Automatic Annotation and Segmentation of Sign Language Videos: Base-level Features and Lexical Signs Classification

Hussein Chaaban, Michèle Gouiffès, Annelies Braffort

Automatic Annotation and Segmentation of Sign Language Videos: Base-level Features and Lexical Signs Classification

Hussein Chaaban, Michèle Gouiffès[a] and Annelies Braffort[b]
Université Paris-Saclay, CNRS, LIMSI, 91400, Orsay, France

Keywords: LSF Videos, Annotations, Lexical Signs, Sign Segmentation.

Abstract: The automatic recognition of Sign Languages is the main focus of most of the works in the field, which explains the progressing demand on the annotated data to train the dedicated models. In this paper, we present a semi automatic annotation system for Sign Languages. Such automation will not only help to create training data but it will reduce as well the processing time and the subjectivity of manual annotations done by linguists in order to study the sign language. The system analyses hand shapes, hands speed variations, and face landmarks to annotate base level features and to separate the different signs. In a second stage, signs are classified into two types, whether they are lexical (i.e. present in a dictionary) or iconic (illustrative), using a probabilistic model. The results show that our system is partially capable of annotating automatically the video sequence with a F1 score = 0.68 for lexical sign annotation and an error of 3.8 frames for sign segmentation. An expert validation of the annotations is still needed.

1 INTRODUCTION

Sign languages (SL) are natural gestural languages that are used mainly by the deaf community. A SL discourse consists of a sequence of signs performed by the hands at a specific location and configuration, accompanied by non-manual components like facial expressions and upper body movements. These sequences of manual and non manual components make it possible to transmit information in parallel during the discourse.

Currently several studies are interested in the problem of the automatic processing of SL, and more particularly of the recognition of the lexical signs, which are only one of the type of signs present in a SL discourse. In the following, for simplification, we will use the term SL recognition for lexical sign recognition. Although a large part of the lexical signs are defined in a dictionary, there is a very large variability linked to the context during their realization. In addition, the signs are often separated by co-articulation movements (transitions). This extreme variability and the co-articulation effect represent an important problem in the automatic processing of SL research field. It is therefore necessary to have a huge number of SL videos annotated, to study linguistic and to build large data-sets for machine learning. SL are not well understood nor described formally. Linguists and researchers in Human Movement attempt to understand which movement, which gesture or which body or face components (from the eyebrows to the hands) are key in forming a given message.

Certain annotation levels of SL videos require to isolate each individual sign before describing the hands movement and the facial/body events that accompany it. To date, the temporal segmentation and motion descriptions of signs are carried out manually using annotation tools such as ANVIL (Kipp, 2001) or ELAN (Wittenburg et al., 2002)]. Though these manual annotations present many problems. In addition to the significant time required for this work, the results are extremely variable because each annotator has its own criteria to estimate the beginning and end of the sign. This leads us to question ourselves on the issue of temporal segmentation and to propose an original approach to speed up this stage of video processing to study the SL. Therefore, in this paper, we propose an automatic annotation system for face and body annotations such as mouthing, head direction, location of signs. Also, the system segments automatically signs, using hands movements without any prior learning phase. In other words, no huge manually annotated data are required for automatic annotation nor for segmentation. After that we propose an

[a] https://orcid.org/0000-0002-7152-4640
[b] https://orcid.org/0000-0003-4595-7714

Chaaban, H., Gouiffès, M. and Braffort, A.
Automatic Annotation and Segmentation of Sign Language Videos: Base-level Features and Lexical Signs Classification.
DOI: 10.5220/0010247104840491
In *Proceedings of the 16th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2021) - Volume 5: VISAPP*, pages 484-491
ISBN: 978-989-758-488-6

algorithm to classify the type of the segmented sign using the automatically generated annotations. In the next section we present some of the works done in the SL processing field. Sections 3, 4 and 5 describes our proposed method. Then, Section 6 discusses the evaluation method, the dataset on which we worked on and the results. In the last section we present our conclusions and the perspective of our work.

2 RELATED WORK

Nowadays, most of the works dedicated to the SL processing are interested in the recognition of lexical signs. Many attempts of SL recognition are conducted on isolated signs (Rastgoo et al., 2020), (Lim et al., 2019). However, in a natural speech of SL, it is often difficult to find precisely the beginning and the end of each sign because of the problems of co-articulation we mentioned before. In addition, most of these works focus on specific datasets, made in controlled environments (uniform background, signer with dark clothes) and dealing with a specific topic, such as weather (Koller et al., 2015). But the real challenge in SL recognition remains in identifying dynamic signs, *i.e.* signs in continuous SL speech, and most importantly independently of the signer (Liang et al., 2018).

In the field of SL recognition, the methods which exploit Hidden Markov Models (HMM) remain among the most used (Fatmi et al., 2017), (Wang et al., 2006). These approaches identify specific information in a signal. It is then a question of comparing the signal of a sign to a previously learned model. The recognition and segmentation processes are then carried out simultaneously. This poses a great limitation in recognizing many types of iconic structures such as the depicting signs, or classifiers (Cuxac, 2000) which cannot be learned due to their excessively large number and their variability depending on the context. Recently, with the great rise of the Convolutional Neural Networks in the machine learning field, there was no exception of deploying these networks for the SL recognition. Many researchers adopted this technique in their works as it showed high recognition rates (Rao et al., 2018)]. But the main problem of this approach is the need of enormous amount of annotated data to train the network. Unfortunately this amount of data is not easily available especially for non-American sign languages. Therefore many attempts of recognition were limited to a certain number of signs (Pigou et al., 2015).

Concerning automatic annotation of SL, some works represent a sign as a series of movements and configuration and try to annotate the segments by describing facial and body events as mouthing, gaze, occlusion, hands locations, hand shapes, and movements without any information about the signs themselves (Naert et al., 2018), (Yang et al., 2006). (Lefebvre-Albaret and Dalle, 2008) worked on the automatic segmentation and annotation of signs to provide reusable data for virtual signers animation. Their algorithm is based exclusively on the 2D movements of the hands in the video. But it requires a human operator to point, during the viewing of the video, a frame (which they call primer) so that each sign has one and only one primer. (Gonzalez Preciado, 2012) proposed a better way to segment signs using hands movements and configurations. The annotations provided are exclusively describing the hands with no information about the other non manual components.

3 FEATURES ANNOTATION

To study LS and use machine learning methods, it is necessary to have loads of annotated videos. In annotations, the linguists extract visual features from the video. Using statistics on these collected manual and non-manual features, linguistic models can be built. These annotations are done manually by linguists or SL experts. They are subject to error, depending on the SL knowledge of the annotator. Furthermore, they are non-reproducible and extremely time consuming. Thus, automating this task would be a saving of time and robustness.

The experts of SL can annotate the manual and non manual components at various levels: from low levels (face and body features and events) to high levels (structure of the discourse). In our automatic annotation, we are interested in the automatic base level features annotation. In Sign languages, these features are the lowest level units are. They are meaningless on their own. But they can be interpreted at a higher level for annotating more complex linguistic units. For example lexical level annotation can be generated using features extracted at the base level. As far now, there is no predefined list of manual and non manual components to annotate. In the literature, experts try to annotate hand shapes, their locations, motion, direction, symmetry between them, mouthing, mouth gestures, gaze and eyebrows.

3.1 Body and Face Landmarks

The first step to annotate these base level features, is extracting the different body and face parts. Many works used image segmentation techniques to locate

these articulators. Those methods usually require specific environments and tools such as uniform backgrounds, colored gloves,... In our work we use OpenPose (Cao et al., 2018) (figure.1), a recent real time body pose estimation library, which provides the coordinates of different body landmarks with high success rates and OpenFace (Baltrušaitis et al., 2018) (figure.2) a toolkit capable of facial landmark detection, head pose estimation, facial action unit recognition, and eye-gaze estimation.

Figure 1: Body landmarks using Openpose.

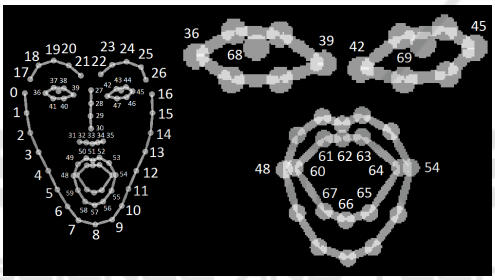


Figure 2: Face landmarks using OpenFace.

3.2 Controlled Vocabulary

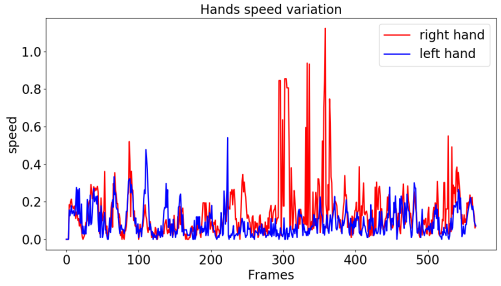
Once we get the coordinates of the different body and face landmarks, we proceed to calculate further features such as mouthing, gaze, etc. In manual annotation software, the ensemble of these features are called controlled vocabulary. Till this date, there is no one standard controlled vocabulary list defined. Each linguist annotates the events that are relevant for his studies. With our automatic annotation system, we annotate the features that are frequently annotated in the literature: mouthing, gaze/head direction, bi-manual motion, signing space and hands-head distance. To annotate mouthing, we calculated the isoperimetric ratio (or circularity) of the interior of the lips using the coordinates provided by OpenFace and the formula: $\mathrm{IR} = \frac{4\pi a}{p^2}$ where a is the area of the interior of the lips and p is its perimeter. The higher the ratio is, the more the mouth is open, which is an indication of a mouthing. The hands-head distance is simply calculated using the coordinates of both wrists $((x_{WRight}, y_{WRight}),(x_{WLeft}, y_{WLeft}))$ and nose (x_{Nose}, y_{Nose}) (center of the face). For the bi-manual motion, we measure the speed and direction of each hand $((V_{WRight},D_{WRight}),(V_{WLeft},D_{WLeft}))$. The correlation of these two information between the two hands indicates whether these latter are symmetrical or moving in an opposite motion. To determine the location of the sign in the space, we compare the abscissa of the neck (x_{Neck}) to the abscissas of both wrists (x_{WRight}, x_{WLeft}) to decide if the sign is centered or not. As for the gaze/head direction, we use OpenFace directly for the task. The calculation of these features is more detailed in our previous article (Chaaban et al., 2019).

4 TEMPORAL SIGN SEGMENTATION

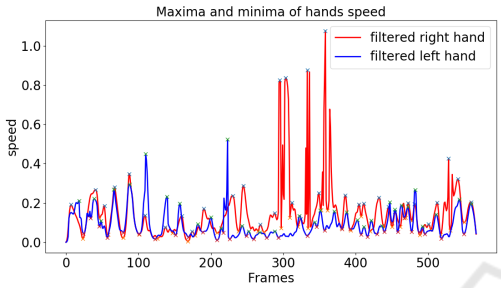
Signs can be produced sequentially, separated by transitions and pauses. In order to study the grammar of the SL or to automatically recognise them, one possible strategy consists in segmenting the video before analyzing the resulting portions of videos. The temporal segmentation of a sign means finding its boundaries, that is when the sign starts and when it ends. In the manual annotation software, the linguists segment the signs manually, which takes hours of work. In addition, the resulting segmentation is, as mentioned previously, non-reproducible since it depends on the subjectivity of the annotator and his/her experience. We propose an algorithm that segments the signs automatically using only hands motions, pauses and hand shapes. The algorithm does not need any learning phase thus annotated data are not required. For the segmentation, we use the coordinates of both wrists to calculate and draw the speed variation of each hand (see figure.3a). Then, a Wiener filter is applied to smooth-in the signal and the local maxima and minima of speed are detected (as shown by figure.3b).

When comparing the detected minima and maxima in the signal to the beginning and the end of manually annotated signs (disregarding the sign category which we will discuss later on) we can see that each sign starts with a maximum of speed of one or both hands and ends with a minimum of speed, as can be shown on Figure.4.

This observation is the basis for our temporal segmentation algorithm. To avoid the over-segmentation of signs that include repetitive movements, we rely on the hand shapes. Since (Stokoe et al., 1976), linguists describe the lexical signs using three parameters that are hands shapes, their locations and their



(a) Hands speed variation.



(b) Filtered speed with local minima and maxima.

Figure 3.

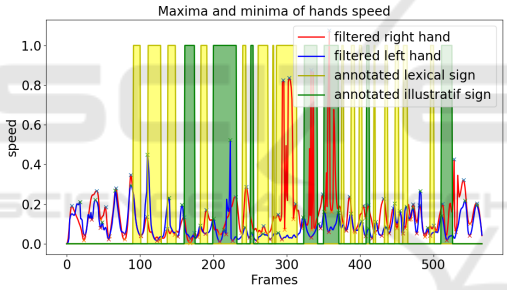


Figure 4: Annotated signs compared to the hands speed.

movements. While in some cases the hands shapes may change during a sign, they remain stable most of the time. Thus, after our initial segmentation using the hands speed, we scan the video sequence a second time and merge consecutive segmented signs of similar hand shape. These latter are mostly segments of a repeated movements forming one sign. For the hand shape detection, we use a pre-trained model that distinguish 60 different hand configurations developed by (Koller et al., 2016). Our algorithm for sign segmentation is represented in the diagram of the figure 5.

5 SIGN TYPE ANNOTATION

In section 3, we presented the base level features annotation. As we described, at that level, the annotated features are meaningless on their own. In this section,

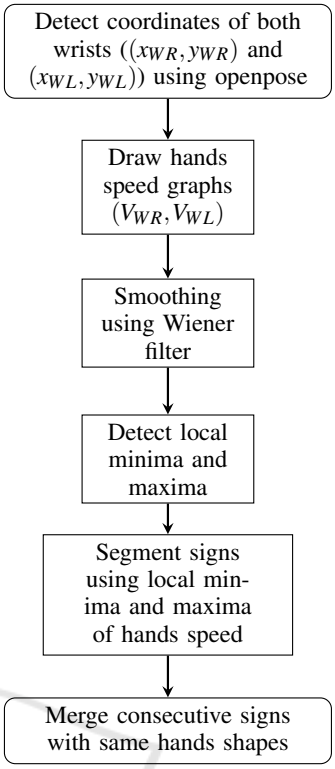


Figure 5: The sign segmentation algorithm.

we use the automatically annotated manual and non manual components to annotate a more meaningful features on a higher level: the lexical level. For this level, we do not annotate the meaning of the signs as we do not have enough data for it, but rather their types. (Cuxac, 2000) described two main types of signs: the lexical signs (LS) which are conventionalized signs that can be found in a dictionary and the highly iconic structures (HIS) (depicting signs or classifiers) which are signs that are partly conventionalized and partly context-dependant, their form changes to describe sizes, shapes, situations and role playing. With a conventionalized form, therefore easier to identify, we are more intrigued in annotating the lexical signs after the segmentation. The automatic annotation of lexical signs will definitely save time for the linguists to study the SL in comparison to manual annotation. Furthermore these annotations will accelerate the automatic recognition of signs since the dictionary of candidate signs will be reduced almost to half the size by dividing it into two classes: lexical signs and non-lexical signs.

In the literature, the automatic annotation of lexical signs is almost absent. Most of the works study the automatic recognition of signs. For that reason, we had to explore the field to discover which manual and non manual features allow us to identify the

lexical signs. We started by an analysis study for each of the features that we automatically annotated. In this study we represented the occurrence of each of the features during the production of each type of signs. The type of the signs were annotated manually by an expert on the video sequences. The resulting histograms showed that the distribution of mouthing, head direction, location of signs in the space and the hand-head distance are normal distributions with two distinct peaks for each type of signs. We drew the histograms for the 4 features. As an example, figure.6 shows the mouthing distribution between two types of signs w.r.t the isoperimetric ratio of the mouth. The other histograms (of the other features) have a similar shape, i.e. two semi separated peaks. The correlation ratio between the various features, displayed in Table.1, shows that they are not correlated, in other words independent.

Table 1: Correlation between the different features (M: Mouthing, H-D: Head Direction, Bi-M: Bimanual Motion, S-L: Sign Location, HH-D: Head-Hand Distance).

Features	M	H-D	Bi-M	S-L	HH-D
M	1	0.102	0.003	0.039	0.175
H-D	0.102	1	0.035	0.293	0.052
Bi-M	0.003	0.035	1	0.060	0.020
S-L	0.039	0.293	0.060	1	0.114
HH-D	0.175	0.052	0.020	0.114	1

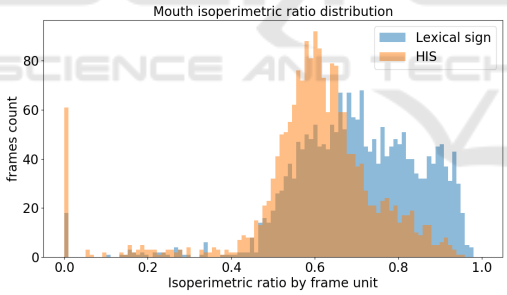


Figure 6: Histogram of mouthing occurrence (represented by the isoperimetric ratio of the mouth) during the production of lexical signs and HIS.

Using these histograms, the average μ_k and the variance σ_k^2 of each feature x_k, we built a probability distribution model (figure.7) with the equation for a normal distribution parameterized by μ_k and : σ_k^2

$$P(x = x_k \mid C) = \frac{1}{\sqrt{2\pi\sigma_k^2}} e^{-\frac{(x_i-\mu_k)^2}{2\sigma_k^2}} \quad (1)$$

Where C is the class of the sign: lexical or not.

To annotate the new frames of the new SL video sequences as lexical sign frames, the models are combined into one decision rule, as follows:

$$P(Lexical \mid F1,F2,F3,F4) = P(Lexical)\prod_{i=1}^{4} P(x_i \mid Lexical). \quad (2)$$

$$P(\overline{Lexical} \mid F1,F2,F3,F4) = P(\overline{Lexical})\prod_{i=1}^{4} P(x_i \mid \overline{Lexical}). \quad (3)$$

where $F1$ is Mouthing, $F2$ is head pose, $F3$ is sign location and $F4$ is hand-head distance (the bimanual motion feature was dropped of our study as it did not add any weight to the classification of the signs).

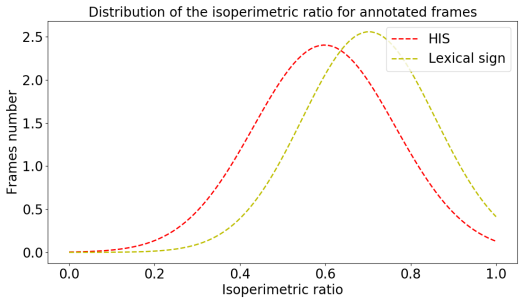


Figure 7: Probability distribution model for mouthing (using the isoperimetric ratio).

Then, we compare (2) and (3). If the result of (2) is higher than the result of (3) the new frame is part of a lexical sign, otherwise it is considered as non-lexical. After the temporal segmentation in section 4, a classifier is applied on every frame of each segmented sign. A segment is classified as lexical when the majority of frames are classified as lexical, otherwise it is considered as non-lexical.

6 EXPERIMENTAL METHOD

In this section, we present the performances of our system to automatically annotate base level features, segmentation and sign category (lexical or not). To our knowledge, there is no public dataset with similar automatic annotations, on which to compare our results. Thus, our evaluation is based on quantitative measures of the segmentation and classification of lexical signs only in comparison to a ground truth dataset manually annotated by an expert.

6.1 Dataset

The dataset used in our work is a portion of MOCAP[1] dataset: a collection of 2D RGB videos in French Sign Language. The dataset includes 49 videos with 4 different signers filmed from hip up face view and

[1]The corpus is downloadable here (for research only): https://www.ortolang.fr/market/corpora/mocap1.

equipped with motion detectors for 3D representation (figure.8) used other study purposes since this dataset was not produced specifically for our work. We note that the presence of the motion detectors did not prevent the detection of face, hands and body keypoints in our 2D RGB image study. The length of the videos

Figure 8: One frame of the MOCAP dataset.

varies from 15 to 34 seconds with an average of 24 seconds, and the dataset contains a total 19.63 min of videos. In each video, the signer was asked to describe an image that represents a scene such as a living room (figure.9) or a forest. 25 images were chosen delicately to have a variation between lexical signs and HIS. The segmentation and the annotation of sign type (Lexical and HIS) were done manually by an expert: 1011 signs were annotated, 709 were lexical signs and 304 were HIS.

Figure 9: Examples of scenes to be described by the signers.

6.2 Evaluation

As described before, our algorithm starts by annotating features such as mouthing, head direction, location of sign. Then, using hand motions, the frames are scanned to find the boundaries of each sign segment. After that, the automatically generated features and a portion of manual annotations provided by an expert are used to create a probabilistic model. The latter classifies the segmented signs into lexical and non lexical signs.

In this section, we measure the performance of the algorithm to locate the lexical signs and find its boundaries in comparison to the ground truth. Since our ground truth is based on manual annotations which are dependent on the annotators knowledge and experience, the evaluation results have to be taken with caution.

As we described in section 6.1, our dataset is divided into training data and testing data. For the evaluation, we compare the automatically generated annotation files to the rest of the manual annotations which were not used in the learning phase to create the classifier. Using the testing dataset, we count the number of true positives (TP) of the detected signs, the false positive (FP), the true negative (TN) and the false negative (FN) ones. We consider that a lexical sign is correctly detected when 3 consecutive frames classified as lexical fall in the range of the manually annotated sign (figure. 10). Then we compute the TP and TN rates (TPR and TNR), i.e. the sensitivity and the specificity, the positive prediction value (PPV) (or precision) and F1-score:

$$TPR = \frac{TP}{TP+FN} \quad TNR = \frac{TN}{TN+FP}$$

$$PPV = \frac{TP}{TP+FP} \quad F_1 = 2\frac{PPV.TPR}{PPV+TPR}$$

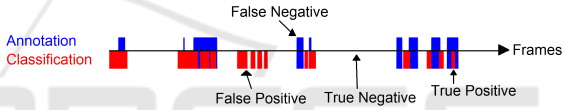


Figure 10: Counting of False/True Positives and False/True Negatives.

We calculated the same measures with frame by frame comparison were each frame from the automatic annotations is compared to the corresponding frame in the manual annotations to determine whether it's a TP, FP, TN, or FN frame to give an approximate measure on the precision of the classification in terms of segmentation. The performance of the proposed segmentation approach has been quantitatively evaluated as well. For each detected sign, we calculate the difference between the first frame of the automatically segmented lexical sign and the first frame of the manually segmented lexical sign. We do the same thing for the last frames of both automatically and manually segmented lexical signs. The mean of the sum of both differences over all the correctly detected lexical signs represents the segmentation error of our algorithm. The results of the segmentation using the hands motion alone and the hands motion corrected by the hands shapes are compared as well.

6.3 Results

The results of our method are evaluated for each signer individually and then combined to check if the classification is independent of the signer.

Intra-signer Study. For each signer, the videos and their corresponding manual annotations are divided into 3 subsets L_1, L_2 and L_3. Two of them $(L_i, L_j) = (L_1, L_2), (L_1, L3), (L_2, L_3)$ are used for learning and the last one for testing. A cross-validation is performed, by collecting the results of each experiment.

Inter-signer Study. Here, the videos are divided into 4 subsets L_1, L_2,L_3 and L_4, each subset includes all the videos from the same signer. Again we tried all the different combinations of subsets for learning and testing with three subsets for learning and one subset for testing.

Firstly, the performance measures for intra and inter-signer studies were calculated after a temporal segmentation using only the hands motion. Then we corrected the segmentation with the hands shapes and recalculated the performance measures. The averages and standard deviations of both experiments and for both studies are shown in the tables 2, 3, 4 and 5.

Table 2: Results for intra-signer classification after a segmentation based on the hands motion only. The shown values are the average of all the results coming from each signer separately.

TPR		TNR		PPV		F1 score	
μ	σ	μ	σ	μ	σ	μ	σ
0.65	0.08	0.60	0.13	0.52	0.09	0.57	0.05

Table 3: Results for intra-signer classification after a segmentation based on the hands motion and hands shapes.

TPR		TNR		PPV		F1 score	
μ	σ	μ	σ	μ	σ	μ	σ
0.78	0.05	0.62	0.12	0.65	0.10	0.70	0.04

Table 4: Results for inter-signer classification after a segmentation based on the hands motion only: averages of all the results coming from all signers combined.

TPR		TNR		PPV		F1 score	
μ	σ	μ	σ	μ	σ	μ	σ
0.70	0.11	0.51	0.14	0.45	0.11	0.53	0.05

Table 5: Results for inter-signer classification after a segmentation based on the hands motion and hands shapes.

TPR		TNR		PPV		F1 score	
μ	σ	μ	σ	μ	σ	μ	σ
0.83	0.08	0.52	0.13	0.56	0.10	0.66	0.05

The tables show that our algorithm is able at a certain level of annotating automatically the lexical signs. The similarity between the results obtained for intra-signer and for inter-signer experiments indicates that our algorithm is signer independent. When comparing the results of classification after segmentation using only hands motion and segmentation using both hands motion and shapes, we can see clearly that the use of the hands shapes to correct the segmentation was beneficial.

To evaluate the performance of our segmentation technique, we calculated the same metrics with frame by frame comparison between manual annotations and automated annotations. The results are shown in table 6. These results though do not reflect the real performance of the segmentation approach as it includes the false positive and the false negative detected signs. Therefore we isolated the true positive detected signs and measured the ability of our algorithm to detect precisely the beginning and the end of each correctly detected sign. We calculated for each sign the number of frames between the detected beginning frame and the annotated beginning frame as well as the difference between the detected end frame and the annotated end frame. The average of this differences was 3.8 frames which represents our error range. Our automatic annotation algorithm is not perfect, and we cannot pretend that our annotations can be used directly for other linguistic studies. An expert validation for the automatic annotations is needed after every processing. Thus our work can be considered as an assistance tool which can with no doubt simplify the annotation of face/body features and lexical signs for the linguists and reduce the time spent for the process.

Table 6: Frame by frame evaluation. The shown values are the average of all the results coming from each sign and all signers combined after a segmentation using hands motion and shapes.

TPR		TNR		PPV		F1 score	
μ	σ	μ	σ	μ	σ	μ	σ
0.62	0.02	0.55	0.04	0.29	0.02	0.39	0.01

The limitations of our algorithm are due to several factors starting with the particularity of the sign language and its grammar, as some lexical signs may change form depending on the context and some were created from the same repeatedly used HIS. Another factor is the subjectivity of the annotations used for training which makes our results subjective as well. Some classification errors were also due to the errors in the detection of facial and body landmarks.

7 CONCLUSIONS

This paper has proposed a tool that will be useful for linguists to pre-annotate Sign Language videos. We started by annotating face and body features such as mouthing, sign location, etc. Then we detailed a temporal segmentation of lexical signs in video se-

quences. The segmentation was carried out by considering only the information coming from the hands, which makes our method applicable to all sign languages. We first used the hands motion calculated using body landmarks provided by OpenPose. Then we corrected the segmentation using the hands shapes. Once we got all signs isolated individually, we built a probabilistic model using a portion of MOCAP dataset which was annotated manually by an expert. The model was used as a classifier to distinguish lexical and non lexical signs. To evaluate our algorithm of classification we used the sensitivity, the specificity, the precision and the F1 score metrics. The results showed that our algorithm was capable of detecting the lexical signs with a F1 score = 0.68 and that the use of hands shapes for segmentation improved the detection (F1 score improved by 0.13). To evaluate the segmentation approach we calculate the average difference between the beginning and end frames of annotated and detected signs (3.8 frames). In the future, we will try to refine both of the segmentation and the classification results by including more features that could be useful for the task. On a parallel axis, we will use the set of the annotated features to create sub-categories of signs based on the similarities between features. Such categorisation would accelerate the process of automatic recognition, make it more efficient.

REFERENCES

Baltrušaitis, T., Zadeh, A., Chong Lim, Y., and Morency, L.-P. (2018). Openface 2.0: Facial behavior analysis toolkit. *IEEE International Conference on Automatic Face and Gesture Recognition*.

Cao, Z., Hidalgo, G., Simon, T., Wei, S.-E., and Sheikh, Y. (2018). OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields. *arXiv preprint arXiv:1812.08008*.

Chaaban, H., Gouiffès, M., and Braffort, A. (2019). *Towards an Automatic Annotation of French Sign Language Videos: Detection of Lexical Signs*, pages 402–412.

Cuxac, C. (2000). La langue des signes française (lsf). les voies de l'iconicité. *Bibliothèque de Faits de Langues*, (15-16). Paris: Ophrys.

Fatmi, R., Rashad, S., Integlia, R., and Hutchison, G. (2017). American sign language recognition using hidden markov models and wearable motion sensors.

Gonzalez Preciado, M. (2012). *Computer Vision Methods for Unconstrained Gesture Recognition in the Context of Sign Language Annotation*. Theses, Université Paul Sabatier - Toulouse III.

Kipp, M. (2001). Anvil - a generic annotation tool for multimodal dialogue. *INTERSPEECH*.

Koller, O., Forster, J., and Ney, H. (2015). Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers. *Computer Vision and Image Understanding 141*, pages 108–125.

Koller, O., Ney, H., and Bowden, R. (2016). Deep hand: How to train a cnn on 1 million hand images when your data is continuous and weakly labelled. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3793–3802, Las Vegas, NV, USA.

Lefebvre-Albaret, F. and Dalle, P. (2008). Une approche de segmentation de la langue des signes française.

Liang, Z.-j., Liao, S.-b., and Hu, B.-z. (2018). 3D Convolutional Neural Networks for Dynamic Sign Language Recognition. *The Computer Journal*, 61(11):1724–1736.

Lim, K., Tan, A., Lee, C.-P., and Tan, S. (2019). Isolated sign language recognition using convolutional neural network hand modelling and hand energy image. *Multimedia Tools and Applications*, 78.

Naert, L., Reverdy, C., Caroline, L., and Gibet, S. (2018). Per channel automatic annotation of sign language motion capture data. *Workshop on the Representation and Processing of Sign Languages: Involving the Language Community, LREC*. Miyazaki Japan.

Pigou, L., Dieleman, S., Kindermans, P.-J., and Schrauwen, B. (2015). Sign language recognition using convolutional neural networks. volume 8925, pages 572–578.

Rao, G., Syamala, K., Kishore, P., and Sastry, A. (2018). Deep convolutional neural networks for sign language recognition. pages 194–197.

Rastgoo, R., Kiani, K., and Escalera, S. (2020). Video-based isolated hand sign language recognition using a deep cascaded model. *Multimedia Tools and Applications*.

Stokoe, W., Casterline, D., and Croneberg, C. (1976). A dictionary of american sign language on linguistic principles (revised ed.). [Silver Spring, Md.]: Linstok Press.

Wang, H., Leu, M., and Oz, C. (2006). American sign language recognition using multidimensional hidden markov models. *Journal of Information Science and Engineering - JISE*, 22:1109–1123.

Wittenburg, P., Levinson, S., Kita, S., and Brugman, H. (2002). Multimodal annotations in gesture and sign language studies. *LREC*.

Yang, R., Sarkar, S., Loeding, B., and Karshmer, A. (2006). Efficient generation of large amounts of training data for sign language recognition: A semi-automatic tool. pages 635–642.