Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 6063–6068
Marseille, 11–16 May 2020

MEDIAPI-SKEL - A 2D-Skeleton Video Database of French Sign Language With Aligned French Subtitles

Hannah Bull[1], Annelies Braffort[2], Michèle Gouiffès[3]
[1,2,3]LIMSI-CNRS, [1,3]University of Paris-Saclay
[1,2,3]LIMSI, Campus universitaire 507, Rue du Belvedère, 91405 Orsay - France
{hannah.bull, annelies.braffort, michele.gouiffes}@limsi.fr

Abstract

This paper presents MEDIAPI-SKEL, a 2D-skeleton database of French Sign Language videos aligned with French subtitles. The corpus contains 27 hours of video of body, face and hand keypoints, aligned to subtitles with a vocabulary size of 17k tokens. In contrast to existing sign language corpora such as videos produced under laboratory conditions or translations of TV programs into sign language, this database is constructed using original sign language content largely produced by deaf journalists at the media company *Média-Pi*. Moreover, the videos are accurately synchronized with French subtitles. We propose three challenges appropriate for this corpus that are related to processing units of signs in context: automatic alignment of text and video, semantic segmentation of sign language, and production of video-text embeddings for cross-modal retrieval. These challenges deviate from the classic task of identifying a limited number of lexical signs in a video stream.

Keywords: Sign Language Processing, Computer Vision, Alignment, Retrieval, Video Embeddings

1. Introduction

There is a relative lack of sign language corpora in comparison to other areas of natural language processing, particularly of large, diverse sign language corpora with high quality native speakers in natural settings. Moreover, much attention in the computer vision literature has been given to automatic detection of a limited number of signs, in comparison to other sign language processing tasks (Bragg et al., 2019). In order to combat these two shortcomings, we propose a new dataset for new challenges.

We provide a new corpus available for public research called MEDIAPI-SKEL[1] (Bull, Hannah and Braffort, Annelies, 2019). Our dataset consists of 368 videos totaling 27 hours of French Sign Language (LSF) with French subtitles, generated from the content of the bilingual LSF-French media company *Média-Pi*[2]. The videos are provided in the form of 2D-skeletons with 135 face, hand and body keypoints, but the original videos can be accessed through a subscription with *Média-Pi*. The subtitles provide an accurate alignment between short segments of text and short segments of sign language video. A frame of this data is shown in Figure 1.

This new corpus allows for challenges for sign language processing at a 'sentence' or 'phrase' level, rather than at the 'word' or 'sign' level. We propose three such machine learning challenges for MEDIAPI-SKEL.

The structure of the article is as follows. Firstly we introduce the unique features of sign languages and the resulting limitations of the lexical approach to sign language processing. Secondly, we discuss differences between MEDIAPI-SKEL and other existing sign language corpora. Thirdly, we justify our particular focus on 2D-skeleton data. Fourthly, we provide information relating to the production and content of the corpus. Finally, we present three data challenges appropriate for MEDIAPI-SKEL.

Figure 1: Frame from MEDIAPI-SKEL

2. Particularities of Sign Languages

2.1. Sign Languages

Sign languages are used by millions of people around the world and are an important means of communication for Deaf communities. They are visual-gestural languages, using the modalities of hand gestures, facial expressions, gaze and body movements. The complexity and richness of sign languages is the same as that of spoken languages.

Sign language is not universal; there are an estimated 144 different sign languages used globally (Eberhard et al., 2019). However, one universal characteristic across sign languages is the strong presence of iconicity. Forms can be naturally depicted using gestures, and thus there is a strong connection between form and meaning in signed languages that is less present in vocal languages (Hadjadj et al., 2018). LSF is used in France, but has a grammatical structure strikingly different to French. Not all signs in LSF are *lexical* signs - many do not have a direct equivalent in written French. There is no standard written form of LSF, and the natural form of recording LSF is through video.

[1] ortolang.fr/market/corpora/mediapi-skel
[2] https://media-pi.fr/

Technologies for sign languages lag behind those for written and spoken languages. Search engines, automatic translation tools and even dictionaries are at primitive stages for LSF compared to the resources available for French.

2.2. Limitations of the Lexical Approach

Much work in computer vision and sign language is based around lexical sign recognition. Pfister et al. (2013) use a large database of sign language-interpreted TV programs by the BBC to learn numerous lexical signs. Pigou et al. (2015) use convolutional neural networks to detect 20 signs in videos of Italian Sign Language.

This is a restricted approach to sign language processing for two main reasons. Firstly, fully lexical signs are only part of a discourse. Partially lexical signs arise from the iconicity of sign language, and include signs such as spacial referents, motion, size and shape of objects (Braffort and Filhol, 2014). These signs may be highly dependent on context. Secondly, signs can occur simultaneously; a discourse in sign language is not a linear sequence of signs (Filhol et al., 2014).

The short phrase 'the big mustache' in LSF is a simple example illustrating these linguistic particularities. There is a sign for 'big', but the isolated sign for 'big' will not appear in the phrase 'the big mustache', and the notion of 'big' will in fact be signed simultaneously with the sign for 'mustache' in a contextually dependent manner. Such particular linguistic structures are best observed from native signers in natural settings, hence the value of MEDIAPI-SKEL.

3. Comparison with Existing Corpora

MEDIAPI-SKEL is distinct from existing sign language corpora in multiple aspects.

Firstly, MEDIAPI-SKEL is a large sign language corpus predominantly produced by deaf journalists. The quantity and quality of original and natural content produced by deaf participants in MEDIAPI-SKEL is difficult to find outside of laboratory-produced corpora. The British Sign Language Corpus (Schembri et al., 2017) is one such linguistic corpus created under laboratory conditions, that contains videos of narratives invented by the participants. The DictaSign corpus (Belissen et al., 2020) contains dialogues in LSF between participants. These corpora are produced in a standard format, with consistent camera angles and uniform background conditions. Such corpora are expensive to acquire, translate and annotate; but conditions can be better controlled. On the other hand, the diversity of scenarios and camera angles in MEDIAPI-SKEL better reflects the diversity of real-world sign language videos.

Secondly, the corpus is not produced by real-time translation of written or spoken text. This is distinct from corpora such as RWTH-PHOENIX-Weather (Forster et al., 2012) and the BBC TV corpus (Pfister et al., 2014), which are acquired from sign language translations of TV programs. Sign language during real-time interpretation tends to closely follow the grammatical structure of the spoken language due to strong time constraints (Leeson, 2005). The spontaneous LSF used in our corpus follows a more natural grammatical structure, and it is the text in the subtitles that is adapted accordingly to align to the LSF.

Thirdly, the alignment between subtitles and video is accurate. Some laboratory-produced corpora contain aligned written translations of sign language, such as the Belgian French Sign Language corpus (Meurant et al., 2016) . This is generally not the case for live interpretations, where the subtitles will be aligned to the speech and the sign language translation appears with a time lag. In the case of RWTH-PHOENIX-Weather, the subtitles are manually realigned to match video segments. In the case of the BBC TV corpus, the subtitles are not aligned to the sign language video.

Finally, we provide 2D-skeleton data for all the videos, rather than the original data. This allows us to publish data without negative impact on the economic model of *Média-Pi*, which relies on offering exclusive content to subscribers. We include hand shapes, body pose and facial keypoints in order to best conserve the the intelligibility of the sign language. We encourage the development of skeleton-based methods of sign language processing for reasons discussed in Section 4.

4. Skeleton-Based Models

The 135 2D-skeleton keypoints of the face, hands and body are sufficient to maintain intelligibility in sign language videos. Ko et al. (2019) use these keypoints to automatically translate a limited range of sentences in Korean Sign Language. In fact, 27 hand and face keypoints are sufficient for participating in discussions in American Sign Language (Tartter and Knowlton, 1981).

We encourage the development of skeleton-based models for sign language processing. This restriction of information with respect to the original video should lead to lighter and faster models, with fewer parameters to train. Moreover, external validity of models is more readily attainable, as sign language processing becomes independent of the background and appearance of the signer. Skeleton data can be normalized such that each person has the same body proportions, which removes some of the variation irrelevant to sign language processing.

Skeleton data has proved valuable in action recognition tasks. Yan et al. (2018) demonstrate that the performance of a 2D-skeleton model is capable of achieving a similar accuracy to models using RGB or optical flow data on action classes strongly related with body motion. The performance of skeleton models is lower for human actions in interaction with the environment. However, unlike actions such as 'playing football', sign language does not involve interaction with external objects, and so skeleton data is particularly appropriate for our case.

Finally, another key area in sign language processing is sign language production using avatars. Motion capture is highly successful in creating realistic animations. Body keypoints are captured from an actor and then transferred onto an animated figure. For example, face, body and hand keypoints can be used to animate avatars signing intelligible isolated signs (Alexanderson and Beskow, 2015). Skeleton models can contribute to this area of research in order to create natural-looking signing avatars.

5. Presentation of Corpus

To constitute this corpus, we use 368 subtitled videos totaling 27 hours of LSF footage and written French produced by *Média'Pi*. The content is in the journalistic genre, providing information on a wide variety of events of public interest.

The mode of production varies depending on the subject matter. For example, news stories of national and international interest are generally presented by one journalist, where factual elements are assembled from written press releases. Coverage of local Deaf-related events may involve discussions and interviews with multiple people at the scene.

In a handful of videos, interviews are conducted with people who use an oral language or a foreign sign language, and these interviews are translated into LSF. Both the interviewee and the interpreter will be shown on the screen, however the subtitles will be aligned with the LSF of the interpreter and not the original language of the interviewee. In the case where spoken French is used, the written content of the subtitles is derived from the spoken French and not from the LSF interpretation, and the audio track is removed in the final video.

The number of videos with one main signer is 295, and the number of videos with multiple signers is 73 (Table 1). This diversity in mode of production and mixture of monologue and dialogue makes MEDIAPI-SKEL a challenging dataset that covers a broad range of journalistic styles.

From the original videos, we extract 25 body keypoints, 2x21 hand keypoints and 70 face keypoints using OpenPose (Cao et al., 2018; Simon et al., 2017). We provide these 135 keypoints for every person in every frame of the 368 videos. Each keypoint includes the X and Y pixel value, as well as a confidence score between 0 and 1. Keypoints which fail to be detected or are occluded are accorded 0 values. Note that the body keypoints of the legs and feet are essentially irrelevant for sign language processing, despite the fact that they are included in our extracted skeleton keypoints.

In addition to the 2D-skeleton video data, we provide the associated subtitles in French with their time tags. The subtitles of this corpus are accurately aligned to the 2D-skeleton video content. Each subtitle corresponds to the associated segment of sign language video. This is a particularly complex task, as the syntax of LSF is very different to the syntax of French. In LSF, contextual elements are generally provided at the start of a discourse and then later referred to, while in written French, contextual elements tend to be spread out throughout a text.

In order to maintain an accurate alignment of video segments and subtitles despite strong ordering differences in LSF and French, the subtitles produced by *Média'Pi* are relatively long. The average length is 4.2 seconds or 11 words (Table 1). This provides enough flexibility to reorder the French phrases in a natural way. Indeed, the final sign of a video segment can correspond to the first word of a subtitle.

The frames at moments of transition between subtitles are semantic breaks in the LSF discourse, often characterized by a deceleration of movement. These semantic breaks are worth studying from a linguistic and machine learning perspective, as described in the challenge in Section 6.2.

Table 1 provides a summary of the size and quality of MEDIAPI-SKEL. The number of signers in each video is roughly estimated by selecting the most likely signer at each subtitle interval using descriptive statistics of hand position and velocity, and then using facial recognition to count the number of unique individuals. We consider a video to have one signer if over 95% of the subtitle texts in that video correspond to the same signer. The vocabulary size is computed by counting the number of unique tokens, omitting punctuation.

Table 2 provides summary statistics for the proposed train-dev-test split for the challenges in Section 6.

Global statistics	
# subtitled videos	368
# hours	27
# frames	2.5 million
Video statistics	
Resolution	1080p (327 videos) 720p (41 videos)
Framerate	30 fps (111 videos) 25 fps (242 videos) 24 fps (15 videos)
Average length of video	4.5 minutes
# signers	> 100
# videos with one main signer	295
# videos with multiple signers	73
Text statistics	
# subtitles	20 187
Average length of subtitle	4.2 seconds 10.9 words
Vocabulary size (tokens)	17 428
Vocabulary size (nouns+verbs+adjectives)	14 383

Table 1: Descriptive statistics

	Train	**Dev**	**Test**
# subtitled videos	278	40	50
# hours	20.9	3.0	3.5
# frames	1980k	277k	323k

Table 2: Train - Dev - Test split

6. Data Processing Challenges

We list three challenges that we intend to pursue using MEDIAPI-SKEL. These challenges are illustrated in Figure 2.

6.1. Alignment

We are interested in automatically aligning subtitles, or short segments of text, with the corresponding segments of sign language video. Given an ordered list of subtitle texts, can we automatically subtitle the video?

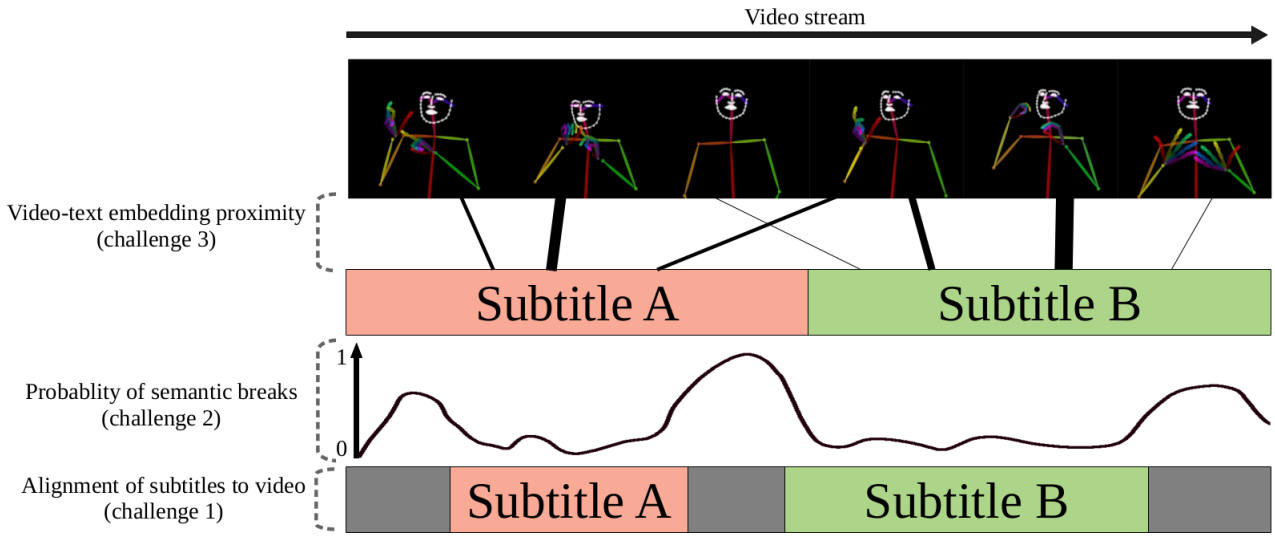

Figure 2: Illustration of challenges: alignment of text and video (challenge 1), semantic segmentation of sign language (challenge 2), and production of video-text embeddings for cross-modal retrieval (challenge 3)

Cui et al. (2017) conduct a similar task, aligning words to lexical sign glosses using recurrent neural networks. Bojanowski et al. (2015) automatically align textual descriptions of sub-tasks to instructional videos. The order of subtasks described in the text annotation follows the order of actions observed in the video, and it is this feature which is exploited in their weakly-supervised learning method. This is also the case for MEDIAPI-SKEL; the order of the subtitles follows the order of the corresponding video segments.

There are numerous applications of automatic alignment of segments of text with segments of video. For example, this task can be used to create an automatic tool for subtitling videos. The process of subtitling a video manually, translating from LSF to written French, takes *Média'Pi* almost 1 hour for each minute of video. This painstakingly long task can be simplified by automatically aligning text with sign language videos. Such tools exist for written languages, for example the software *aeneas*[3], which automatically aligns text with segments of video.

Furthermore, this task can be used to create a bilingual concordancer similar to DeepL's *Linguee*[4]. A bilingual concordancer aligns phrases in one language with phrases in another language. Such a concordancer is a translation aide tool, displaying words and phrases in their context. Whilst *Linguee* aligns text phrases with text phrases, we aim to construct an alignment between segments of text and segments of sign language video. With a concordancer, a translator can quickly search for previously translated segments of text (or even search for signs).

Finally, we can enhance existing corpora such as the BBC TV Corpus (Pfister et al., 2014) by aligning the subtitles to the video stream.

[3] https://www.readbeyond.it/aeneas/
[4] https://www.linguee.fr/

6.2. Semantic Segmentation

In this challenge, we are interested in semantically segmenting a video into short units ('clauses') showing signs in their context. Concretely, we aim to detect the moments between the end of one subtitle and the beginning of the next. This challenge is useful for segmenting a video into bite-sized pieces, each of which could be translated separately. Breaking up a translation task from sign language to written language in this way can speed up the process of translation.

Moreover, this task can be considered as an intermediate task in achieving automatic alignment. A prior segmentation of a sign language video can be used to discretise the problem of matching text segments to continuous video.

The segmentation of sign language into sentence units is discussed in the linguistic literature, and automatic detection of the semantic breaks chosen by *Média'Pi*'s subtitlers can contribute to this discussion. Different ways of defining sentences in sign language are discussed by Crasborn (2007), and both signers and non-signers are capable of recognizing visual cues of the start and end of sentences (Fenlon et al., 2007). These visual cues could be automatically detected by using skeleton-based neural networks, such as the graph convolutional network proposed by Yan et al. (2018). Our challenge can help to quantitatively measure the visual cues of semantic segmentation.

6.3. Video-Text Embeddings

Our third challenge is to find joint vector representations of segments of sign language video and segments of text for video-text retrieval. The goal is to find video embeddings and text embeddings in the same high-dimensional vector space, and then compute the distance between them in that space. This distance represents the semantic distance between the LSF content and the French content.

Mithun et al. (2018) present a method for video-text cross-modal retrieval, which they apply to two datasets containing short videos with textual annotations: the Microsoft Re-

search Video to Text dataset (Xu et al., 2016) and the Microsoft Video Description dataset (Chen and Dolan, 2011). The method is evaluated using rank-based performance in finding the video segment that matches with a text segment, or vice versa.

One possible application of this challenge is a search engine for sign language that finds segments of video given textual input. Another application is to use the distance between video and text embeddings as a measure of loss for the task in Section 6.1., which aims to find the closest match between text segments and video segments.

7. Conclusion

We present MEDIAPI-SKEL, a new 2D-skeleton database of sign language content accurately aligned with subtitles. This corpus can be freely downloaded for public research on Ortolang[5], a language data repository (Bull, Hannah and Braffort, Annelies, 2019).

MEDIAPI-SKEL is appropriate for training a number of sign language processing tasks beyond the classical task of sign spotting. In particular, we aim to develop skeleton-based methods for sign language processing.

Additionally, the corpus can be used for linguistic purposes. For example, one can quantitatively measure visual cues for semantic or grammatical structures, such as questions or lists of items. It could also be used in avatar animation from body keypoints.

8. Acknowledgements

This work has been partially funded by the ROSETTA project, financed by the French Public Investment Bank (Bpifrance). Additionally, we thank *Média-Pi* for providing us with access to their quality content and for allowing us to publish this version of MEDIAPI-SKEL.

9. Bibliographical References

Alexanderson, S. and Beskow, J. (2015). Towards fully automated motion capture of signs – development and evaluation of a key word signing avatar. *ACM Trans. Access. Comput.*, 7(2):7:1–7:17, June.

Belissen, V., Braffort, A., and Gouiffès, M. (2020). Dicta-sign-lsf-v2: Remake of a continuous french sign language dialogue corpus and a first baseline for automatic sign language processing. In *Proceedings of the Twelfth International Conference on Language Resources and Evaluation (LREC'20)*, Marseille, France, May.

Bojanowski, P., Lajugie, R., Grave, E., Bach, F., Laptev, I., Ponce, J., and Schmid, C. (2015). Weakly-supervised alignment of video with text. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, pages 4462–4470, Santiago, Chile, December. IEEE Computer Society.

Braffort, A. and Filhol, M. (2014). Sign language: Constraint-based sign language processing. In *Constraints and Language*, chapter 9, pages 191–218. Cambridge Scholars Publishing.

[5] https://www.ortolang.fr/

Bragg, D., Koller, O., Bellard, M., Berke, L., Boudreault, P., Braffort, A., Caselli, N., Huenerfauth, M., Kacorri, H., Verhoef, T., et al. (2019). Sign language recognition, generation, and translation: An interdisciplinary perspective. In *ASSETS '19: The 21st International ACM SIGACCESS Conference on Computers and Accessibility*, pages 16–31, Pittsburgh, PA, USA, October. ACM.

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

Chen, D. L. and Dolan, W. B. (2011). Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 190–200, Portland, Oregon, June. Association for Computational Linguistics.

Crasborn, O. A. (2007). How to recognise a sentence when you see one. *Sign Language & Linguistics*, 10(2):103–111.

Cui, R., Liu, H., and Zhang, C. (2017). Recurrent convolutional neural networks for continuous sign language recognition by staged optimization. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1610–1618, Venice, Italy, July.

Eberhard, D. M., Simons, G. F., and Fennig, C. D. (2019). *Ethnologue: Languages of the world*. SIL International, Dallas, Texas, 22nd edition.

Fenlon, J., Denmark, T., Campbell, R., and Woll, B. (2007). Seeing sentence boundaries. *Sign Language & Linguistics*, 10(2):177–200.

Filhol, M., Hadjadj, M. N., and Choisier, A. (2014). Non-manual features: the right to indifference. In *6th Workshop on the Representation and Processing of Sign Language (LREC'14)*, Reykjavik, Iceland, May.

Forster, J., Schmidt, C., Hoyoux, T., Koller, O., Zelle, U., Piater, J. H., and Ney, H. (2012). Rwth-phoenix-weather: A large vocabulary sign language recognition and translation corpus. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 3746–3753, Istanbul, Turkey, May. European Language Resource Association (ELRA).

Hadjadj, M. N., Filhol, M., and Braffort, A. (2018). Modeling french sign language: a proposal for a semantically compositional system. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC'18)*, pages 4253–4258, Miyazaki, Japan, May. European Language Resource Association (ELRA).

Ko, S.-K., Kim, C. J., Jung, H., and Cho, C. (2019). Neural sign language translation based on human keypoint estimation. *Applied Sciences*, 9(13):2683.

Leeson, L. (2005). Making the effort in simultaneous interpreting. In *Topics in Signed Language Interpreting: Theory and Practice*, volume 63, chapter 3, pages 51–68. John Benjamins Publishing.

Meurant, L., Gobert, M., and Cleve, A. (2016). Mod-

elling a parallel corpus of french and french belgian sign language. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4236–4240, Portorož, Slovenia, May. European Language Resource Association (ELRA).

Mithun, N. C., Li, J., Metze, F., and Roy-Chowdhury, A. K. (2018). Learning joint embedding with multimodal cues for cross-modal video-text retrieval. In *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval*, pages 19–27, Yokohama, Japan, June. ACM.

Pfister, T., Charles, J., and Zisserman, A. (2013). Large-scale learning of sign language by watching tv (using co-occurrences). In *Proceedings of the British Machine Vision Conference*, Bristol, UK, September. BMVA Press.

Pfister, T., Simonyan, K., Charles, J., and Zisserman, A. (2014). Deep convolutional neural networks for efficient pose estimation in gesture videos. In *Asian Conference on Computer Vision (ACCV)*, pages 538–552, Singapore, November. Springer.

Pigou, L., Dieleman, S., Kindermans, P.-J., and Schrauwen, B. (2015). Sign language recognition using convolutional neural networks. In *Computer Vision - ECCV 2014 Workshops*, pages 572–578, Zurich, Switzerland, September. Springer International Publishing.

Schembri, A., Fenlon, J., Rentelis, R., and Cormier, K. (2017). British Sign Language corpus project: A corpus of digital video data and annotations of British Sign Language 2008–2017 (Third Edition). *http://www.bslcorpusproject.org*.

Simon, T., Joo, H., Matthews, I., and Sheikh, Y. (2017). Hand keypoint detection in single images using multiview bootstrapping. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4645–4653, Venice, Italy, July.

Tartter, V. C. and Knowlton, K. C. (1981). Perception of sign language from an array of 27 moving spots. *Nature*, 289(5799):676.

Xu, J., Mei, T., Yao, T., and Rui, Y. (2016). Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5288–5296, Las Vegas, USA, June.

Yan, S., Xiong, Y., and Lin, D. (2018). Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*, New Orleans, Louisiana, USA, Feburary.

10. Language Resource References

Bull, Hannah and Braffort, Annelies. (2019). *MEDIAPI-SKEL*. Limsi, distributed via ORTOLANG (Open Resources and TOols for LANGuage), Limsi resources, 1.0, ISLRN 184-726-682-550-4.