Optimization of a gesture representation network for Sign Language analysis

Valentin Belissen, Michèle Gouiffes, Annelies Braffort

Optimization of a gesture representation network for Sign Language analysis

Valentin Belissen
LIMSI, CNRS, Université Paris Saclay, Orsay,France

Michèle Gouiffès
LIMSI, CNRS, Orsay,France

Annelies Braffort
LIMSI, CNRS, Orsay,France

Abstract

This paper presents the manufacturing and optimization of a convolutional-recurrent neural network, in order to jointly learn the detection of numerous Sign Language linguistic features in ordinary RGB videos.

The proposed architecture can learn generic temporal-gestural features from a compact representation of people producing *continuous* Sign Language. These generic features make it possible to detect both lexical signs and higher-level linguistic patterns simultaneously. New pattern types can be added to the model and accurately detected without retraining the gestural features, that is with few training instances.

The network is trained and tested on a continuous dialog corpus of French Sign Language. It gets localized F1-scores up to 80%, depending on the optimization of the network architecture.

1 Introduction

As a natural language for the Deaf, Sign Languages (SL) are still to be thoroughly described and understood in terms of linguistics [2]. They make use of lexicon as well as more complex linguistic structures, such as pointing signs or classifiers [25]. To the authors' knowledge, these high-level structures are usually ignored in the automatic Sign Language Recognition (SLR) literature, even though they play a crucial role in SL production. As more and more SL material is available on public platforms, social networks etc., the ability to automatically detect high-level linguistic features as well as specific lexicon would be of great assistance to the linguistic community, as there is a strong demand from language experts.

In this paper, we present two major contributions to the field of continuous SLR:

1. The manufacturing of a convolutional-recurrent feedforward neural network that uses a simple and generic modeling of a signer, and that can output many different types of SL linguistic predictions.

2. A general temporal-gestural representation that can be used to learn the detection of rare linguistic events.

This paper is organized as follows: in Section 2, state-of-the-art in continuous SLR is presented and limitations are discussed. A generalizable modeling of a signer is introduced in Section 3. Then, the gesture representation network that we propose is presented and detailed in Section 4. Subsequently, experiments are conducted and their results are summarized in Section 5. Finally, we end with conclusions and future work in Section 6.

2 Related work and limitations

In this section, we start with a brief description of SL linguistics and associated hypotheses. We then reflect on usual SL automatic analysis architectures and their potential drawbacks.

Frame	Comment
1	Lexical sign "Paris"
2, 3	Lexical sign "Eiffel Tower"
4	The right hand produces the lexical sign/verb "Can". Although it is usually produced as a two-handed sign, the left hand is already being used as a *fragment buoy* – here a fragment of the tower –, which is a non-lexical SL function helping the interlocutor understand that what is being said still relates to the same scene.
5, 6	The right hand has a typical hand shape for a person – known as a *proform* – and the straight motion from the bottom to the middle of the tower indicates the action of using the elevator. The left hand is still used as a fragment buoy.
7	Pointing sign to a precise location, at the middle of the tower. It indicates the location of what is going to be introduced.
8, 9	The two hands are used to depict an outer shape. Its base is a square, and it is rather slim – which is stressed by the crinkled eyes.
10	Lexical sign "Restaurant". The location and shape that have just been described thus apply to it.
11	Lexical sign "Good quality". Even though this sign can be found in dictionaries, its subjective nature is such that a great variability in terms of implication of the signer is observed, related to the continuous appreciation from *plain good* to *outstanding*.
12	Lexical sign "Also/In addition". It is almost always produced with two hands, but here a slight deformation is observed on the left hand: is it actually back to the fragment buoy referring to the Eiffel Tower.
13, 14	The hands configuration is the same as in frame 6, but now climbing to the top of the tower, then the right hand produces the lexical sign "To look/To observe" while the left hand keeps its fragment buoy function. What is interesting is the body and head posture: the shoulders and head are slightly lifted, the chin up, and she is clearly staring into the distance. She is actually executing a *role play*, or *role shift*, showing the interlocutor that someone at the top of the Eiffel Tower will then have a good perspective.

Figure 1: French Sign Language sequence example (duration: 4 seconds). The general topic is about traveling. Possible translation: *In Paris, if you climb the Eiffel Tower, you will find a great restaurant at the middle floor. Also, you can see very far from the top.*

2.1 Linguistics of Sign Languages

2.1.1 Fundamental hypotheses and SL complexity

A lot of past and current work has focused on recognizing lexical signs[1] that are realized in an isolated way, usually called *citation-form* lexical SLR [18, 32, 33, 38, 8]. Since signs are not achieved similarly in continuous discourse compared to when isolated, continuous SLR is actually a much more challenging task, with a lot of variability and continuous transitions between successive signs. It has only been addressed from a lexical perspective and with the – usually unstated – hypothesis that SL production can be reduced to a sequence of lexical signs [22, 5, 9].

However, SL actually have – at least – three strong characteristics that make them fundamentally different from unidimensional sequential languages:

1. They are **multi-channel**: information is conveyed through hand motion, shape and orientation, body posture and motion, facial expression and gaze;
2. They are strongly **spatially organized**: events, objects, people and other entities are placed in the signing space and related to each other in a visual way. The grammar of SL is structured by the use of space;
3. They allow signers to generate new signs – that would not appear in a dictionary – on the go, in an iconic way, or even modify lexical signs. More generally, **SL do not only consist of lexical signs** but they also make use of more complex iconic structures.

In the authors' opinion, the above-mentioned papers actually deal with continuous *lexical sign* recognition, while continuous *sign language* recognition is still to be addressed. We highlight the fact that lexical signs only account for a portion of SL production. This can be seen on the random example of Fig. 1. This example shows that SL production should by all means not be seen and analyzed as a succession of citation-form lexical signs. Many other gestural units are used in SL, and this research intends to show that some of them can be dealt with.

2.2 Generalizability and model architecture

2.2.1 Input data

A lot of SL recording and subsequent analysis has been done in controlled environments, with specific conditions, like RGB-D setup [31, 33, 38, 43] or very high recording frame rate [13, 12, 42]. In a less controlled environment with more general conditions (RGB images and 25 frames per second), a lot of research has been conducted on the *RWTH-PHOENIX-Weather 2014* dataset [19]. The usual approach on this corpus has been to start with a Convolutional Neural Network in order to derive features on the images [10, 23, 43, 4, 20]. The features derived from a CNN might be prone to a lack of generalizability, for instance if applied to videos where scale or appearance are changed. The signer modeling presented in Section 3 is a direct result of this discussion.

2.2.2 Model architecture

As just stated, most recent architectures dealing with continuous SL recognition use CNNs as a preprocessing layer. Recurrent Neural Networks are usually used [10, 43, 9, 22], with a Hidden Markov Model (HMM) embedding [10, 23, 21, 40]. This common architecture makes learning rather time-consuming and requires a lot of data – the CNN features must be trained with SL data. Furthermore, it is not possible to add new features to the model without retraining it altogether. In Section 4, we present a RNN network that directly outputs linguistic features probabilities, and learns gestural features.

Section 3 details a modeling of a signer that is both generalizable and compact.

3 Generalizable signer modeling

Modeling a signer in a SL video in a generalizable way is a challenging task. In this section, we briefly present the choices that were made and the final modeling that is used in this paper. It was decided to make use of several open-source programs. The pre-computed input vector that results from the modeling presented thereafter is of size $N_f^{in} = 420$.

[1] From [17], we use this commonly accepted definition: *fully-lexical signs are highly conventionalised signs in both form and meaning in the sense that both are relatively stable or consistent across contexts. Fully-lexical signs can easily be listed in a dictionary.* Lexical signs can be used as verbs, nouns, adjectives, etc.

3.1 Body pose

Convolutional Neural Networks (CNN) have emerged as a very effective tool to get relevant features from images. OpenPose [6, 34, 39] is a powerful open source library, with real-time capability for estimating 2D body pose. A 2D to 3D model was then trained on motion capture data from the French Sign Language (LSF) corpus MOCAP1 [26], only on upper body pose, following [44]. It is to be noted that instead of using raw 3D upper body pose estimation, we compute handcrafted features: every joint angle, orientation and their dynamics (speed and acceleration); every joint position relative to the parent joint – for instance left elbow relative to left shoulder), and their dynamics; relative position, speed, acceleration and distance of one hand to the other.

3.2 Facial landmarks

A 3D face estimate is directly obtained from video frames thanks to a CNN model trained on $230,000$ images [3]. As for body pose, we compute handcrafted features instead of raw data: each rotation angle, speed and acceleration for axes X, Y and Z of the centroid of the head; horizontal and vertical mouth openness; relative motion of the eyebrows to the eyes.

3.3 Hand modeling

While 3D hand pose estimation [41, 15, 35] on real-life RGB videos has not appeared to be reliable enough to this day, a SL-specific model was developed in [20]. This CNN model classifies cropped hand images into 61 predefined hand shapes classes. Whereas this model focuses on hand shape – which only accounts for a portion of the information conveyed through the hands –, we decided to use it as our hand modeling system. Thus, for each frame and each hand, we scale hand data down to a vector of 61 probabilities.

Our goal is to be able to detect any type of lexical or non-lexical feature in natural continuous SL like the Dicta-Sign corpus. For that purpose, we present further below a recurrent architecture that was built, and the results it achieved.

4 Manufacturing and training a generalizable gesture representation network

In this section, we focus on manufacturing a recurrent neural network that uses the signer modeling presented in Section 3. This network must be able to learn a generalizable gesture representation, in order to jointly learn the detection of different SL features.

4.1 Network architecture and parameters

With time $t = 1 \dots T$, the input and output sequences are defined as follows: x_t as a flattened input vector, its size N_f^{in} corresponding to the total number of pre-computed motion features (computed from body pose and facial landmarks) and hand features (see Section 3); y_t as the output of the model – it consists of N_f^{out} predictions, one for each output type of the model. The model then learns the conditional probabilities $f_t^j\left((x_t)_{t=1\dots T}\right) = \mathbb{P}\left(y_t^j \middle| (x_t)_{t=1\dots T}\right)$.
The full architecture of the model is presented on Fig. 2:

- A 1D-convolutional layer is first applied on the input x_t, which is a common operation on temporal data and action recognition, see for instance [24]. Its parameters are the kernel size N_k^c, the stride length of the convolution N_s^c and the number of filters N_f^c.
  $N_k^c = 3$ on Fig. 2.
- A first attention layer is applied on the convolved input. Attention enables the network to focus on relevant parts of input sequences [27]. The attention can be *feature-independent* when only one weight per time step is used, or *feature-dependent* if one weight per feature of each time step is calculated.
- Recurrent layers are then added to the network. Since this work does not target real-time applications, the recurrent layers have been set as bidirectional. Long Short-Term Memory (LSTM) are preferred, as they tackle vanishing gradient issues [14], which in the case of high frequency data like video frames is critical. The total number of LSTM layers is N_l^{LSTM}, and the number of hidden units of each LSTM layer is $N_h^{LSTM_i}$.
  $N_l^{LSTM} = 2$ on Fig. 2.

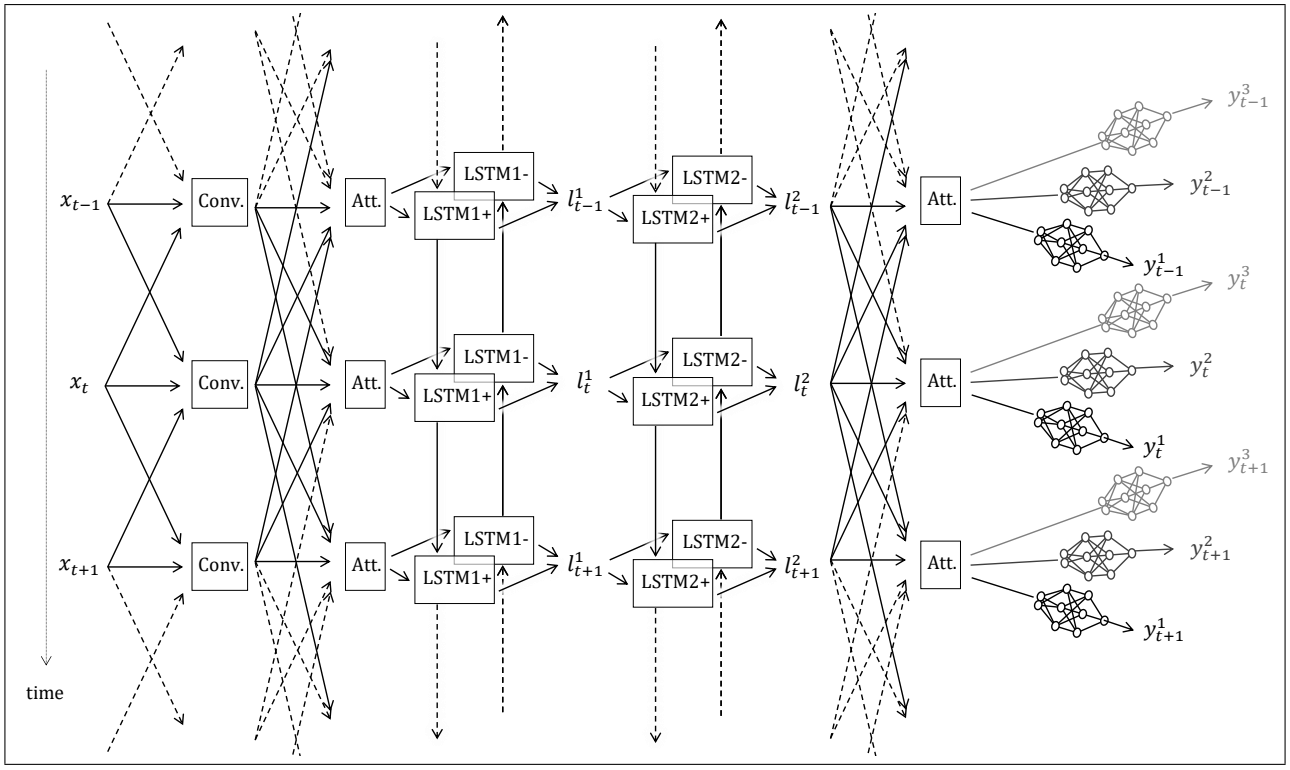

Figure 2: Simplified unfolded scheme of a RNN network for SL linguistic features learning. From left to right, the followings layers are stacked: input, 1D-convolution, attention, bidirectional LSTM (2×), attention, multi-layer perceptron before each output feature;

- A second attention layer is used, either feature-independent or feature-dependent.
- Last, before each predicted output y^i is calculated, a multilayer perceptron (MLP) is connected to the output of the second attention layer. The parameters are the number of layers N_l^{MLP} and the number of hidden neurons N_h^{MLP}.
  $N_f^{out} = 3$ on Fig. 2.

Dropout is used to prevent overfitting in the RNN and MLP layers [36]. Attention and output layers use softmax activation, while all other layers use Rectified Linear Unit activation [30]. RMSProp optimizer is used [37], and the whole model was built with Keras [7] on top of Tensorflow [1]. The architecture optimization is discussed in Section 5.4.

4.2 Specifics of this model

As stated previously, this model is aimed at jointly learning different SL features. As a matter of fact, all these features share the same layers, apart from the last softmax classifier and possibly the MLP layers of each feature. That is, the output of the last LSTM layer – l_t^2 on Fig. 2 – can be read as a general gestural-temporal representation.

If the network is already trained for the detection of N_f^{out} features, it is thus possible to add a $(N_f^{out}+1)^{th}$ feature to the network and only train one classifier, from the pre-trained gestural-temporal representation. This is examined in Section 5.5.

5 Experiments

In this section, we present the experiments that we conducted on the Dicta-Sign corpus, in order to evaluate the relevance of the learning architecture. The features we tested are detailed, and the performance measure is discussed.

5.1 Dicta-Sign: a relevant SL dataset

A number of criteria were taken into account in order to pick an appropriate dataset to train and test our model:

- French Sign Language (LSF) was preferred. Indeed, most research has focused on American [12, 43, 42, 29], German [10, 5, 23], Greek [8] and Chinese [33, 38] SL. Our goal is to draw inspiration from these works, while producing a general enough model that could be applied to other Sign Languages than LSF;
- A continuous SL recording, with the lowest possible restriction in terms of language, in order to get realistic data;
- The video resolution and frequency was to be relatively low so that our model could be applied in most use cases;
- We wanted a corpus that was not annotated only on the lexical level.

The SL corpus of Dicta-Sign contains dialogs in four different Sign Languages, including French Sign Language (LSF) [28]. Fig. 3 shows the setup of dialog recorded for the corpus. For this work, only the French part was retained, containing about five hours of annotated dialogs from 16 native signers with the following annotations according to [16]:

- *Fully Lexical Signs (FLS)*: see footnote 1 on page 3.
- *Partially Lexical Signs (PLS)*: they are also referred to as classifier signs or classifier predicates (see [25]). Their definition is close to what is called *iconic signs* in [11]. They include *Depicting Signs (DS)*, *Pointing signs (PT)* and *Fragment buoys*.
- *Non Lexical Signs (NLS)*: Here NLS comprise finger-spelling (FS) and numbering (N).

The image resolution of this corpus is 480p, while the frame rate is 25 *fps*. Furthermore, the environment is loosely controlled, consequently different narrative and signing styles were observed.

Figure 3: Dicta-Sign corpus [28]: setup for the recording of two signers who are facing each other. Two front views are recorded along with a side view.

Convol.	Attention	LSTM	MLPs	FLS	PT	PT1	PT2	N	N_{epochs}
✗	✗	1	✗	0.48	0.55	0.71	0.30	0.75	250
✓	✗	1	✗	0.53	0.57	0.70	0.31	0.78	300
✗	✓	1	✗	0.53	0.60	0.78	0.31	0.80	2000
✗	✗	2	✗	0.50	0.60	0.75	0.33	0.78	350
✗	✗	1	✓	0.50	0.56	0.72	0.31	0.76	800
✓	✓	2	✓	0.57	0.61	0.80	0.35	0.81	2500

Table 1: Localized F1-score for 5 different linguistic outputs, with different network options. The number of epochs until convergence is indicated in the last column.

5.2 Set of tested features

In these experiments, we built a network to jointly detect 5 lexical signs (FLS) – "Also/Same", "Wire/Cable", "Yes", "No", "Center/Middle" –, Pointing signs (PT), Pointing signs to the 1^{st} person (PT1) and the 2^{nd} person (PT2), and Numbering (N) – for instance dates.

The detection of these features could help build a global model aimed at understanding a SL discourse – conversely to existing models that focus on lexicon only.

5.3 Performance measure

Although the models we trained output frame-wise predictions, most frame labels are "blank", so that frame-wise accuracy is close to 100%. For sake of clarity, it was then decided to present, for each output feature y^j, localized precision P^j, recall R^j and $F1^j$-score, defined as $F1^j = 2P^jR^j/(P^j + R^j)$.

Localized precision and recall are computed from the evaluation of *true positives*, *true negatives* and *false positives* within a time window of 1 second. We thought this value was large enough so that the indefinite nature of the beginning and end of a sign is not an issue, and small enough so that localization is acceptable.

Last, we highlight the fact that the training, validation and test sets are signer-independent, which is known to make learning more difficult, as stated in [22][2].

Trained layers	Localized F1-score
Whole network	0.78
Only last classifier	0.26
Last classifier and a LSTM layer	0.74

Table 2: Localized F1-score for the "Numbering" output. The whole network is either trained on several features including "Numbering", or it is trained on the other features, then frozen, then used to generate gestural features that subsequently only train a classifier/a classifier preceded by a LSTM layer.

5.4 Architecture optimization

Network optimization is conducted and the results are presented in Table 1. Convolution[3], Attention[4], depth of LSTM layers[5] and the influence of adding MLP layers[6] before each output is examined.

Performance is increased with the presence of convolution, attention, additional LSTM layers and additional MLPs. However, one can note that training is much longer with attention, while the gain of performance is limited. Conversely, convolution does not lengthen training but still increases performance.

5.5 Addition of new features after training

In Section 4.2, we indicated that this network should learn general temporal-gestural features, since different types of linguistic features are connected to the output of the last LSTM layer.

In order to test this ability, we trained the network without the "Numbering" feature output, then we froze all the weights of the network and connected a "Numbering" output (softmax classifier) on top of the LSTM. We then trained this classifier only – keeping all the other weights frozen. The results of this experiment are given in Table 2. The configuration that was used for this experiment was: Convolution, no Attention, one LSTM layer, no MLP before the feature outputs. We also tested the same configuration with a LSTM layer added along with the softmax classifier. As can be seen in Table 2, the configuration including a LSTM and a classifier enables to accurately detect "Numbering" signs. However, more exploration is still needed to optimize this learning transfer.

6 Conclusions and perspectives

In this paper, we have experimented the joint learning and detection of several Sign Language linguistic features on a continuous French Sign Language corpus. With a generalizable signer modeling as input, a gesture representation network was built and optimized. This network is convolutional and recurrent, and directly outputs linguistic features probabilities. These features can be lexical or non-lexical, and we have demonstrated that new features can be added to the model, even after the gestural representation has been learned.

Future works include a better signer modeling – especially focusing on the hands – and a thorough analysis of false positives and false negatives, in order to further optimize the model. Transferring learning is also a way to explore further, since it allows the detection of rare linguistic events. Last, we intend to test this model on another corpus, in order to verify its generalizability.

References

[1] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: a system for large-scale machine learning. In *OSDI*, volume 16, pages 265–283, 2016.

[2] A. Braffort. *La Langue des Signes Française (LSF): Modélisations, Ressources et Applications.* Collection Sciences cognitives. ISTE/Hermes Science Publishing, 2016.

[2] From [22]: *the signer independent setting poses a much more difficult problem*

[3] $N_k^c = 3$, $N_s^c = 1$, $N_f^c = 200$

[4] Feature-independent

[5] $N_h^{LSTM_i} = 50$

[6] $N_l^{MLP} = 2$, $N_h^{MLP} = 30$

[3] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2D & 3D Face Alignment problem? (and a dataset of 230,000 3D facial landmarks). In *International Conference on Computer Vision*, 2017.

[4] Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, and Richard Bowden. Subunets: End-to-end hand shape and continuous sign language recognition. In *IEEE International Conference on Computer Vision (ICCV)*, 2017.

[5] Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. Neural Sign Language Translation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[6] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. In *CVPR*, 2017.

[7] François Chollet et al. Keras, 2015.

[8] Helen Cooper, Eng-Jon Ong, Nicolas Pugeault, and Richard Bowden. Sign language recognition using sub-units. *Journal of Machine Learning Research*, 13(Jul):2205–2231, 2012.

[9] Runpeng Cui, Hu Liu, and Changshui Zhang. Recurrent Convolutional Neural Networks for Continuous Sign Language Recognition by Staged Optimization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7361–7369, 2017.

[10] Runpeng Cui, Hu Liu, and Changshui Zhang. A deep neural framework for continuous sign language recognition by iterative training. *IEEE Transactions on Multimedia*, 2019.

[11] Christian Cuxac. *La langue des signes française (LSF): les voies de l'iconicité*. Number 15-16. Ophrys, 2000.

[12] Mark Dilsizian, Dimitris Metaxas, and Carol Neidle. Linguistically-driven Framework for Computationally Efficient and Scalable Sign Recognition. *LREC*, 2018.

[13] Mark Dilsizian, Zhiqiang Tang, Dimitris Metaxas, Matt Huenerfauth, and Carol Neidle. The Importance of 3D Motion Trajectories for Computer-based Sign Recognition. *7th Workshop on the Representation and Processing of Sign Languages: Corpus Mining, The 10th International Conference on Language Resources and Evaluation (LREC 2016)*, 2016.

[14] Felix A Gers, Jürgen Schmidhuber, and Fred Cummins. Learning to forget: Continual prediction with LSTM. *IET*, 1999.

[15] Umar Iqbal, Pavlo Molchanov, Thomas Breuel, Juergen Gall, and Jan Kautz. Hand Pose Estimation via Latent 2.5 D Heatmap Regression. *arXiv preprint arXiv:1804.09534*, 2018.

[16] Trevor Johnston and L De Beuzeville. Auslan corpus annotation guidelines. *Centre for Language Sciences, Department of Linguistics, Macquarie University*, 2014.

[17] Trevor Johnston and Adam Schembri. *Australian Sign Language (Auslan): An introduction to sign language linguistics*. Cambridge University Press, 2007.

[18] Hamid Reza Vaezi Joze and Oscar Koller. MS-ASL: A Large-Scale Data Set and Benchmark for Understanding American Sign Language. *arXiv preprint arXiv:1812.01053*, 2018.

[19] Oscar Koller, Jens Forster, and Hermann Ney. Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers. *Computer Vision and Image Understanding*, 141:108–125, December 2015.

[20] Oscar Koller, Hermann Ney, and Richard Bowden. Deep Hand: How to Train a CNN on 1 Million Hand Images When Your Data is Continuous and Weakly Labelled. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3793–3802, June 2016.

[21] Oscar Koller, O Zargaran, Hermann Ney, and Richard Bowden. Deep Sign: Hybrid CNN-HMM for Continuous Sign Language Recognition. In *Proceedings of the British Machine Vision Conference 2016*, September 2016.

[22] Oscar Koller, Sepehr Zargaran, and Hermann Ney. Re-Sign: Re-Aligned End-to-End Sequence Modelling with Deep Recurrent CNN-HMMs. In *IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, HI, USA, July 2017.

[23] Oscar Koller, Sepehr Zargaran, Hermann Ney, and Richard Bowden. Deep Sign: Enabling Robust Statistical Continuous Sign Language Recognition via Hybrid CNN-HMMs. *International Journal of Computer Vision*, 126(12):1311–1325, 2018.

[24] Colin Lea, Rene Vidal, Austin Reiter, and Gregory D Hager. Temporal convolutional networks: A unified approach to action segmentation. In *European Conference on Computer Vision*, pages 47–54. Springer, 2016.

[25] Scott K Liddell. *An investigation into the syntactic structure of American Sign Language*. University of California, San Diego, 1977.

[26] Limsi and CIAMS. MOCAP1, 2017. ORTOLANG (Open Resources and TOols for LANGuage) –www.ortolang.fr.

[27] Thang Luong, Hieu Pham, and Christopher D. Manning. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbon, Portugal, September 2015. Association for Computational Linguistics.

[28] S. Matthes, T. Hanke, Anja Regen, J. Storz, Satu Worseck, E. Efthimiou, N. Dimou, Annelies Braffort, J. Glauert, and E. Safar. Dicta-Sign – Building a Multilingual Sign Language Corpus. In , pages 117–122, Istanbul, Turkey, 2012.

[29] Dimitris N Metaxas, Bo Liu, Fei Yang, Peng Yang, Nicholas Michael, and Carol Neidle. Recognition of Nonmanual Markers in American Sign Language (ASL) Using Non-Parametric Adaptive 2D-3D Face Tracking. In *LREC*, pages 2414–2420, 2012.

[30] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814, 2010.

[31] Lionel Pigou, Sander Dieleman, Pieter-Jan Kindermans, and Benjamin Schrauwen. Sign language recognition using convolutional neural networks. In *Workshop at the European Conference on Computer Vision*, pages 572–578. Springer, 2014.

[32] Lionel Pigou, Aäron Van Den Oord, Sander Dieleman, Mieke Van Herreweghe, and Joni Dambre. Beyond temporal pooling: Recurrence and temporal convolutions for gesture recognition in video. *International Journal of Computer Vision*, 126(2-4):430–439, 2018.

[33] Junfu Pu, Wengang Zhou, Jihai Zhang, and Houqiang Li. Sign language recognition based on trajectory modeling with HMMs. In *International Conference on Multimedia Modeling*, pages 686–697. Springer, 2016.

[34] Tomas Simon, Hanbyul Joo, Iain Matthews, and Yaser Sheikh. Hand Keypoint Detection in Single Images using Multiview Bootstrapping. *arXiv preprint arXiv:1704.07809*, 2017.

[35] Adrian Spurr, Jie Song, Seonwook Park, and Otmar Hilliges. Cross-modal Deep Variational Hand Pose Estimation. In *CVPR*, 2018.

[36] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *J. Mach. Learn. Res.*, 15(1):1929–1958, January 2014.

[37] T. Tieleman and G. Hinton. Lecture 6.5—RmsProp: Divide the gradient by a running average of its recent magnitude. COURSERA: Neural Networks for Machine Learning, 2012.

[38] Hanjie Wang, Xiujuan Chai, Xiaopeng Hong, Guoying Zhao, and Xilin Chen. Isolated Sign Language Recognition with Grassmann Covariance Matrices. *ACM Transactions on Accessible Computing (TACCESS)*, 8(4):14, 2016.

[39] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4724–4732, 2016.

[40] Di Wu, Lionel Pigou, Pieter-Jan Kindermans, Nam Do-Hoang Le, Ling Shao, Joni Dambre, and Jean-Marc Odobez. Deep Dynamic Neural Networks for Multimodal Gesture Segmentation and Recognition. *IEEE transactions on pattern analysis and machine intelligence*, 38(8):1583–1597, 2016.

[41] Donglai Xiang, Hanbyul Joo, and Yaser Sheikh. Monocular Total Capture: Posing Face, Body, and Hands in the Wild. *arXiv preprint arXiv:1812.01598*, 2018. https://github.com/xiangdonglai Pas de modèle dispo pour l'instant https://www.youtube.com/watch?v=-7rQSPYZRNw.

[42] Hee-Deok Yang and Seong-Whan Lee. Robust sign language recognition by combining manual and non-manual features based on conditional random field and support vector machine. *Pattern Recognition Letters*, 34(16):2051–2056, 2013.

[43] Yuancheng Ye, Yingli Tian, Matt Huenerfauth, and Jingya Liu. Recognizing American Sign Language Gestures from within Continuous Videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 2064–2073, 2018.

[44] Ruiqi Zhao, Yan Wang, and Aleix M Martinez. A simple, fast and highly-accurate algorithm to recover 3d shape from 2d landmarks on a single image. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):3059–3066, 2018.