Polish Sign Language Gestures to Text Conversion Using Machine Learning

Łukasz Lemieszewski
The Jacob of Paradies University, Department of Technology, Gorzów Wielkopolski, Poland *llemieszewski@ajp.edu.pl*

Marek Nowak
PH03NIX Software
Gorzów Wielkopolski, Polska *marek.n6666@gmail.com*

Jarosław Becker
The Jacob of Paradies University, Department of Technology, Gorzów Wielkopolski, Poland *jbecker@ajp.edu.pl*

Abstract

There are around 50-100 thousand deaf people in Poland, their main language is Polish sign language. It can be challenging for them to communicate with the rest of society and there is a gap in Polish sign language gestures to text conversion. Although some research has been done before, no research paper or product solves this problem.

The primary objective is to develop a concept of an intelligent application that can convert Polish sign language from either a video or a live feed. To achieve this, research was conducted on other sign languages, which helped in selecting the most promising hybrid models of deep neural networks. Subsequently, tests were conducted and the best model was chosen. Finally, the best model was trained on the dataset of Polish sign language, using weights (transfer learning) trained on the MS American Sign Language dataset.

Keywords: Machine Learning (ML), Convolution Networks (CN), Sequence data classification, Sign Language (SL).

1. Introduction

Deaf individuals are a significant group in Polish society - according to a 2014 report on the situation of deaf people in Poland [4], their number is estimated at 50,000-100,000. One of the challenges they face is communication. They use Polish sign language to communicate on a daily basis, but the problem is that very few people in Polish society are proficient in this language.

According to the collective work of the 2020 Commission of Experts on Deaf People [1], the supposed number of speakers of this language is 30-50 thousand people, and according to the previously mentioned report [4], this number is estimated at a value in the range of 50-100 thousand. This means that these people are not able to communicate with the majority of Polish society using this language.

The purpose of this paper is to facilitate the process of communicating between these people and those unfamiliar with the language, to investigate which model of hybrid deep neural network architecture is the best to solve the problem and to create an application that will make it easy for non-technical individuals to use the model.

In order to achieve the desired outcome, three models were selected and trained on a subset of the MS-ASL dataset. A test was then conducted to determine the best solution, which was subsequently trained on a larger subset of the dataset. The pretrained weights of this model were later utilized for transfer learning to improve the accuracy of the model for Polish sign language. For the purpose of further training of this model, a set of selected Polish Sign Language words was created. To simplify the use of the model, an application with a graphic user interface was developed. This application allows users to make predictions from selected

mp4 files or live video feeds.

1.1. Related work

After searching through various sources, no solution was found to the problem of translating Polish sign language into text. One of found attempts related to this topic is the hear.ai project, but it seems that the project never reached the gesture-to-text conversion phase. Instead, it only developed a tool to convert HamNoSys annotations to numerical labels for specific initial body features and hand positions. Details are described in work [11]. Currently, the website associated with the project - https://www.hearai.pl does not exist (the domain is up for sale on the Internet domain exchange) and in the source code repository [15] the last change was on 19.12.2022 (as of 22.03.2024).

In 2023, a paper was published describing an application to learn Polish Sign Language [14] which focused on developing a symbol recognition system using machine learning with selected letters, auxiliary characters and digits rather than videos.

A dataset of labeled Polish Sign Language gestures was not found, while numerous collections of American Sign Language gestures exist, including MS-ASL.

Analogous solutions for other sign languages, such as American Sign Language and German Sign Language, can be divided into two categories:

- translating only gestures that represent single letters of the manual alphabet, which are then used to create sentences,
- translating complex gestures, which represent whole words (such as "hi", "you", "good", and so on). This category can be further divided into two subcategories: translating isolated gestures and translating gesture while taking context into consideration. A single gesture can have many meanings depending on the context in which it is used.

Examples of the first approach can be found in [17], articles realizing sign language to text conversion in real time using transfer learning and [6] presenting the results real-time American Sign Language recognition with convolutional neural networks and also projects: Sign Language to Text Conversion [7], Unvoiced [12].

Second approach can be found in article [8], which describes MS-ASL – a large-scale data set and benchmark for understanding American Sign Language. It presents a comparative analysis of various solutions to a problem by evaluating their average top-five accuracy for the MS-ASL dataset (marked in Table 1 as ASL and divided into four groups whose name refers to the number of classes).

Results reported in Table 1 indicate that only I3D model (*Two-Stream Inflated 3D ConvNet*, presented in [3]) achieved promising results with a top-five accuracy of 81.08% for a subset comprising 1000 classes. The other examined solutions achieved significantly worse results:

- the Hierarchical Co-occurrence Network (HCN) model, which utilizes the coordinates of 137 points on the human body to make predictions, demonstrated a performance of only 32.50% for the ASL1000 dataset,
- model VGG (Visual Geometry Group) + LSTM (Long Short-Term Memory) achieved a top-five accuracy of only 5.86% for the ASL500 set (no data for ASL1000),
- naive Classifier for ASL1000 has achieved a top-five accuracy of 0.58%.

Table 1. Average per-class top-five accuracy.

Method	ASL100 [%]	ASL200 [%]	ASL500 [%]	ASL1000 [%]
Naive Classifier	4.86	2.49	1.05	0.58
VGG+LSTM	33.42	21.21	5.86	-
HCN	73.98	60.29	43.86	32.50
I3D	95.16	93.79	89.80	81.08

Promising results were also obtained in [2] describing American Sign Language recognition using deep learning and computer vision, where a set of 150 classes with five examples per class was used to train a model consisting of an Inception convolutional network [16] and

a recurrent network consisting of LSTM cells. The results were (depending on the output in the convolutional model whose value is passed to the recursive network) 91% accuracy with the SoftMax Layer and 55% accuracy with the Pool Layer.

An extensive summary and overview of current (at least for the 2020 publication year) developments in the problem area is contained in [13] describing research on sign language recognition using a deep survey.

1.2. Models under consideration

Three hybrid neural network models were developed to compare their results (output) and select the best one. It was important to test a range of possible solutions in order to evaluate the results of already conducted studies and get the idea of the best architecture for the problem. The focus was on creation and comparison of quite diverse models to get overall idea, which architecture might be the best and open the way for more detailed research on most promising solution instead of focusing on details from the beginning.

Model A was based on the architecture specified in the article [5] focusing on training a neural network with an image sequence. The chosen model requires the use of a 2D convolutional network, and it was decided to choose a model from the "Keras Applications" collection. These are deep neural networks tested on the ImageNet collection, which comprises approximately 14 million images. The MobileNet model was chosen because of its relatively high accuracy (top-1: 70.4%, top-5: 89.5% [10]) while maintaining high performance (3.4 ms per inference step for GPU [10]).

This network is placed in the TimeDistributed layer, which allows to apply a layer (in this scenario MobileNet model) to every selected frame of a video. Results of this wrapper are passed to the recurrent network – composed from Gated recurrent units (GRUs). The result is then fed into a decision network, which comprises of three Dense layers, which are separated by Dropout layers that prevent over-fitting and one Dense layer with a SoftMax activation function returning the probabilities of matching a class to an example. The model A is shown in Figure 1A.

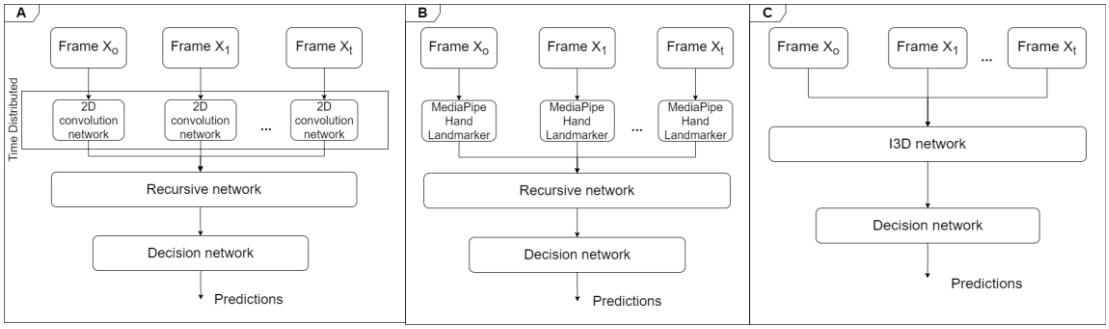

Fig. 1. Diagram of model A, model B and model C.

This solution identifies the coordinates of various points on the hand, such as the wrist, fingertips, etc. It supports up to two hands and the outcomes of this network are coordinates for 21 points of each hand (LH, RH), each consisting of x and y coordinates.

Results produced by this network for each selected frame are then transferred to a recurrent network, which is composed from GRUs. Output of these layers are transferred to a decision network that looks analogous to the one in model A - three Dense layers, separated by Dropout layers and one Dense layer with a SoftMax activation function. This model B is shown in Figure 1b.

Model C uses an I3D model described in [3] presenting a new model and the Kinetics dataset pre-trained on ImageNet and Kinetics collections. The outcome is fed into a decision network consisting of a 3D convolution network and a SoftMax activation layer. A diagram of the model C is shown in Figure 1c.

Although the models mentioned above were created based on the basis of existing research on sign language translation, as stated in section 1.1, there are currently no studies on developing a solution (consisting of a translation model and application) for Polish sign

language. Therefore, it can be assumed that this study on Polish sign language is one of the first of its kind. Meanwhile, the evaluation of models for translating American sign language expands current knowledge and verifies the results presented in other papers.

2. Models study

In order to determine the best model, it was necessary to conduct a study that involved training each model, evaluating the results, and comparing them. This process enabled the identification of the most effective model, based on its top 1 and top 5 accuracy in relation to the others.

2.1. Data set

MS-ASL data set was used to train and evaluate the results of models. It's a dataset detailed in [8] which describes the use of a large-scale dataset and benchmark to understand American Sign Language. In this set videos are hosted on a YouTube service and in shared files there are only hyperlinks to them (with details like the timeframe of the gesture). Unfortunately, lots of videos were unavailable (private or deleted) so the final data set contained 978 classes and 11642, 3359, and 2682 videos with gestures for training, validation and test.

2.2. Preprocessing

Before the training phase, each video was processed by SSD (Single Shot Detector) network (to be precise ssd-mobilenet-v2 was used). The network was used to crop the video to the bounding box inside which the person was present.

In addition, during training, 64 frames were selected with a random number of frames skipped between each (0 to 2). If the output was not long enough (it ends before 64 frames were selected), the process was repeated from the beginning, adding selected frames to already existing output (so basically video was looped). After this, the obtained collection of frames was randomly flipped and rotated (-10 to 10 degrees).

Each frame was also normalized. For model A and B, each pixel was normalized to be in a range of -1 to 1. For model B, preprocessing was built into the MediaPipe package. For models A and C the frame size was additionally changed to 224x224.

2.3. Models of training and results

To speed up the training process only classes with more than 30 examples were used. Ended up with 38 classes and 1262, 342 and 184 for train, validation and test – this set is later called 30 plus.

Each model was trained for 20 epochs with Stochastic gradient descent (SGD) optimizer and momentum 0.9. After that evaluation was performed on the test set. Two metrics were used: top 1 accuracy and top 5 accuracy [9]. Results are reported in Table 2. These two metrics allow for a more nuanced evaluation of the model. "Accuracy top 1" is a strict metric that requires the model to precisely hit the exact category. On the other hand, "Accuracy top 5" provides a more forgiving assessment, accepting that for some tasks, especially those with many similar categories, simply finding the correct category among the top five predictions may be good enough. Utilizing both of these metrics gives a fuller picture of the model's classification abilities and helps in adjusting the training process to achieve better results.

Table 2. Accuracy of the considered models for the test set.

Model	Model name	Accuracy top 1 [%]	Accuracy top 5 [%]
A	Mobilenet+GRU	3.25	13.19
B	Mediapipe+GRU	22.67	58.34
C	I3D	68.57	89.14

Results suggest that the best suited to predicting highly connected sequential data in the form of images is model C based on the I3D network. The results of model B are also promising

but falling behind I3D. Maybe with more data than only points on hands (hand gesture plays a dominant role in sign language but there is also information in body, head, mount, eye movement and position), the model it could achieve better results. Model A is clearly the worst one with only 3.25% for top 1 accuracy and does not look like there is anything that could make it relevant in this field of problem.

2.4. Results of training I3D based model on larger subsets

As the model C based on I3D achieved the best results it was selected for further investigation. It was additionally trained for 40 epochs on mentioned in 3.1 set of 978 classes (from now on called full set) and a set called 20 plus where the number of examples per class was a minimum of 20 - this resulted in a subset of 142 classes with 3741, 1045, 642 examples for training, validation and test. Results are reported in Table 3.

The more classes are presented in the subset the worse accuracy gets. It is worth pointing out that larger sets are unbalanced, in full set examples per class range from 2 up to 39. It probably has an impact on a result as standard deviation is more than 70% higher for the full set than for 30 plus set. There is also possibility that this model is not good enough to handle larger amounts of classes.

Table 3. Results of training model C based on I3D evaluated on test set.

Subset name	Number of classes	Accuracy top 1 [%]	Accuracy top-5 [%]	Mean top-1 accuracy [%]	Standard deviation top-1
30 plus	38	78.81	96.73	79.21	21.05
20 plus	142	65.24	87.83	62.31	25.48
Full	978	46.53	71.70	42.88	36.24

Comparing the set of 20% of classes with the worst accuracy (for testing) to the set of 20% of classes with the fewest examples (for training), there is 44 percent overlap for the full set, 32% for the 20-plus set and 0% overlap for the 30-plus set. Based on this, it can be concluded that the number of examples for a class affects its prediction results, but most likely the complexity of the gesture and similarity to other classes also play a role here.

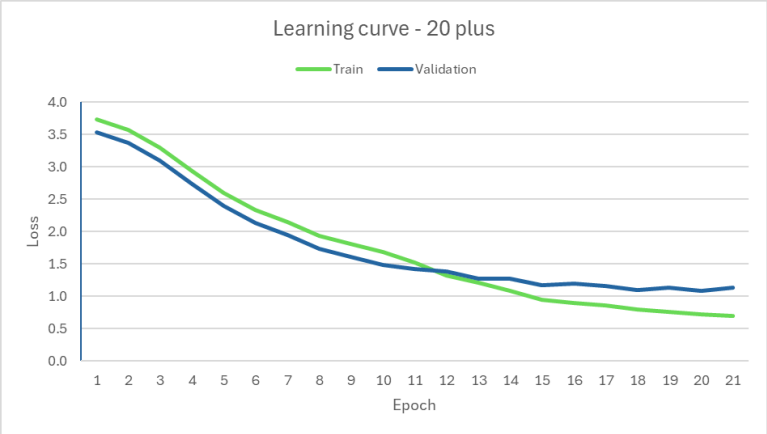

Fig. 2. Learning curve for 20 plus subset

As can be seen in Figure 2 around epoch 12 for subset 20 plus model starts overfitting - loss of validation set is not decreasing, while training loss is falling down.

3. Creating solution for Polish Sign Language

The problem with Polish Sign Language in this context is lack of labeled data set of gestures. The newly created dataset is small in comparison to MS-ASL and contains only

15 classes and 151, 63 and 25 examples for training, validation and test subsets. Distribution of examples between classes and collections is presented in Table 4.

Table 4. Distribution of examples between classes and collections for polish sing language data set.

Polish Sign Language gesture (English translation)	Number of examples for the training set	Number of examples for the validation set	Number of examples for the test set
Co (What)	6	3	2
Cześć (Hi)	20	8	3
Dobranoc (Goodnight)	9	3	1
Dobrze (Good)	6	3	1
Do widzenia (Goodbye)	18	7	3
Dziękuje (Thank you)	9	4	1
Dzień dobry (Good morning)	22	9	3
Gdzie (Where)	10	4	2
Ja (I)	9	4	2
Nie (No)	5	2	1
On (He)	7	3	1
Proszę (Please)	8	3	1
Przepraszam (Sorry)	8	4	1
Tak (Yes)	7	3	2
Ty (You)	8	3	1

Considering this, transfer learning was used to prevent over-fitting and handle the possibility of a lack of examples needed to effectively train the.

3.1. Results

Weights received from learning model on 20 plus subset were used for transfer learning. Only decision network was overwritten with random weights. Other layers were frozen, except 0, 5 and 10 last layers (counting from the layers behind the decision network) which were unfrozen.

Results of training are reported in table 5. In addition to the top 1 and top 5 accuracy, the table also includes accuracy and loss for validation set, since the test subset is small in this dataset and for this reason the results may not be reliable.

Table 5. Results obtained by the model depending on the number of unfrozen layers.

Number of unfrozen layers	Accuracy top 1 [%]	Accuracy top 5 [%]	Accuracy for validation subset [%]	Loss for validation subset
0	80.00	92.00	81.00	0.68
5	84.00	88.00	83.00	0.68
10	76.00	88.00	76.00	0.91

The results show that unfreezing 5 layers has a small impact on accuracy. Top 1 accuracy is better by 4% points for top 1 in comparison to 0 layers but gets worse result for top 5 accuracy. It might suggest that those layers weights overfit by a bit. For 10 unfrozen layers the result for top 1 accuracy was much worse – only 76% and loss for validation subset was higher than for 0 and 5 unfrozen layers. It suggests that for this data set no more than around 5 layers can be unfrozen, because for 10 unfrozen layers accuracy and loss was getting worse and the “knowledge” transferred was fading.

4. Application

In order to allow an easy and convenient way to use trained model an application was created.

This is a desktop application created with Tkinter library and supports Linux operating system. It contains an implemented preprocessing algorithm, which returns data compatible with a model. Algorithm is similar to the one described in section 2.2, 64 frames are selected from the provided collection, each is clipped to the bounding box obtained from the SSD network and the pixel values are brought to a value between -1 and 1.

The user only needs to provide input data. Providing it is also simplified, application allows to select any mp4 file with single or multiple gestures through a friendly dialog box. The selected video and predicted class are then displayed.

Fig. 3. One of the application windows (6 different recognized sign language gestures).

Additionally, there is an option to record new gesture or use live feed (e.g. from webcam) to get predictions. Also in the main menu is an option called "Showcase" which uses a test subset to show all videos and predictions. The user can navigate between them using the "next" or "previous" button. Examples of this window is presented in a Figure 3 (this one support recognizing gestures based on live video feed - 6 different recognized sign language gestures).

5. Conclusion

There is currently no automated solution to help the deaf community in Poland to communicate with the rest of society. This makes it difficult for them to handle daily affairs such as meetings at the bank without the assistance of a translator, which can be expensive and not always available.

One possible solution is to develop translation software for Polish sign language. There hasn't been much research devoted specifically to this area, but some helpful insight is present in the field of translation of other sign languages. Based on that and the research conducted in this paper, the most promising is the use of an I3D model, which is effective in predicting accurately hundreds of classes.

One of the major challenges faced in creating a functional solution for Polish sign language is the unavailability of a sufficient dataset. The dataset that was created for the purpose of this research was only adequate to develop a proof-of-concept application and requires significant expansion to create a fully functional solution. Despite this limitation, the research conducted, and the development of this solution provides a solid foundation and direction for future research.

Future research should focus on expanding the dataset of gestures and further development of the application. Creating versions for mobile systems like Android and IOS, as well as web browsers, will make the application more accessible. Additionally, moving the prediction

process to the cloud will speed up predictions on devices with low processing power and simplify the process of updating the model. This means that adding support for new gestures would not require any action from the user, such as updating the application.

There is still room for improvement in the model, and using other 3D convolutional network architectures specifically tailored to the problem may lead to better results. It's worth exploring other architecture options in the pursuit of optimal outcomes.

References

1. A collective work of the Expert Committee on the Deaf: Osoby gluche w Polsce 2020. Wyzwania i rekomendacje (2020), https://bip.brpo.gov.pl/sites/default/files/Osoby_Gluche_w_Polsce_2020_Wyzwania_i_Rekomendacje.pdf, Accessed October 2, 2023
2. Bantupalli K., Xie Y.: American Sign Language Recognition using Deep Learning and Computer Vision. In: IEEE International Conference on Big Data (Big Data), pp. 4896-4899. IEEE (2018)
3. Carreira J., Zisserman A.: Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 6299-6308. IEEE (2017)
4. Expert Committee on Persons with Disabilities and the g/Deaf Panel of the Ombudsman.: Sytuacja osob gluchych w Polsce (2014), https://bip.brpo.gov.pl/sites/default/files/Raport_Sytuacja_osob_poz%203_srodki_2%20XII.pdf, Accessed October 2, 2023
5. Ferlet P.: Training a Neural Network with an Image Sequence — example with a video as input (2019), https://medium.com/smileinnovation/training-neural-network-with-image-sequence-an-example-with-video-as-input-c3407f7a0b0f, Accessed October 18, 2023
6. Garcia B., Viesca A. S.: Real-time American Sign Language Recognition with Convolutional Neural Networks. Convolutional Neural Networks for Visual Recognition, 2.225-232: 8. (2016)
7. Gupta N.: Source code repository “Sign Language to Text Conversion”, https://github.com/emnikhil/Sign-Language-To-Text-Conversion, Accessed October 28, 2023
8. Joze, H. R. V., Koller, O.: MS-ASL: A Large-scale Data Set and Benchmark for Understanding American Sign Language. arXiv preprint arXiv:1812.01053, (2018).
9. Kandel I., Castelli M.: Transfer Learning with Convolutional Neural Networks for Diabetic Retinopathy Image Classification. A Review. Applied Sciences (2020)
10. Keras: Keras Applications, https://keras.io/api/applications/#usage-examples-for-image-classification-models, Accessed October 18, 2023
11. Majchrowska S., Plantykow M., Olech M.: Handling sign language transcription system with the computer-friendly numerical multilabels. arXiv preprint arXiv:2204.06924. (2022)
12. Nagaraj A.: Source code repository “Unvoiced”, https://github.com/grassknoted/Unvoiced, Accessed October 28, 2023
13. Rastgoo, R., Kiani, K., & Escalera, S.: Sign language recognition: A deep survey. Expert Systems with Applications. Expert Systems with Applications, 164, 113794 (2021)
14. Slian, A., Czajkowska J., Bugdol M.: Mobile Application for Learning Polish Sign Language. In: Polish Conference on Biocybernetics and Biomedical Engineering. Cham: Springer Nature Switzerland, pp. 95-104. Springer (2023)
15. Source code repository HearAI, https://github.com/hearai/hearai, Accessed October 28, 2023
16. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S.E., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition, pp.1-9. IEEE (2015)
17. Thakar S., Shah S., Shah B., Nimkar A. V.: Sign Language to Text Conversion in Real Time using Transfer Learning. In: IEEE 3rd Global Conference for Advancement in Technology (GCAT), IEEE (2022)