Sign Language Video Analysis For Automatic Recognition and Detection

Valentin Belissen

Sign Language Video Analysis For Automatic Recognition and Detection

Valentin Belissen
LIMSI, CNRS, Université Paris Saclay
Orsay, 91400, France
+33 (0)1 69 15 58 02
valentin.belissen@limsi.fr

ABSTRACT

UPDATED—February 13, 2019. This Research Project Summary presents ongoing work on automatic Sign Language Recognition (SLR) and detection, carried out as a PhD research project at the LIMSI, CNRS, France.

This work tackles the issue of extracting relevant information from a Sign Language (SL) video, and developing systems to detect lexical signs or even high-level linguistic features. It will benefit the Deaf community, because it can help indexing or searching through SL videos, as well as querying lexical signs databases.

CCS Concepts

•Human-centered computing → Accessibility technologies; Accessibility systems and tools; Gestural input; •Applied computing → Language translation; •Social and professional topics → People with disabilities;

Author Keywords

Sign Language Recognition; Lexical Signs; Video Analysis; Deep Learning; Recurrent Neural Networks; Natural Language Processing.

THESIS RESEARCH SUMMARY

Addressed problem and motivation

Sign Languages (SL) are undoubtedly the most natural language for Deaf people. One characteristic though, is that SL do not have a written equivalent, like commonly used languages do, which makes it impossible to search amongst SL videos, like one would do for text documents. However, with the development of the Internet, there is more and more SL videos available to the Deaf, on public platforms, social networks, etc. It is thus a strong need inside the Deaf community to develop tools that would enable one to do such searches or queries, possibly without the use of any written language.

ACM ASSETS'18, Galway, Ireland

DOI: http://dx.doi.org/10.475/123_4

A lot of past and current work has focused on the problem of recognizing signs that are realized in an isolated way, usually called *citation-form* lexical SLR [7, 8, 15, 17, 19]. Although it might be a step towards true SLR, it has strong limitations, if only that signs are not achieved similarly in real talks compared to when isolated. In real talks, SLR is a much more difficult task, with more variability and continuous transitions between successive signs. It has been addressed on specific corpus and in a sequential way [9, 10, 11, 4, 6], although they actually have strong characteristics that make them fundamentally different from unidimensional sequential languages:

- they are multi-channel: information is conveyed through hand motion, shape and orientation, body posture and motion, and facial expression;
- they are strongly spatially organized: events, objects, people or other entities are placed in the signing space and related to each other in a visual way;
- they allow signers to generate new signs – that would not appear in a dictionary – on the go, in an iconic way, or even modify standard signs.

An other important source of difficulty in automatically analyzing a SL video relies in the fact that, conversely to an audio signal, the relevant information to be extracted – human motion and expression – only accounts for a small fraction of the media file size. One should also note that SL are fundamentally 3-dimensional, while a classic RGB video is a 2D projection with no direct depth information.

Last, but not least, is the fact that due to their inherent flexibility, SL show an important variability between different signers in terms of style, dynamics. In particular, it makes it difficult to properly segment SL videos in an automated way.

Proposed solution and methodology

– status of the research

Dimensionality reduction and 3D reconstruction

The first goal of this research project has been to tackle the issue of extracting relevant information in SL videos. It consists of focusing on body pose, facial pose and hand pose. The whole pipeline is summarized on Figure 1.

- We decided to use state of the art 2D body pose detection [5, 16, 18] and then train a Deep Neural Network on Motion

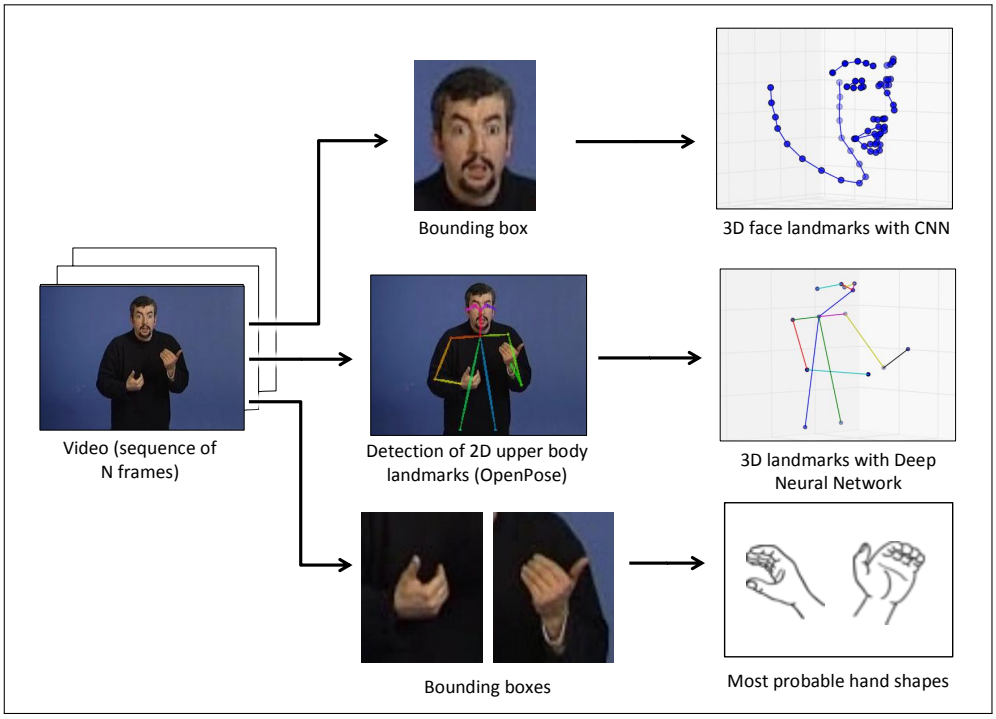

Figure 1. Relevant data extraction pipeline for a 2D RGB sign language video

Capture data recorded by the LIMSI [12, 1] to estimate the 3D body pose [20].

- In order to extract facial information, we used a Convolutional Neural Network (CNN) [3] that estimates the 3D position of about 70 facial landmarks.

- Unfortunately, it has not yet been possible to find or develop a model able to get a reliable estimate on the 3D hand pose. It was then decided – for now – to use a trained CNN model [9] that classifies hand images into 60 SL handshapes.

Citation-form Lexical Sign Recognition

This research project has given some interest to the problem of isolated (citation-form) lexical sign recognition. We have been trying to tackle it with the goal of making a system able to find the signs as close as possible to a query sign, without any need for training or predetermined classes (which is quite different from learning approaches that need to set a certain number of classes *a priori*). For that purpose, we have been developing a modified Dynamic Time Warping algorithm [2] that already gives promising results, and we intend to compare it to traditional learning algorithms.

Sign Language Recognition in real tasks

Our goal is to train a modern Recurrent Neural Network (RNN) on the extracted features described before. For that purpose, we intend to use the French part of DictaSign, a SL corpus with about 8 hours of annotated dialog from 18 native speakers [13]. This network should be able to detect lexical signs in context. We also intend to try training the network to detect higher-level linguistic features, like iconic structures or spatial information, in order to go towards more realistic Sign Language Translation (SLT).

Envisioned contributions

- Proposing an isolated lexical sign recognition system, able to be deployed on any Sign Language, with no need to train – only at least one example for each sign. This will make it possible to query SL videos databases like *Ocelles* [14], with the possibility to add classes without retraining the whole system. It could also be used to regroup signs that appear to be similar in terms of a chosen set of features, like hand motion, hand shape, facial expression, body dynamics, etc.

- Proposing a RNN-based SLR system that will make it possible to search through many SL videos, like one would do to search through web search engines with a text entry.

Interest of the Doctoral Consortium to this research

This consortium would enable us to get valuable feedback on the chosen approach and methods. In our view, accessibility is a proper approach to Sign Language studies, and we would be honored to interact with the faculty and the other participants on this matter. In particular, we are willing to learn from other students and researchers how they managed to understand the fundamental accessibility issues of the communities they work with.

REFERENCES

1. Mohamed-el-Fatah Benchiheub, Bastien Berret, and Annelies Braffort. 2016. Collecting and Analysing a Motion-Capture Corpus of French Sign Language. In *7th*

Workshop on the Representation and Processing of Sign Languages. Portoroz - SI, 7–12.

2. Donald J Berndt and James Clifford. 1994. Using dynamic time warping to find patterns in time series.. In *KDD workshop*, Vol. 10. Seattle, WA, 359–370.
3. Adrian Bulat and Georgios Tzimiropoulos. 2017. 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*.
4. Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, and Richard Bowden. 2017. Subunets: End-to-end hand shape and continuous sign language recognition. In *IEEE International Conference on Computer Vision (ICCV)*.
5. Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2017. Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. In *CVPR*.
6. Runpeng Cui, Hu Liu, and Changshui Zhang. 2017. Recurrent Convolutional Neural Networks for Continuous Sign Language Recognition by Staged Optimization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7361–7369.
7. Mark Dilsizian, Dimitris Metaxas, and Carol Neidle. 2018. Linguistically-driven Framework for Computationally Efficient and Scalable Sign Recognition. *LREC* (2018).
8. Mark Dilsizian, Zhiqiang Tang, Dimitris Metaxas, Matt Huenerfauth, and Carol Neidle. 2016. 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).
9. Oscar Koller, Hermann Ney, and Richard Bowden. 2016. 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*. 3793–3802.
10. Oscar Koller, O Zargaran, Hermann Ney, and Richard Bowden. 2016. Deep Sign: Hybrid CNN-HMM for Continuous Sign Language Recognition. In *Proceedings of the British Machine Vision Conference 2016*.
11. Oscar Koller, Sepehr Zargaran, and Hermann Ney. 2017. 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.
12. Limsi and CIAMS. 2017. MOCAP1. (2017). `https://hdl.handle.net/11403/mocap1/v1` ORTOLANG (Open Resources and TOols for LANGuage) –www.ortolang.fr.
13. S. Matthes, T. Hanke, Anja Regen, J. Storz, Satu Worseck, E. Efthimiou, N. Dimou, Annelies Braffort, J. Glauert, and E. Safar. 2012. Dicta-Sign – Building a Multilingual Sign Language Corpus. In . Istanbul, Turkey, 117–122.
14. Cédric Moreau, Anne Vanbrugghe, Sandrine Rincheval, and Anne-Sophie Destrumelle. 2013. Culture (s) et bilinguisme: Ocelles, les enjeux d'une plateforme collaborative en LSF. *La nouvelle revue de l'adaptation et de la scolarisation* 2 (2013), 225–235.
15. Junfu Pu, Wengang Zhou, Jihai Zhang, and Houqiang Li. 2016. Sign language recognition based on trajectory modeling with hmms. In *International Conference on Multimedia Modeling*. Springer, 686–697.
16. Tomas Simon, Hanbyul Joo, Iain Matthews, and Yaser Sheikh. 2017. Hand Keypoint Detection in Single Images using Multiview Bootstrapping. *arXiv preprint arXiv:1704.07809* (2017).
17. Hanjie Wang, Xiujuan Chai, Xiaopeng Hong, Guoying Zhao, and Xilin Chen. 2016. Isolated Sign Language Recognition with Grassmann Covariance Matrices. *ACM Transactions on Accessible Computing (TACCESS)* 8, 4 (2016), 14.
18. Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. 2016. Convolutional pose machines. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4724–4732.
19. Hee-Deok Yang and Seong-Whan Lee. 2013. 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 (2013), 2051–2056.
20. Ruiqi Zhao, Yan Wang, C Fabian Benitez-Quiroz, Yaojie Liu, and Aleix M Martinez. 2016. Fast and precise face alignment and 3D shape reconstruction from a single 2D image. In *European Conference on Computer Vision*. Springer, 590–603.