Detecting Palms Open Sign with Decision Tree Classifier

s1010384
Radboud University

ABSTRACT

Introduction Sign language is used as a way of communicating through facial expression, body language and hands. A often-used sign is the Palms-Open (PO) sign. This research executes binary classification in detecting the PO sign by means of video processing on a subset of the NGT (Nederlandse Gebarentaal) Corpus.
Methods With a Decision Tree (DCT) Classifier, three videos of the NGT are processed in order to detect the PO sign. Features are extracted for each video separately and are also combined, leading to four data subsets for classification. Feature extraction approaches include the VGG19 model and the use of *Open Pose* keypoints.
Results Results have shown that the occurrence of the PO sign with respect to other signs is not as high as the literature denotes. Furthermore, with a F1-score of 0.99, the decision tree classifier is able to detect the PO sign almost perfectly.
Conclusion and Discussion In conclusion, we can state that - based on our results - the VGG19 model in combination with *Open Pose* keypoints create a sufficient feature subset for binary classification of the PO sign. We recommend to explore a larger subset of the NGT Corpus for further research.
For full code, please refer to `https://github.com/Josien94/TxMM`.

1 INTRODUCTION

Sign language is a used as a way of communicating through facial expression, body language and hands [16]. It is widely used by the deaf community and/or people who are hard of hearing. A sign can either be expressed as a gesture, or by means of fingerspelling (also called *dactylology*).

One sign of which the semantically and lexical meaning is not often well-defined is the *Palms-Open* (PO) sign, also referred to as the *Palms-Up* sign [6]. This sign (illustrated in Figure 1), is formed by the rotation of one or both open hands towards an upward position. This sign is not only used in sign language, but also as a common gesture with speech. A research conducted by Chu et al. [5] describes the sign frequency of the PO sign. Results have shown that of more than 8000 gestures, 24% of them involved the PO sign. This is also supported by a research of Johnston [11], which describes the 300 most frequent signs used in Australian Sign Language (Auslan). With 40% this sign (annotated with `G(5-UP):WELL`) is the second most frequent used sign.

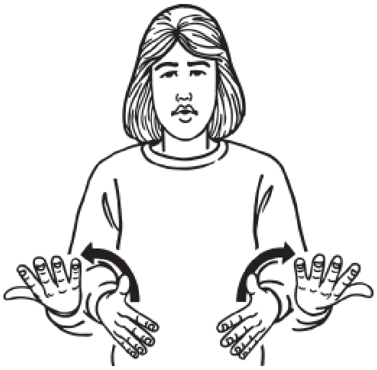

Figure 1: Palms-Open sign [15].

When annotating gestures - not only the PO sign - several difficulties arise. Variation in two levels of sign language make annotation (either manually or automatically) challenging: 1) Phonological variation and 2) Lexical variation [18].

Phonological variation has to do with the form of signs. For example. location of the sign, the movement and orientation of the hands [14].Lexical variation has to do with the use of different signs for the same word, often due to region, ethnicity or educational background [19].

When looking at these challenges in the light of the PO-sign, there are several phonological variations that stand out. First of all, a sign can be signed with one or two hands. During the PO-sign, the non-dominant hand will drop. A research of McKee and Wallingford [15] have shown that 63% of tokens were produced two-handed, and 37% one-handed. The signs surrounding the token - in this research referred to as *context* - is of direct influence of possible hand-dropping. Secondly, the handshape of the PO sign can be different between sign language speakers. The basic form is - as aforementioned - an open palm, oriented upwards. However, the PO sign can also be quite subtle, as is described in research of McKee and Wallingford [15]. A third difficulty is observed when taking the location of the sign into account. The PO-sign is most often signed in front of the torso, but it is found that the PO-sign is often articulated at different heights [15].

In this research, we try to overcome these challenges and use machine learning techniques in order to detect the PO signs in Dutch Sign Language (NGT; Nederlandse Gebarentaal). At the Radboud University, a Sign Language research group at the Centre for Language Studies conducts many research to NGT [1]. A project from 2006-2009 *Corpus NGT* focused on annotation NGT. These annotations were mainly created in *ELAN* software [2] [23]. The main reason for this, is the ability of multiple annotation layers. The group is currently working on the project *Corpus NGT 2* to - amongst

TxMM'20-'21, Radboud University, Nijmegen, Netherlands

others - add more sentence-level translations. In this research, a subset of this (annotated) corpus is used in order to perform binary classification to detect the PO sign. We used a Decision Tree Classifier (DCT) as the main model to predict the presence of a PO sign in the frames of these videos [22].

2 RELATED WORK

Several research has been conducted into the use of NGT. Research of Crasborn and van der Kooij [8] gives more insight in the (general) phonological variations of NGT. Examples are variations in eye-contact, raise of brows and head and body position.

A developing field of research is the development of Sign Language Recognition (SLR) systems. These systems are developed to "translate" sign language to text, using video processing. This turns out to be a pretty hard tasks and several challenges occur when performing continuous SLR [10]. Especially, the semantics in Sign Language is a well-known problem [20].

In the past, several models have been developed to use video processing in order to extract signs. An example is the use of Conventional Neural Networks and other deep learning approaches by Adaloglou et al. [3]. Other authors used use of Kinect, which construct 3D models using Markov models [13]. The authors make use of software tracking the body and collecting skeleton data. Another software which makes use of skeleton data is *OpenPose* [9]. This software is able to detect joints of the human body, face and hands, based on video-input. For each frame, this is then translated in so-called *keypoints*. These keypoints indicate a particular joint in the body, face or hand. OpenPose outputs a JSON format with a mapping from the body to keypoints, where the location coordinates for each keypoint in space are described. This can either be in 2D (i.e. (x, y)) or 3D (i.e. (x, y, z)). OpenPose is used in a research by Ko et al. [12], where results have shown that quite a robust SLR system by human keypoint estimation.

3 METHODS

The general approach in this research can be subdivided into 4 tasks: 1) Collecting data, 2) Feature extraction, 3) Preprocess data, 4) Tune hyperparameters, 5) Perform and Evaluate classification. The following sections will describe these tasks in more detail respectively.

3.1 Collecting data

The video data - originated from [7] - are publicly available. A subset of the corpus exiting of 385 videos are stored in *MPEG* format on a designated server. For this subset, the *ELAN* annotated data is also made available for this research. This research makes use of 3 video's of the NGT corpus, with respectively 4115, 4115 and 8532 frames. The videos are hereafter referred to by their filename: 1) *CNGT0001_S003*, 2) *CNGT0001_S004* and 3) *CNGT0004_S003*.

As aforementioned, the *ELAN* annotated data of these videos are also made available for this research. This data contains the frame numbers showing the PO sign.

3.2 Feature extraction

Detecting the PO sign is done by means of binary classification of each frame, where each frame is a sample. To extract the right features from the frames of the videos, two resources are used: 1) VGG19 model and 2) *Open Pose*. The VGG19 model is a convolutional neural network (CNN), consisting of 19 layers [21]. It is a pre-trained model, trained on ImageNet, a database of millions of images [17]. The model makes transfer learning possible, where features of new data can be defined. The architecture of the VGG19 model is shown below in Figure 2.

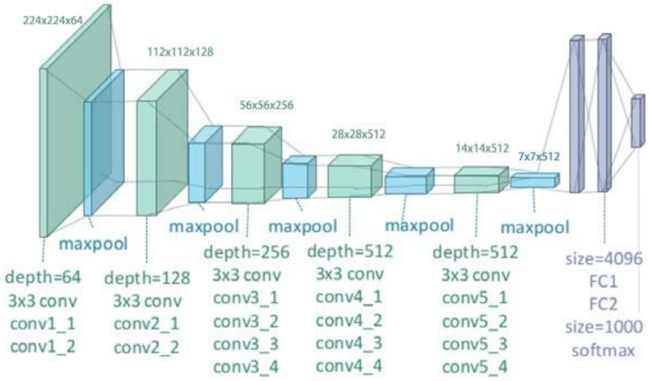

Figure 2: Architecture VVG19 model. FC indicates Fully Connected [24].

For feature extraction, the input of the second-last layer is used, meaning 4096 features for each frame.

The second resource includes the keypoints of the *Open Pose* software, as described in section 2. Here, the 2D keypoints of the face, body and hands are extracted for each frame. This adds another 274 features for each frame.

This results in the following dimensions of the features for each video:

Dataset	Dimension features
CNGT0001_S003	4014 × 4370
CNGT0001_S003	4014 × 4370
CNGT0001_S003	8431 × 4370
Combined	16459 × 4370

Table 1: Dimension features for binary classification for the four datasets.

3.3 Preprocess data

The features of each frame are split into a training- and test set, with a ratio of 0.8:0.2. With the help of the annotated data, the label of each frame could be determined. As the PO sign is underrepresented compared to all other frames, we performed oversampling using SMOTE (Synthetic Minority Oversampling Technique). This technique creates synthetic replicas of existing samples, leading to slight alterations of the existing samples [4]. As a result, there is an equal amount of samples classified with the PO sign being present or absent.

3.4 Tune Hyperparameters

The hyperparameter of a DCT classifier is the maximum depth of the tree. For this step, the training data is split further into a train- and validation dataset, again with a ratio of 0.8:0.2. With stratified K-fold cross validation - with $k = 10$ - the optimal depth of the tree is determined. For this, the depth is varied between 10 and 100 with intermediate steps of 10. For each depth, the data is fitted on the training data and tested on the validation data. For each fold, the Root Mean Square Error (RMSE) is computed. This evaluation metric is defined as $\sqrt{\frac{1}{m}\sum_i(\hat{y}_i - y_i)^2}$, with $\hat{y}_i$ the predicted labels and y_i the known labels. This leads to an (averaged) RMSEA for each depth, where a lower RMSE means a better fit of the classifier.

Four separate classifications are performed: on each of the three videos separately (*CNGT0001_S003*, *CNGT0001_S004* and *CNGT0004_S003*) and on all three videos combined. The optimal depth of the classifications with corresponding RMSE are shown below:

Dataset	Optimal depth	RMSE
CNGT0001_S003	60	0.0056
CNGT0001_S003	80	0.021
CNGT0001_S003	70	0.078
Combined	50	0.073

Table 2: Optimal depth DCT classifier for the four datasets.

3.5 Perform and evaluate classification

With the optimal depths obtained in the previous step, the final classification is performed on the (oversampled) test set. The results of the classification are evaluated by means of the F1-score. This metric - defined as the harmonic mean of the precision and recall - is calculated for each of the four classifications. The score is based on the confusion matrix for each classification.

4 RESULTS

To have an overview of the used data, first off the distribution of the PO sign is plotted for the three videos on which binary classification is performed. These results are shown in Figures 3, 4 and 5 respectively. We see that the distribution of the PO sign compared to all other frames is quite low. It also supports the decision to use oversampling in order to increase the performance of binary classification.

With the depths that are described in 3, four classifications are performed. The confusion matrices of these classifications are shown in Figures 6, 7, 8 and 9 respectively.

This leads to the following F1 scores, displayed in Table 3.

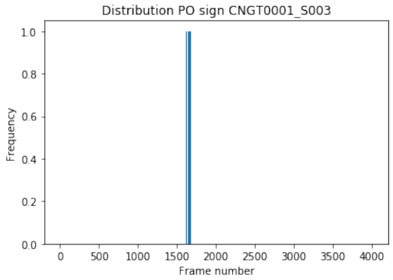

Figure 3: Distribution PO sign *CNGT0001_S003*

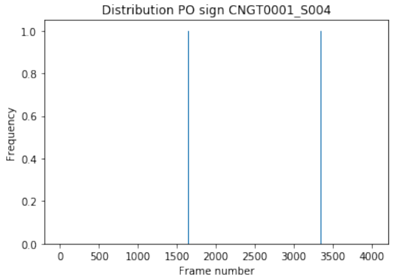

Figure 4: Distribution PO sign *CNGT0001_S004*

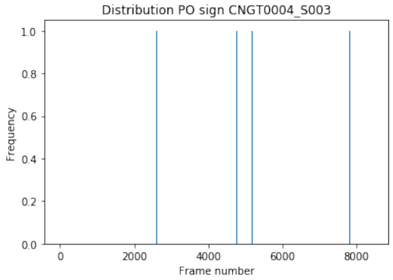

Figure 5: Distribution PO sign *CNGT0004_S003*

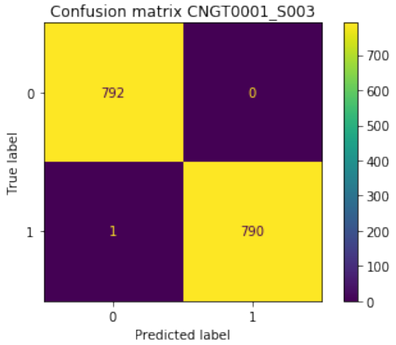

Figure 6: Confusion matrix binary classification of *CNGT0001_S003* dataset.

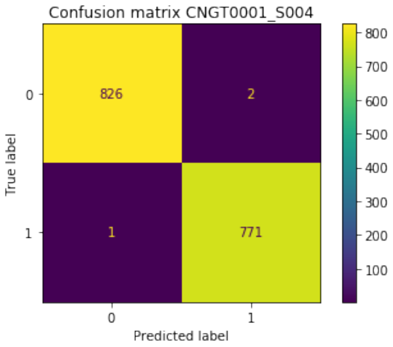

Figure 7: Confusion matrix binary classification of *CNGT0001_S004* dataset.

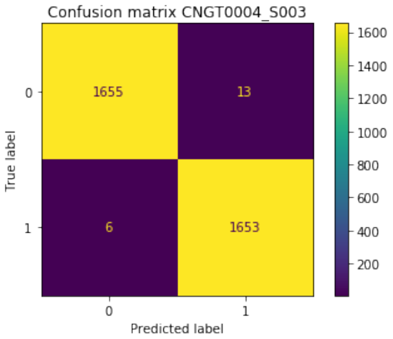

Figure 8: Confusion matrix binary classification of *CNGT0004_S003* dataset.

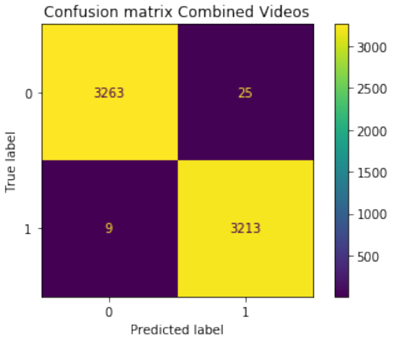

Figure 9: Confusion matrix binary classification of combined dataset.

Dataset	F1-score
CNGT0001_S003	0.9987
CNGT0001_S003	0.9981
CNGT0001_S003	0.9976
Combined	0.9947

Table 3: F1-score for classification of four datasets.

As we can see from the results, the F1-score of all four datasets is quite high. The highest score is obtained with the data of *CNGT001_S003*. However, the differences are really small and in general, for each of the four classification, a minimum F1-score of 0.99 is achieved.

4.1 Conclusion and Discussion

This research includes the binary classification of the PO sign by means of video processing. Features for classification includes features extracted from the VGG19 model and *Open Pose* keypoints. With these features, a DCT classifier is created and four classifications are performed with optimal depth for each classification.

The results have shown that for each classification, a minimum F1-score of 0.99 is achieved. We thus can state that the results are really satisfying for the purpose of our research.

Looking at the distribution of the PO sign in the used subset of the NGT Corpus, we don't see a similar occurrence of the sign as opposed to what the literature denotes [5] [11]. However, with (only) three videos, the subset is not representative for the whole corpus and thus this cannot be seen as a hard conclusion.

For further research, we advise to extent the research to a bigger subset of the NGT Corpus and look further into the distribution of the PO sign. Furthermore, we can state that - based on our results - the VGG19 model in combination with *Open Pose* keypoints create a sufficient feature subset for binary classification of the PO sign. For further research, these feature extraction methods can be elaborated also on a bigger subset of the NGT Corpus.

REFERENCES

[1] n.d. Centre for Language Studies. https://www.ru.nl/cls/our-research/research-groups/sign-language-linguistics/ Accessed on: 2020-11-26.

[2] 4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies 2010. *Glossing a multi-purpose sign language corpus.* 4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies, Language Resource and Evlauation Conference (LREC).

[3] N. Adaloglou, T. Chatzis, I. Papastratis, A. Stergioulas, G. Papadopulos, V. Zacharopulou, G J. Xydopoulous, K. Atzakas, D. Papazachariou, and P. Daras. 2020. A Comprehensive Study on Sign Language Recognition Methods. (7 2020). arXiv:2007.12530.

[4] N.V. Chawla, K.W. Bowyer, L.O. Hall, and W.P. Kegelmeyer. 2002. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research* 16 (2002). doi: https://doi.org/10.1613/jair.953.

[5] M. Chu, A. Meyer, L. Foulkes, and S. Kita. 2014. Individual Differences in Frequency and Saliency of Speech-Accompanying Gestures: The Role of Cognitive Abilities and Empathy. *Journal of Experimental Psychology: General* 143, 2 (2014), pp. 694–709. doi: 10.1037/a0033861.

[6] K. Cooperrider, N. Abner, and S. Glodin-Meadow. [n.d.]. The Palm-Up Puzzle: Meanings and Origins of a Widespread Form in Gesture and Sign. ([n. d.]).

[7] O. Crasbon and I. Zwitserlood. n.d. The Corpus NGT. https://www.ru.nl/corpusngt/de_filmpjes/download-filmpje/ Accessed on: 2020-11-26.

[8] O. Crasborn and E. van der Kooij. 2013. The phonology of focus in Sign Language of the Netherlands. *Journal of Linguistics* 1, 3 (2013), pp. 1–51. doi: 10.1017/S0022226713000054.

[9] G. Hidalgo, Z. Cao, T. Simon, S. Wei, H. Joo, and Y. Sheikh. 2020. CMU-Perceptual-Computing-Lab - Open Pose. https://github.com/CMU-Perceptual-Computing-Lab/openpose GitHub repository.

[10] N. Ibrahim, H. Zayed, and M. Selim. [n.d.]. Advances, Challenges, and Opportunities in Continuous Sign Language Recognition. *Journal of Engineering and Applied Sciences* 15, 5 ([n. d.]), 1205–1227. doi: jeasci.2020.1205.1227.

[11] T. Johnston. 2012. Lexical Frequency in Sign Languages. *The Journal of Deaf Studies and Deaf Education* 17, 2 (2012), pp. 163–193. doi: 10.1093/deafed/enr036.

[12] S. Ko, C. Kim, Jung., and C. Cho. 2019. Neural Sign Language Translation based on Human Keypoint Estimation. (6 2019). arXiv:1811.11436.

[13] S. Lang, M. Block, and R. Roja. 2012. Sign Language Recognition Using Kinect, Vol. 7267. International Conference on Artificial Intelligence and Soft Computing, Lecture Notes in Computer Science, pp. 394–402. doi: 10.1007/978-3-642-29347-4_46.

[14] C. Lucas and R. Bayley. 2005. Variation in ASL: The Role of Grammatical Function. *Sign Language Studies* 6, 1 (2005), pp. 38–75.

[15] R L. McKee and S. Wallingford. 2011. "So, well, whatever": discourse functions of palm-up in New Zealand Sign Language. *Sign Language Linguistics* 14 (2011), pp. 213–247. doi: 0.1075/sll.14.2.01mck.

[16] National Institute on Deafness and Other Communication Disorders (NIDCD. 2019. American Sign Language. https://www.nidcd.nih.gov/health/american-sign-language Accessed on: 26-10-2020.

[17] O. Russakovsky, J. Hung, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. Berg, and L. Fei-Fei. 2014. ImageNet Large Scale Visual Recognition Challenge. doi: 10.1007/s11263-015-0816-y.

[18] A. Schembri, K. Cormier, T.A. Johnston, and D. Mckee. 2010. Sociolinguistic variation in British, Australian and New Zealand Sign Languages. In *Sign Languages*, D. Brentari (Ed.). Cambridge University Press, Chapter 21, pp. 476–498. doi: 10.1017/CBO9780511712203.022.

[19] A. Schembri and T. Johnston. 2012. Sociolinguistic aspects of variation and change. In *Sign Language*, R. Pfau, M. Steinbach, and B. Woll (Eds.). De Gruyter Mouton, Chapter 33, pp. 788—-816. doi: 10.1515/9783110261325.788.

[20] P. Schlenker. 2018. Sign Language Semantics: Problems and Prospects. *Theoretical Linguistics* 44, 3-4 (11 2018). doi: 10.1515/tl-2018-0022.

[21] K. Simonyan and A. Zisserman. 2015. Very deep convolutional networks for large-scale image recognition.

[22] P.H. Swan and H. Hauska. 1977. The decision tree classifier: Design and potential. *IEEE Transactions on Geoscience Electronics* 15, 3 (1977), 142–147. doi: 10.1109/TGE.1977.6498972.

[23] The Language Archive. n.d.. https://archive.mpi.nl/tla/elan Accessed on: 2020-11-26.

[24] Y. Zheng, C. Yang, and A. Merkulov. 2018. Breast cancer screening using convolutional neural network and follow-up digital mammography. In *Proceedings Volume 10669, Computational Imaging III.* doi: 10.1117/12.2304564.