Towards an Automatic Annotation of French Sign Language Videos: Detection of Lexical Signs

Hussein Chaaban, Michèle Gouiffès, Annelies Braffort

Towards an Automatic Annotation of French Sign language Videos: Detection of Lexical Signs

Hussein CHAABAN[1,2,3,4], Michèle GOUIFFÈS[1,2,3,4], and Annelies BRAFFORT[2,3,4]

[1] Paris Sud University
[2] Paris-Saclay University
[3] Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur LIMSI
[4] Centre national de la recherche scientifique CNRS

Abstract. This paper presents an approach towards an automatic annotation system for French Sign Language (LSF). Such automation aims to reduce the processing time and the subjectivity of manual annotations done by linguists in order to study the sign language and simplify indexing for automatic signs recognition. The described system uses face and body keypoints collected from 2D RGB standard LSF videos. A naive Bayesian model was built to classify gestural units using the collected keypoints as features. We started from the observation that, for many signers, the production of lexical signs is very often accompanied by mouthing. Effectively, the results showed that the system is capable of detecting lexical signs, with highest success rate, using only information about mouthing and head direction.

Keywords: LSF Videos · Annotations · Lexical Signs · Mouthing

1 Introduction

Sign Languages (SL) are visuo-gestural languages used mainly by the deaf community. Very few linguist studies have been produced to explain and formalize their rules and grammar. The first contemporary linguist to study SL was William C. Stokoe [17] who described the language in terms of phonemes (or cheremes) and built a written transcription of it. This work has laid the groundwork and paved the way for deeper research on SL. Today, linguists collect and annotate videos of signers in natural contexts in order to extract knowledge from them. Currently, most of the SL videos are manually annotated using a software like ELAN [21] or ANVIL [10]. Though this process consumes a lot of time and the produced annotations are usually non-reproducible since they depend on the subjectivity and the experience of the annotator. An automatic annotation system could certainly accelerate the work and enhance the reproducibility of the results.

Cues about the hands (shape and motion), face expressions, gaze orientation, mouthing are useful to be annotated. When talking about SL, most of the

non SL signers tend to think directly of the hands. In fact, two communication channels exist: Manual Components (MC) consisting of the hand shapes, orientations and motions and, Non-Manual Components (NMC) consisting of face features and body pose. Cuxac's model [5] presents two ways of signifying using a combination of 4 different MC and 4 NMC:
1) "saying and showing" with an illustrative intent which consists of **Highly Iconic Structures (HIS)** that include Transfers in Sizes and Shapes (TSS) of objects, in Situations (ST) and in Persons (PT);
2) "without showing" which consists of **Lexical Signs (LS)**, *i.e.* predefined signs in a dictionary, and pointing. More than 65% of the signs are lexical [8].

Thus, distinguishing these two classes, LS and HIS, would be a first step before applying a dedicated processing for each of them. To do so, this paper proposes to determine the more relevant body and face features to detect LS in a SL discourse. Then, it illustrates this result by testing a classification method. The experiments are made on a French Sign Language (LSF) video dataset consisting of standard RGB videos. Thus, it is not required to use any specific sensor or wearable device, which enlarges its possible use-cases. This pre-annotation intends to facilitate the work of the linguists and can be useful to constitute annotated data for further deep learning strategies.

The remainder of the paper is structured as follows. The next section discusses the previous work conducted on annotation of SL videos. Section 3 describes the datasets that are studied in this paper. Then, Sections 4 and 5 describe respectively the features and classification methods. To finish, Section 6.3 discusses the results.

2 Related Work

In the literature, few papers explored the automatic annotation of SL. The majority of these works on SL annotation study the American Sign Language (ASL) as it is the richest one in terms of databases. The conducted studies on ASL may not be necessarily suitable nor applicable to LSF since SL are not universal languages: each country has its own SL and grammar.

The first attempts of SL recognition were conducted on isolated signs. The general idea was to extract some features from images in order to identify signs using a classifier such as SVM [16] [7], neural networks [9], HMM [1], KNN [15]. These works were mainly focusing on the MC as features as it was believed that hands had the main information in an SL speech. Nowadays, many image processing and object segmentation techniques were developed [20] and with the revolution in machine learning, the systems are capable of estimating and tracking face [18] and body [19] in real time with high success rate using only 2D image features. Most works on SL recognition focus on specific datasets, made in controlled environments (uniform background, signer with dark clothes) and dealing with a specific topic, such as weather [11]. But the real challenge in SL recognition remains in identifying dynamic signs, *i.e.* signs in real SL speech, and

most importantly independently of the signer [12]. Such work requires a huge annotated dataset, which is not available for LSF.

Concerning automatic annotation, most of the proposals try to annotate the segments by describing facial and body events as mouthing, gaze, occlusion, hands placements, handshapes, and movements [14]. Few of them go further and exploit these events by combining MC and NMC to add a second level annotation such as LS and HIS. In fact, [6] succeeded to annotate pointing in LSF videos by combining MC and NMC. In [13], the MC and NMC are tracked in order to categorize LS. However, an actual annotation of LS was lacking, and the tracking of NMC was done on controlled videos of the head. In our work, we tested some combinations of MC and NMC to figure out which components are the most effective to classify LS.

3 Data

The dataset is a portion of MOCAP dataset, which collects RGB videos in LSF produced in our lab for other purposes [5]. The videos show the signer from hip up face view. These videos are standard (2D, 720x540 pixels, 25 FPS). We used 49 videos with 4 different signers with randomly picked combinations for learning and test sets. The length of videos varies between 15 to 34 seconds (average of 24 seconds, 19.63 min in total). In the videos the signers were asked to describe what they see in an image (Figure.1). The given images represented 25 different

Fig. 1. Sequence of lexical sign "Salon"

scenes (Fig.2) such as a living room, a forest, a wine store, a library, a city, a monument, a construction site... The images were chosen to have a variety of LS and HIS. All the videos were annotated manually by one expert. The annotations include gaze, LS and HIS. 1011 signs were annotated, 709 were LS and 304 were HIS.

[5] Because of privacy policies, these videos are not available online https://www.ortolang.fr/market/corpora/mocap1

Fig. 2. Examples of scenes to be described by the signers

4 Features extraction

As far now, linguists did not establish a unified way for annotation nor a predefined list of MC and NMC to track. Checking the literature, most papers were interested in studying the handshapes, their placements, motion, direction and symmetry between them as MC and mouthing, mouth gestures, gaze and eyebrows as NMC.

To extract the features, we use OpenPose [4], a recent real time pose estimation library for face and body, which provides the coordinates of keypoints (body articulators and face elements). We have processed these coordinates to provide more evolved features described hereafter.

Mouthing Based on the work of [3] which proves that mouth features are important indicators of LS, our first work has consisted in tracking the mouthing. Then other MC and NMC features were successively added to see how the classification improves. OpenPose provides the coordinates of 20 points that define the outer-line of the lips (Fig.3 (a)).

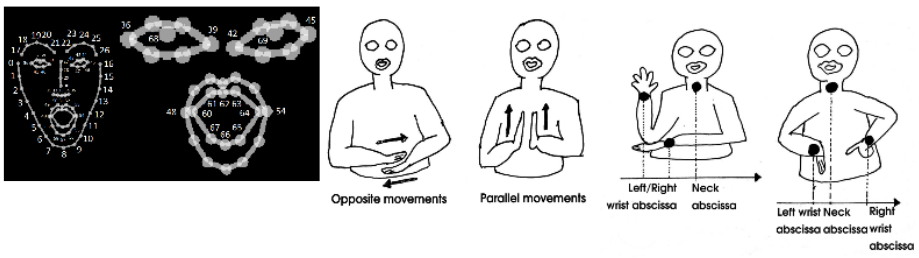

Fig. 3. (a) Facial Keypoints of OpenPose. (b) Relative movements of hands. (c) Placement of signs.

We assume that a mouthing is detected whenever the signer opens his mouth due to the pronunciation of a vowel, which is not the case for mouth gesture.

To detect the opening of the mouth, we calculated the isoperimetric ratio (or circularity) of the interior of the lips using the formula: $\text{IR} = \frac{4\pi a}{p^2}$ where a is the area and p is the perimeter. The higher the ratio is, the more the mouth is open, which is the sign of a mouthing.

However, mouth is often occluded when the signs are formed in front of the head. To handle the problem, a temporal analyzing window of 5 frames is used in which the last relevant IR value is kept, *i.e.* $\text{IR} \leq 1$. The occlusion is detected when the distance between hands and mouth falls under a threshold (80 pixels in our dataset) and when lips coordinates are null.

Gaze/Head direction Gaze plays an important role in HIS, when the signer places objects in the signing space in front of him, and wants to draw the attention of the partner on something in the signing space.

Our facial features are detected using OpenFace [2]. Theoretically, the gaze could be tracked from this model. However, because of the low resolution of the images under consideration, we had to use only the head direction (which is generally close). We define the head direction as a ratio, where 0 refers to the signer head in center position, and negative/positive values stands for left and right respectively.

Bi-manual motion During HIS, the signer can draw objects in the signing space, generally with both hands moving in a symmetrical or opposite way. With OpenPose, we can get with high precision the coordinates of both wrists and elbows. Using these coordinates we deduce the velocity and direction of the hands movements to create a motion characteristic vector for each arm. The correlation of the two vectors of the two arms can give us an information about the relative movements of arms : symmetrical (both velocity and direction are similar), opposite (similar velocity and opposite directions).

Signing space The LS, generally known by the interlocutor, are mostly made in front of the signer. Contrary to the HIS (transfers), they require less placement of objects in the signing space (left and right). The abscissas of neck and wrists, found in each frame by OpenPose (Fig. 3 (b) and (c)) are used to evaluate this location. Therefore we simply tested if the abscissa of the neck is between the abscissas of both wrists. If it is the case then the sign is centered if not the sign is either to the left or to the right.

5 Lexical classification

Since we do not have a huge dataset for learning, a simple classifier has been chosen, instead of convolutional neural networks. The first step in building our classifier was finding the decision rule. Using the extracted features from the learning data and combining them with the annotations of LS and HIS, we drew the distribution of each parameter between the two types of signs along the

frames of the videos. Since the values of our features are continuous, we took the assumption that their distributions are normal with mean μ_k and variance σ_k^2. In Fig.4, it can be seen how the features values (for instance IR) distributed

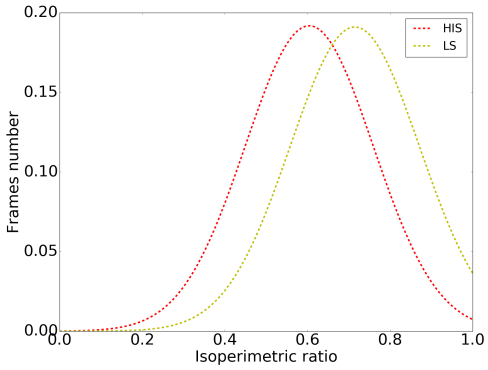

Fig. 4. Distribution of isoperimetric ratio IR between the two types of signs (LS and HIS.

between LS and HIS for a specific learning set. These functions represent the probability distribution of each feature (x) given a sign type (C). $P(x_i \mid C)$ can be computed by plugging x_i into the equation for a Normal distribution parameterized by μ_k and σ_k^2

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

The discussion so far has derived the independent feature model, that is, the naive Bayes probability model. The naive Bayes classifier combines this model with a decision rule. One common rule is to pick the hypothesis that is the most probable; this is known as the maximum *a posteriori* or MAP decision rule. The corresponding Bayes classifier assigns a class label $\hat{y} = C_k$ for some k as follows:

$$\hat{y} = \text{argmax}_{k \in \{1,\ldots,K\}} \; P(C_k) \prod_{i=1}^{n} p(x_i \mid C_k). \tag{2}$$

with the believe that all the features are independents.

After creating our model using the learning dataset, for each new frame in the testing dataset we calculate:

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

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

where $F1$ is Mouthing, $F2$ is head pose, $F3$ is hands symmetry and $F4$ is sign placement. Then we compare (3) and (4), if the result of (3) is bigger then the result of (4) the new frame is part of LS if not then it is part of HIS sign.

6 Experiments and Results

6.1 Preliminary analysis

The manual annotations provided in MOCAP were useful in a first time to establish some statistics about the signs. We were most interested in the signs frequencies and their lengths. We discovered that 69.99% of signs in the database are lexical where 30% are HIS and that the standard length of a sign is between 3 and 10 frames, as shown by the distribution of the sign lengths on Fig.5.

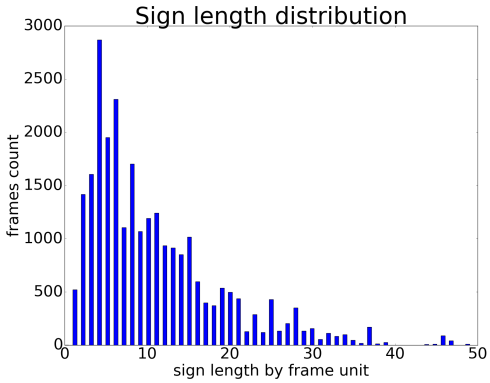

Fig. 5. The distribution of sign lengths in the dataset MOCAP.

6.2 Evaluation

The proposed method is applied to the dataset detailed in section 3.

The classification results of LS are compared to the manual annotations. Figure.6 shows, for one of the videos, and for each frame, an example of classification result (in red) compared to the annotation (in blue). Because of the subjectivity of the annotations, an annotated LS is considered as correctly detected when 3 consecutive frames (smallest sign length of a LS) classified as lexical fall in the range of the annotated sign.

For the evaluation metrics, we counted the true positives (TP) among detected lexical signs, false positives (FP), true negatives (TN) and false negatives

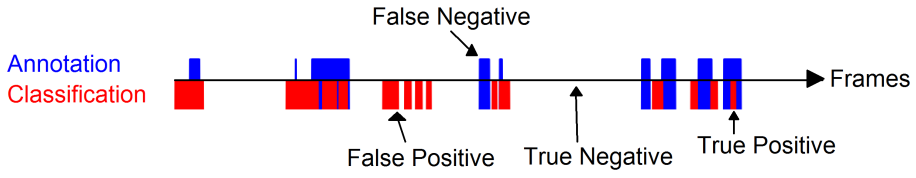

Fig. 6. Counting of False/True Positives and False/True Negatives

(FN) in each video in the test dataset. Then we compute the TP and TN rates (TPR and TNR), the positive prediction value (PPV) and F1-score :

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

6.3 Classification results

First, the results of our method are evaluated for each signer individually and then combined to check if the classification is independent of the signer. For each experiment, the Mouthing (M) is tested alone, and the other ones are successively added: Head direction (H), Bi-manual motion (B) and Sign placement (S).

Intra-signer study. For each signer, the videos 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. The averages and standard deviations of the results are shown in Table.1.

Table 1. The evaluation of the results for intra-signer classification using the features Mouthing (M), Head direction (H), Bi-manual motion (B) and Sign placement (S)). The shown values are the average of all the results coming from each signer separately

Features	TPR		TNR		PPV		F1 score	
	μ	σ	μ	σ	μ	σ	μ	σ
M	0.24	0.09	0.80	0.15	0.55	0.24	0.32	0.10
M + H	0.57	0.11	0.56	0.19	0.48	0.17	0.50	0.13
M + H + B	0.57	0.13	0.56,	0.18	0.46	0.17	0.48	0.14
M + H + B + S	0.57	0.10	0.55	0.18	0.46	0.16	0.48	0.12

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 and the results are shown in Table.2.

Table 2. The evaluation of the results for inter classification using the features Mouthing (M), Head direction (H), Bi-manual motion (B) and Sign placement (S)). The shown values are the average of all the results coming from all signers combined

Features	TPR		TNR		PPV		F1 score	
	μ	σ	μ	σ	μ	σ	μ	σ
M	0.25	0.09	0.80	0.08	0.52	0.15	0.33	0.10
M + H	0.61	0.11	0.57	0.19	0.50	0.16	0.53	0.09
M + H + B	0.62	0.12	0.57	0.17	0.51	0.15	0.53	0.09
M + H + B + S	0.58	0.12	0.57	0.17	0.49	0.14	0.50	0.09

By analysing the tables 1 and 2, mouthing and head orientation appear to be the most relevant features for distinguishing LS from HIS. While, the bimanual signing and the placement of signs seem not adding any relevant information for this task. The similarity between the results obtained for intra-signer and for inter-signer experiments confirms the generality of our approach. The performance of the results seems to be low compared to more standard gesture recognition applications. This is explained by the huge variety of the motion made for signs, the imperfection and subjectivity of the annotations and the error margin of OpenPose and OpenFace during the features extraction since we are working on low resolution videos. However, our application consists of a semi automatic annotation of SL. It will be of great help for linguists, who will just have to confirm or not the correctness of the classification.

6.4 Impact of the segmentation

As mentioned previously, the manual annotations of the videos are both subjective and imprecise. Each annotator has his own rules to define the beginning and the end of each sign. We wanted to find how many of the False Positive classified LS actually refer to a neighbour existing sign in the annotation to test the hypothesis that this classified sign was considered as False detection due to the subjectivity of the the annotation and a delay of between the annotation and the detection (Fig. 7). Thus we enlarged each detected sign by 3 frames (small-

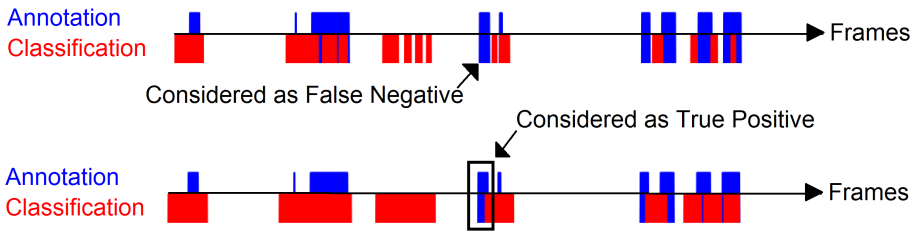

Fig. 7. Before enlarging detected signs (upper image) and after (lower image)

est length of a LS) at the beginning and the end of the sign and recalculated the evaluation results. The new values in Table.3 show an improvement of the classification rate. Even if the improvement is not that high it definitely makes us more curious about the importance of the segmentation of detected signs.

Table 3. The evaluation of the results for inter and intra classification after enlarging the detected signs

Intra-signer

Features	TPR		TNR		PPV		F1 score	
	μ	σ	μ	σ	μ	σ	μ	σ
M + H	0.54	0.15	0.74	0.14	0.64	0.16	0.57	0.12
M + H + B + S	0.54	0.16	0.73	0.13	0.63	0.14	0.56	0.12

Inter-signer

Features	TPR		TNR		PPV		F1 score	
	μ	σ	μ	σ	μ	σ	μ	σ
M + H	0.57	0.15	0.73	0.14	0.63	0.12	0.57	0.08
M + H + B + S	0.52	0.16	0.73	0.13	0.63	0.13	0.54	0.11

7 Conclusion

This paper has proposed a tool that will be useful for linguists to pre-annotate Sign Language (SL) videos, in order to alleviate the annotation burden. This first step distinguishes temporal segments that correspond to lexical signs from other segments, such as the highly iconic ones. According to the study made on the features, it has been shown that mouthing and head orientation are the most discriminant features for this task. This work has several perspectives. First, the impact of other features will be tested and other classifiers such as SVM will be used just to compare the results and observe the impact of the classification system on the results. Then, once a lexical sign is detected in the video, we will have to refine the temporal segmentation around this detection. After segmentation, it will be possible to launch a sign recognition algorithm on the resulting LS segments. It will be interesting also to test our approach on other SL, in order to test its universality.

References

1. Assaleh, K., Shanableh, T., Fanaswala, M.: Persian sign language (PSL) recognition using wavelet transform and neural networks. Journal of Intelligent Learning Systems and Applications **2**, 19–27 (January 2010). https://doi.org/10.4236/jilsa.2010.21003
2. Baltrušaitis, T., Zadeh, A., Chong Lim, Y., Morency, L.P.: Openface 2.0: Facial behavior analysis toolkit. IEEE International Conference on Automatic Face and Gesture Recognition (2018)
3. Balvet, A., Sallandre, M.A.: Mouth features as non-manual cues for the categorization of lexical and productive signs in french sign language (LSF). 6th Workshop on the Representation and Processing of Sign Languages: Beyond the manual channel (May 2014), halshs-01079270, reykjavik, France
4. Cao, Z., Hidalgo, G., Simon, T., Wei, S.E., Sheikh, Y.: OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields. arXiv preprint arXiv:1812.08008 (2018)
5. Cuxac, C.: La langue des signes française (lsf). les voies de l'iconicité. Bibliothèque de Faits de Langues (15-16) (2000), paris: Ophrys
6. Garcia, B., Sallandre, M.A., Schoder, C., L'Huillier, M.T.: Typologie des pointages en langue des signes franaise (lsf) et problmatiques de leur annotation (2011)
7. Huang, C.L., Tsai, B.L.: A vision-based Taiwanese Sign Language Recognition. 20th International Conference on Pattern Recognition, ICPR pp. 3683–3686 (August 2010). https://doi.org/10.1109/ICPR.2010.1110, istanbul, Turkey
8. Johnston, T.: Lexical frequency in sign languages. The Journal of Deaf Studies and Deaf Education **17**(2), 163–193 (Spring 2012), https://doi.org/10.1093/deafed/enr036
9. Karami, A., Zanj, B., KianiSarkaleh, A.: Persain sign language (psl) recognition using wavelet transform and neural networks. Expert systeme with Applications **38**(3), 2661–2667 (March 2011). https://doi.org/10.1016/j.eswa.2010.08.056
10. Kipp, M.: Anvil - a generic annotation tool for multimodal dialogue. INTERSPEECH (2001)
11. Koller, O., Forster, J., Ney, H.: Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers. Computer Vision and Image Understanding 141 pp. 108–125 (December 2015)
12. Liang, Z.j., Liao, S.b., Hu, B.z.: 3D Convolutional Neural Networks for Dynamic Sign Language Recognition. The Computer Journal **61**(11), 1724–1736 (November 2018). https://doi.org/10.1093/comjnl/bxy049
13. Marek, H., Krňoul, Z., Campr, P., Mülle, L.: Towards automatic annotation of sign language dictionary corpora. Text, Speech and Dialogue. TSD, Lecture Notes in Computer Science **6836**, 331–339 (2011). https://doi.org/10.1007/978-3-642-23538-2-42, springer, Berlin, Heidelberg
14. Naert, L., Reverdy, C., Caroline, L., Gibet, S.: Per channel automatic annotation of sign language motion capture data. Workshop on the Representation and Processing of Sign Languages: Involving the Language Community, LREC (May 2018), https://hal.archives-ouvertes.fr/hal-01851404, miyazaki Japan
15. Nandy, A., Prasad, J.S., Mondal, S., Chakraborty, P., Nandi, G.C.: Recognition isolated indian sign language gesture in real time. Information processing and management Communications in Computer and Information Sciences **70**, 102–107 (2010)

16. Rashid, O., Al-Hamadi, A., Michaelis, B.: Utilizing invariant descriptors for finger spelling american sign language using svm. In: Bebis G. et al. (eds) Advances in Visual Computing. ISVC 2010. Lecture Notes in Computer Science **6453** (2010). https://doi.org/10.1007/978-3-642-17289-2-25, springer, Berlin, Heidelberg
17. Stokoe, W., Casterline, D., Croneberg, C.: A dictionary of american sign language on linguistic principles (revised ed.) (1976), [Silver Spring, Md.]: Linstok Press
18. Viola, P.A., Jones, M.J.: Robust real-time face detection. International Journal of Computer Vision **57**, 137–154 (2004). https://doi.org/10.1023/B:VISI.0000013087.49260.fb
19. Wei, S., Ramakrishna, V., Kanade, T., Sheikh, Y.: Convolutional pose machines. CoRR **abs/1602.00134** (August 2016), http://arxiv.org/abs/1602.00134, dblp computer science bibliography, https://dblp.org
20. Wiley, V., Lucas, T.: Computer vision and image processing: A paper review. International Journal of Artificial Intelligence Research **2**(1), 28–36 (June 2018). https://doi.org/10.29099/ijair.v2i1.42
21. Wittenburg, P., Levinson, S., Kita, S., Brugman, H.: Multimodal annotations in gesture and sign language studies. LREC (2002)