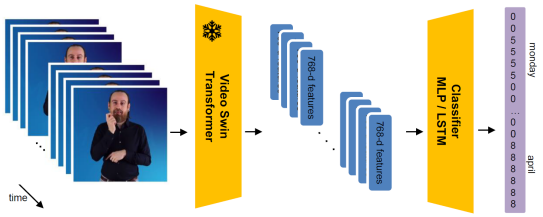


Figure 4: Model Architecture.

5.3. Training Setup

The classifier is trained with batches corresponding to non-shuffled features of videos for 15 epochs, using Adam optimization with an initially fixed learning rate of 1e-4. We also used L2 penalty (weight decay = 1e-5). The Video Swin Transformer was frozen and we initialized the classifier neural network’s weights with Xavier Initialization.

Loss. We used the cross-entropy cost function, which is particularly suitable for multi-class classification models. Since the dataset is highly imbalanced (90% of the images are annotated as 0), we applied weights to the cost function. The weights w_c assigned to each class c are defined as follows: $w_c = 1 - \frac{\text{number of examples for class c}}{\text{total number of examples}}$.

Metrics. To assess the quality of the models, we measure accuracy, F1-score, and recall, as follows. First, F1-scores $F1_c^i$ (or recall R_c^i) are computed for each video i and each class c present in the ground truth of video i. For each class c, these scores are averaged to get $F1_c$ (or recall R_c). The final F1-score (or recall R) is finally obtained by averaging the $F1_c$ (or R_c).

As mentioned previously, the ground truth annotation is partial since not all occurrences are identified. However, when annotated, the signs are well segmented and reliable. Therefore, during training, we choose the model with the best recall to minimize the likelihood of missing true positives.

Sign Classifier. We tested several architectures of the classifiers, considering both MLPs and LSTMs with one or two layers and hidden layers of 100, 200, or 300 neurons.

In this study, we focus on experiments involving a 2-layer MLP with 200 neurons. We introduced a Normalization layer, used the ReLU activation function after the first layer, and applied a softmax at the output. For the evaluation, a smoothing function is used to eliminate isolated signs.

6. Experiments

We carry out several experiments on this model. The first experiments aim to study the impact of the expert analysis and their modification of the dictionary on the supervised annotation. This is made both quantitatively and qualitatively. A second experiment aims to increase the size of the initial vocabulary, in order to evaluate the scalability of our procedure.

6.1. Expert versus non expert

This experiment explores the contribution of experts in the data sorting process.

Data. In the concluding phase of the initial stage (Section 3), we organized, for each word of our list, a set of automatically clustered videos. Subsequently, these videos underwent a preliminary manual sorting process, involving the removal of clusters corresponding to detection errors and the adding of potential variants. This sorting was carried out by non-experts[4] in a first step, and then by experts (as detailed in Section 4). We consequently obtained a non-expert and an expert dictionaries $D1$ and $D2$, from which we acquired annotated videos (Table 1).

	nb. classes	nb. signs	nb. annot. videos
w/o expert	37	3137	2657
w expert	45	1773	1613

Table 1: Data quantification - w/o and w expertise. In each case, there is an additional class corresponding to a null class.

Note that the scenario involving expertise is more challenging because there are more classes and fewer occurrences per class.

Quantitative results. Table 2 presents the results obtained for two classifiers trained with the setup described in Section 5.3, using data sorted with and without expert involvement. In both cases, the data was divided into training, validation and test sets. For consistency, the same videos were selected for the validation and the test set (respectively 227 and 225 videos).

[4]Non-experts: Machine Learning computer scientists who, through working with sign language videos, are presumably capable of comparing sign videos and decide if the signs correspond to the same lexical unit. They do not have the expertise to determine whether a sign will be segmented perfectly, nor to distinguish fine variations of signs.

Data	Recall	F1	Accuracy
w/o expert	0.85 (± 0.008)	0.77 (± 0.003)	0.95 (± 0.005)
w expert	0.85 (± 0.017)	0.78 (± 0.01)	0.95 (± 0.004)

Table 2: Scores on Test set of the classifiers trained on data with and without expertise.

As explained before, the non-expert dictionary $D1$ contains 36 labels, while the expert one $D2$ contains 44 labels. The new labels are created by separating variants of form or meaning. In some cases, the differences in the forms can be tricky to perceive, which is why the first automatic step grouped them in a single class.

This is the case for example for the $D1$ class "mercredi" (Wednesday) that has been split by experts into 3 labels in $D2$, which are "mercredi" (Wednesday), "mercredi dernier" (previous Wednesday), "mercredi prochain" (next Wednesday). These three signs with different meanings differ only in the strong hand movement. In SL, time is expressed along the camera axis, with the past to the rear, the present at the level of the signer and the future forwards. What differs on this axis alone is of course more complicated to distinguish in video-type data, which raises a greater challenge to the classifier. However, neglecting this expertise step can lead to major errors which will subsequently have a detrimental effect on task performance.

Thus, our two classifiers are trained on a dictionary $D1$ with fewer classes and more occurrences, but less precision on both form and meaning, and a dictionary $D2$ with more classes and fewer occurrences, but more precision on form and meaning. The performance of the two classifiers is very promising and shows that, despite the fact that $D1$ contains more data, using $D2$ produces similar scores. There is no difference despite the more challenging conditions of the expertised lexicon and, above all, much greater precision.

Qualitative analysis. Figure 5 shows an example of automatic annotation of lexical units on a test video with the subtitle "But the G7 countries - Canada, France, Germany, Italy, Japan, the United Kingdom and the United States - reached an agreement on Saturday". For this qualitative study, an annotation by a LSF expert has been done on the video using Elan software[5].

In both cases, all signs are recognized, and are relatively close to the ground truth. Sign segmentation differs slightly between the two classifiers. In this example, the "with expert" classifier is able

[5]https://archive.mpi.nl/tla/elan - Max Planck Institute for Psycholinguistics (Nijmegen)

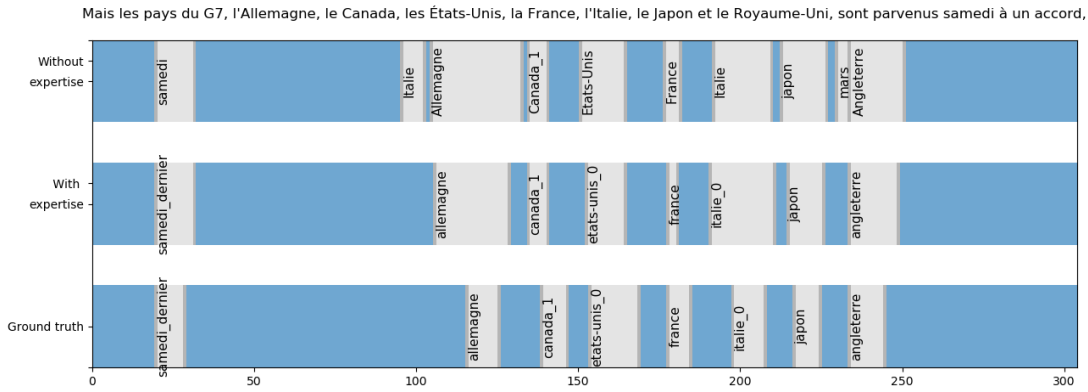


Figure 5: Comparison between the predictions of the non-expert (top), the expert (middle) classifiers and a ground truth (bottom) on a test video.

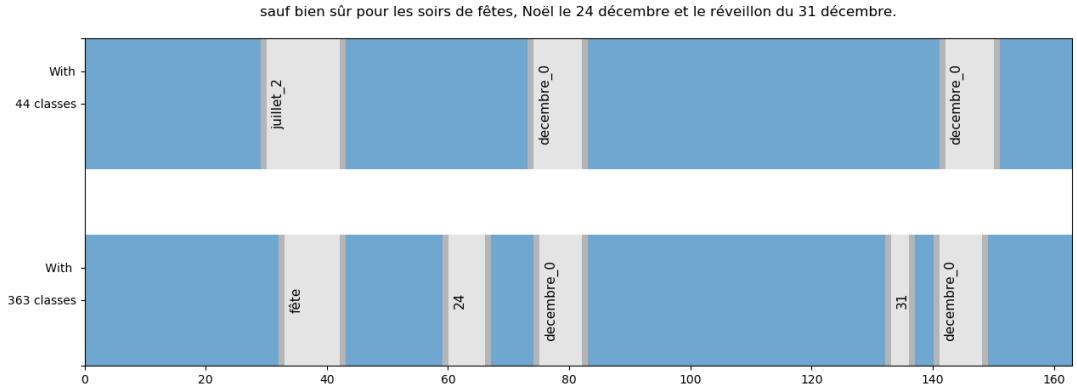


Figure 6: Comparison between the predictions of the 45 and the 364 classes classifier.

to eliminate two insertions present in the version "without expert" classifier (insertions of "Italy" and "March").

6.2. Towards a much larger dictionary

The experiment is extended by increasing the number of the dictionary entries, following theses steps:

- Creation of a dictionary comprising 363 labels: 44 sorted by experts (same as in 6.1), to which we added 319 labels sorted by non-experts[6]. In total, 7339 occurrences of signs were collected.
- Annotation of 6047 videos using this dictionary.
- Training of a 364-classes classifier using the training setup described in Section 5.3.

The model achieved an accuracy of 0.93, a recall of 0.65, and a F1-score of 0.63 on the test set.

The figure 6 illustrates the predictions of the expert 45-classifier from Section 6.1 and the predictions of the new 364-classifier on a test video with the subtitle: "except, of course, for Christmas Eve on 24 December and New Year's Eve on 31 December."

The 364-classifier predicts five positive signs, while the 45-classifier only three. "juillet_2", a variant of "juillet" (July), corresponds to the same sign as "fête" (celebration)[7]. This suggests that the 364-dictionary contains two labels for the same form with a different meaning. This is usually not recommended, but the classifier appears to perform well.

7. Conclusion and prospects

The paper has presented a system designed for the automatic annotation of lexical units in LSF videos, with an initial vocabulary of 36 labels. This lexicon has been extended to 44 and then 363 labels. Our Pytorch implementation is publicly available.

The proposed method highlights the transferability of a SL video encoder from one SL (BSL) to another one (LSF).

A non expert dictionary has been compared to

[6]The sorting of videos was conducted by non-experts due to time constraints, but we nevertheless believe it would be beneficial for this step to be carried out by experts.

[7]This is due to the celebration of "14 juillet".

an expert one, in the context of sign recognition in continuous videos. Without expertise, results are very convincing. Yet it hides a problem, which is a lack of precision to distinguish certain signs, notably when they differ according to the motion along the camera axis (e.g. last Wednesday versus next Wednesday). It has shown that, even when using elaborated video encoders such as Video Swin Transformer, not all the subtleties of SL are caught, such as the use of space, which can change the meaning of signs . In our experiments, the expertise has provided a refinement of the classes which is overriding to keep the meaning of the utterances.

Progress is underway, with the next step being to expand the dictionary, coupled with expert review to achieve a vocabulary of 1000 words. The final goal is to annotate the entire Mediapi-rgb corpus as finely as possible, while simultaneously creating a sufficiently large dictionary to train specific video encoding models for LSF.

8. Acknowledgements

We would like to thank Media'Pi! for allowing us to use the invaluable bilingual resources they produce for our research. We thanks also Diandra Fabre from Gipsa-Lab for her help with the non-expert data cleaning.

9. Bibliographical References

S. Albanie, G. Varol, L. Momeni, T. Afouras, J. S. Chung, N. Fox, and A. Zisserman. 2020. Bsl-1k: Scaling up co-articulated sign language recognition using mouthing cues. In *ECCV*, volume 12356, pages 35–53.

V. Belissen, A. Braffort, and M. Gouiffès. 2020a. Dicta-Sign-LSF-v2: Remake of a continuous French Sign Language dialogue corpus and a first baseline for automatic sign language processing. In *LREC*, pages 6040–6048, Marseille, FR.

V. Belissen, A. Braffort, and M. Gouiffès. 2020b. Experimenting the automatic recognition of non-conventionalized units in sign language. *Algorithms*, 13(12):310.

P.J. Buehler, A. Zisserman, and M. Everingham. 2009. Learning sign language by watching tv (using weakly aligned subtitles). *IEEE CVPR*, pages 2961–2968.

H. Bull. 2023. *Learning sign language from subtitles*. Ph.D. thesis. Université Paris-Saclay.

J. Carreira and A. Zisserman. 2017. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. pages 4724–4733. IEEE CVPR.

J. Devlin, M-W. Chang, K. Lee, and K. Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Conf. of the NAACL association: Human Language Technologies*, page 4171–4186. ACL.

T. Jiang, N.C. Camgoz, and R. Bowden. 2021. Looking for the Signs: Identifying Isolated Sign Instances in Continuous Video Footage. In *IEEE FG*, pages 1–8, Jodhpur, India.

M. Kopf, M. Schulder, and T. Hanke. 2022. The Sign Language Dataset Compendium: Creating an overview of digital linguistic resources. In *LREC Work. on the Repr. and Proc. of Sign Languages*, pages 102–109, Marseille, France.

D. Li, C. Xu, X. Yu, K. Zhang, B. Swift, H. Suominen, and H. Li. 2020. Tspnet: Hierarchical feature learning via temporal semantic pyramid for sign language translation. In *NeurIPS*, volume 33, pages 12034–12045.

Z. Liu, J. Ning, Y. Cao, Y. Wei, Z. Zhang, S. Lin, and H. Hu. 2022. Video swin transformer. In *CVPR*, pages 3202–3211, New Orleans, USA. IEEE.

L. Momeni, H. Bull, K. R. Prajwal, S. Albanie, G. Varol, and A. Zisserman. 2022. Automatic dense annotation ofnbsp;large-vocabulary sign language videos. In *ECCV October 23–27*, page 671–690.

L. Momeni, G. Varol, S Albanie, T. Afouras, and A. Zisserman. 2020. Watch, read and lookup: learning to spot signs from multiple supervisors.

Y. Ouakrim, D. Beautemps, M. Gouiffès, T. Hueber, F. Berthommier, and A. Braffort. 2023. A Multistream Model for Continuous Recognition of Lexical Units in French Sign Language. In *GRETSI 2023*, Grenoble, France.

Y. Ouakrim, H. Bull, M. Gouiffès, D. Beautemps, T. Hueber, and A. Braffort. 2024. Mediapi-rgb: Enabling technological breakthroughs in french sign language (lsf) research through an extensive video-text corpus. In *20th International Conference on Computer Vision Theory and Applications (VISAPP)*.

K. R. Prajwal, H. Bull, L. Momeni, S. Albanie, G. Varol, and A. Zisserman. 2022. Weakly-supervised fingerspelling recognition in british sign language videos. In *BMVC*, London, UK.

K. Renz, N. C. Stache, S. Albanie, and G. Varol. 2021. Sign language segmentation with temporal convolutional networks. In *IEEE ICASSP*, pages 2135–2139, Toronto, Canada. IEEE.

P. J. Rousseeuw. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65.

H-D. Yang, S. Sclaroff, and S-W. Lee. 2009. Sign language spotting with a threshold model based on conditional random fields. *IEEE Trans. on PAMI*, 31(7):1264–1277.

10. Language Resource References

Belissen, V. and Braffort, A. and Gouiffès, M. 2020. *Dicta-Sign-LSF corpus*. ISLRN 442-418-132-318-7.

Bull, H. and Ouakrim, Y and Lascar, J. and Braffort, A. and Gouiffès, M. 2024. *Mediapi-RGB corpus*. ISLRN 421-833-561-507-6.