Sign Classification in Sign Language Corpora with Deep Neural Networks

Lionel Pigou, Mieke Van Herreweghe, Joni Dambre
Ghent University
{lionel.pigou, mieke.vanherreweghe, joni.dambre}@ugent.be

Abstract

Automatic and unconstrained sign language recognition (SLR) in image sequences remains a challenging problem. The variety of signers, backgrounds, sign executions and signer positions makes the development of SLR systems very challenging. Current methods try to alleviate this complexity by extracting engineered features to detect hand shapes, hand trajectories and facial expressions as an intermediate step for SLR. Our goal is to approach SLR based on feature learning rather than feature engineering. We tackle SLR using the recent advances in the domain of deep learning with deep neural networks. The problem is approached by classifying isolated signs from the Corpus VGT (Flemish Sign Language Corpus) and the Corpus NGT (Dutch Sign Language Corpus). Furthermore, we investigate cross-domain feature learning to boost the performance to cope with the fewer Corpus VGT annotations.

Keywords: sign language recognition, deep learning, neural networks

1. Introduction

SLR systems have many different use cases: corpus annotation, in hospitals, as a personal sign language learning assistant or translating daily conversations between signers and non-signers to name a few. Unfortunately, unconstrained SLR remains a big challenge. Sign language uses multiple communication channels in parallel with high visible intra-sign and low inter-sign variability compared to common classification tasks. In addition, publicly available annotated corpora are scarce and not intended for building classifiers in the first place.

A common approach in SLR is to get around the high dimensionality of image-based data by engineering features to detect joint trajectories (Charles et al., 2013), facial expressions (Liu et al., 2014) and hand shapes (Ong and Bowden, 2004) as an intermediate step. Data gloves (Oz and Leu, 2011), colored gloves (Wang and Popović, 2009) or depth cameras (Chai et al., 2013) are often deployed in order to obtain a reasonable identification accuracy.

In recent years, deep neural networks achieve state-of-the-art performance in many research domains including image classification (Szegedy et al., 2014), speech recognition (Graves et al., 2013) and human pose estimation (Pfister et al., 2014). The deep learning models that we use in this work are based on convolutional neural networks (CNNs) (Lecun et al., 1998). A CNN is a model with many parameters that are adjusted iteratively using optimization algorithms (= *learning*) and a large amount of annotated data.

In previous work (Pigou et al., 2015), we showed that deep neural networks are very successful for gesture recognition and gesture spotting in spatiotemporal data. Our developed system is able to recognize 20 different Italian gestures (i.e., emblems). We achieved a classification accuracy of 97.23% in the *Chalearn 2014 Looking At People* gesture spotting challenge (Escalera et al., 2014). This gives us an indication that deep neural networks can be useful for SLR.

In this work, the problem is approached by classifying isolated signs from the Corpus VGT (Van Herreweghe et al., 2015), the Flemish Sign Language Corpus, and the Corpus NGT (Crasborn et al., 2008; Crasborn and Zwitserlood, 2008), the Dutch Sign Language Corpus. Furthermore, we investigate cross-domain feature learning to boost the performance to cope with the fewer Corpus VGT annotations.

2. Methodology

2.1. Data

The two corpora used to explore SLR (Corpus VGT and Corpus NGT) have similar camera setups and use very similar gloss annotation rules with identical software (ELAN). Both corpora consist of Deaf signers that perform tasks such as retelling comic strips, discuss an event and debating on chosen topics. For each corpus, the 100 most frequently used signs are extracted together with their gloss. The data is split into three sets: 70% training set, 20% test set and 10% validation set. The training set is used to optimize the

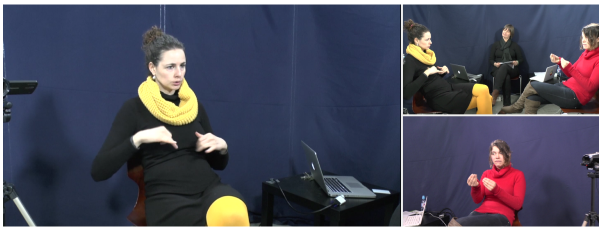

Figure 1: A sample from the Corpus VGT (Ghent University), filmed from three viewpoints.

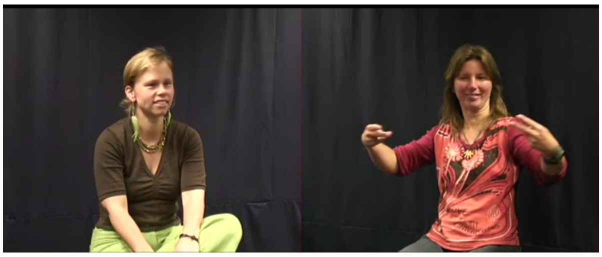

Figure 2: A sample from the Corpus NGT (Radboud University Nijmegen), filmed from two viewpoints.

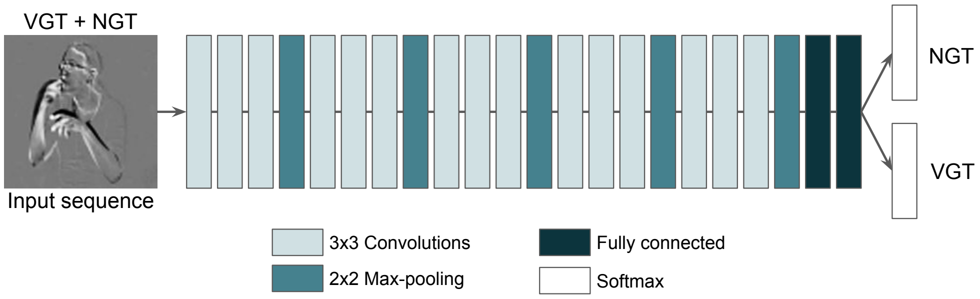

Figure 3: The architecture overview of the deep neural network used in this work. All layers are shared among corpora, except for the softmax classifier. This will boost the performance for the Corpus VGT, as it learns better features using the Corpus NGT with more annotations.

neural networks, the validation set is used for evaluation during training and the test set is used to evaluate the final models.

The Corpus VGT (Figure 1) uses Flemish Sign Language. The project started in Juli 2012 and ended in November 2015 at Ghent University, in collaboration with the Linguistics Group VGT of KU Leuven Campus Antwerp, and promoted by Prof. Dr. Mieke Van Herreweghe (Ghent University) and Prof. Dr. Myriam Vermeerbergen (KU Leuven Campus Antwerp). The corpus contains 140 hours of video and a small fraction is annotated. After cleaning the data, we extracted a total of 12599 video-gloss pairs from 53 different Deaf signers.

The Corpus NGT (Figure 2) contains Deaf signers using Dutch Sign Language from the Netherlands. This project was executed by the sign language group at the Radboud University Nijmegen. Every narrative or discussion fragment forms a clip of its own, with more than 2000 clips. We extracted a total of 55224 video-gloss pairs from 78 different Deaf signers.

As Figure 4 shows, there is a class imbalance for both corpora. This means that accuracy measures will be highly skewed. For example, only predicting the most common sign (which is “ME”) for every sample across the whole dataset already results in 30.9% and 11.2% accuracy for the Corpus NGT and the Corpus VGT respectively.

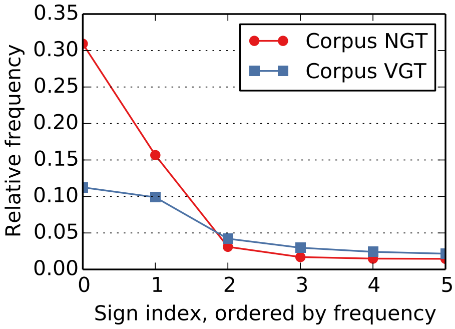

Figure 4: The relative frequency for the five most common signs in both corpora. The class imbalance is significant in both corpora, but is especially prevalent for the Corpus NGT.

2.2. Convolutional Neural Network (CNN)

CNNs are models that allow to learn a hierarchy of layered features instead of manually extracting them. They are among the most successful techniques in deep learning, a domain in machine learning that has proven to be very successful at recognizing patterns in high dimensional data such as images, videos and audio. These artificial networks are inspired by the visual cortex of the human brain. The neurons in a CNN will connect to a local region of the image, called a receptive field. This is accomplished by performing discrete convolutions on the image with filter values as trainable weights, which are optimized using the *gradient descent* algorithm. A second important building block in a CNN is a pooling scheme, where only the interesting information of the feature maps is pooled together.

These base operations are performed in multiple layers as illustrated in Figure 3. This architecture is inspired by (Simonyan and Zisserman, 2014). Three convolutional layers are stacked before performing max-pooling (only the maximum activation of each region remains) on non-overlapping 2x2 spatial regions. The input image sequence consists of 8 frames of size 128x128. Each frame is subtracted from the previous frame to remove static information. These frames are rotated, shifted and stretched randomly during training to artificially increase the amount of data in order to learn more generalized features. This technique is called *data augmentation*.

3. Results

3.1. Corpus NGT

The resulting model, with the highest score on the validation set, is illustrated in Figure 3 (without the VGT branch). The shorthand notation of the full architecture is as follows: C^3_{32}-P-C^3_{64}-P-C^3_{128}-P-C^3_{256}-P-C^3_{512}-P-D_{2048}-D_{2048}-S, where C^a_b denotes a stacked convolutional layers with b feature maps and 3x3 filters, P a max-pooling layer with 2x2 pooling regions, D_c a fully connected layer with c units and S a softmax classifier.

The top-N accuracy is a measure indicating the probability that the correct answer is within the model’s N best guesses. The top-N accuracies of the test set for the Corpus NGT are depicted in Figure 5. The CNN achieves a top-1, top-3 and top-5 accuracy of 56.2%, 75.7% and 82.1% respectively for 100 signs. This is especially interesting for automatic corpus

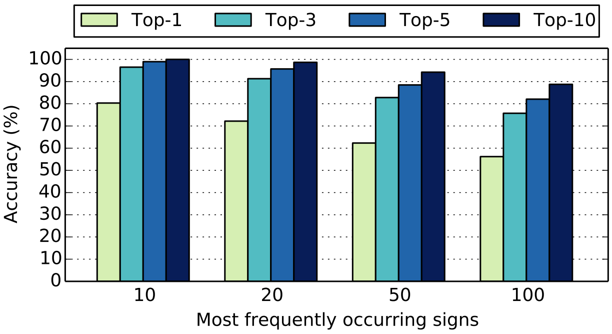

Figure 5: **Corpus NGT** top-N accuracies. A measure indicating the probability of the correct answer being within the model's N best guesses.

annotation, where providing a list with the N best guesses is appropriate.

As mentioned above, we have to keep in mind the class imbalance. The confusion matrix shows the fraction of true positives for each class (each sign) on the diagonal. It also tells us which classes it gets confused with. To have a better insight into the model's performance, we show the confusion matrix in Figure 6. Not surprisingly, almost all classes get confused with frequently occurring ones. The CNN learned to bet on common glosses when it is unsure about a certain input, because more often than not it will get rewarded for that. Other misclassification is due to signs that are hard to distinguish from each other.

3.2. Corpus VGT

To cope with the smaller amount of annotations for the Corpus VGT compared to the Corpus NGT, we train a shared model on both corpora (Figure 3). This cross-domain learning is a form of *transfer learning*, where the knowledge of one or more domains (in this case the Corpus NGT) is useful for other domains. Our motivation is that the learned features for both domains should be similar, except for the softmax classifier. All sign languages have similar visual features: they consist of hand, arm, face and body expressions. We hope to capture these generic building blocks in order to boost the performance for the Corpus VGT.

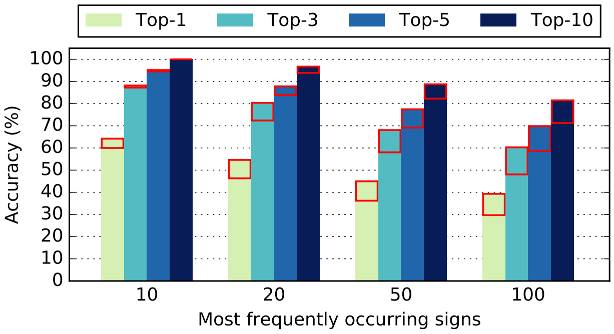

Figure 7: **Corpus VGT** top-N accuracies with cross-domain learned features. The red outline shows the improvement compared to the accuracies without cross-domain learning.

In Figure 7, the top-N accuracies are shown. It achieves a top-1, top-3 and top-5 accuracy of 39.3%, 60.3% and 69.9% respectively for 100 signs. To show the improvement using the cross-domain learning, the sensitivity (true positive rate) increase for each class is depicted in Figure 9. We clearly see a significant improvement for most signs, but a few classes are negatively affected by it. The resulting confusion matrix is shown in Figure 8. The errors are more spread out than the ones for the Corpus NGT, because the class imbalance is less prevalent.

4. Conclusion and Future Work

We show that CNNs are capable of learning features from image sequences across linguistic sign language corpora.

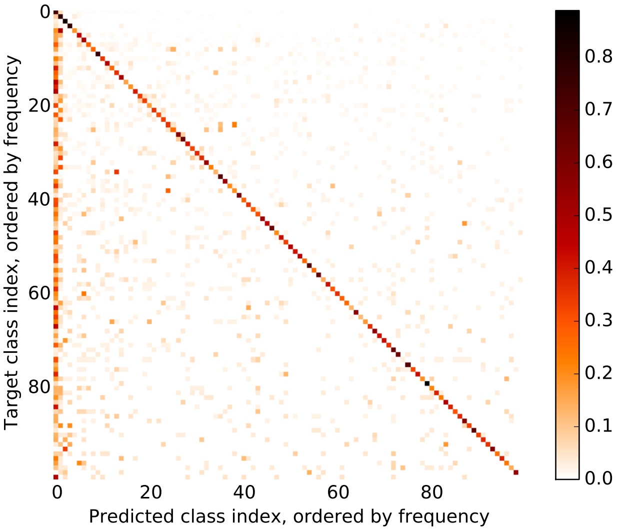

Figure 6: **Corpus NGT** confusion matrix indicating the classification performance of the deep neural network.

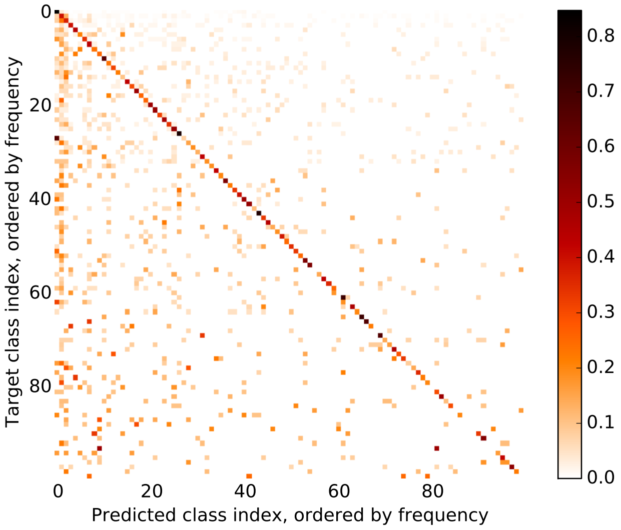

Figure 8: **Corpus VGT** confusion matrix with cross-domain learned features.

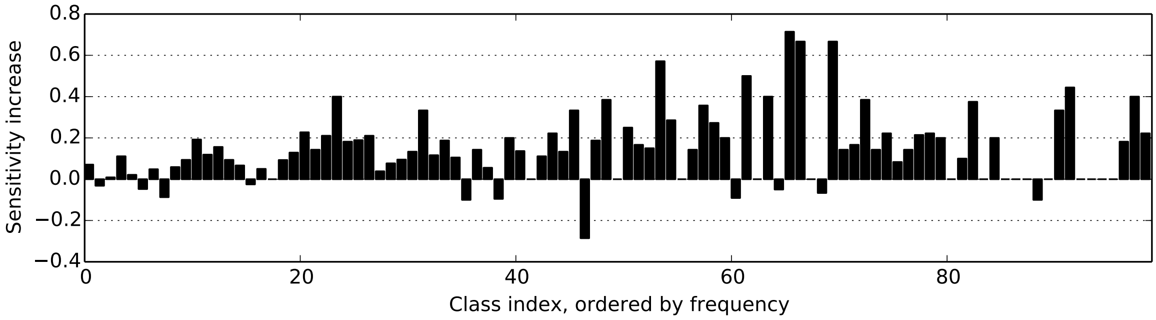

Figure 9: **Corpus VGT** sensitivity (true positive rate) increase compared to the model without cross-domain feature learning, depicted for each sign. Some signs are negatively affected by it. Further research will be required to determine the reason.

Our models achieve an accuracy of 39.3% with the Corpus VGT and 56.2% with the Corpus NGT for the 100 most common signs. We also show that the knowledge learned from the Corpus NGT can be passed on to boost the performance of the Corpus VGT.

Given the high dimensionality of video, the fact that these corpora are not tailored for machine learning and the fast and subtle movements of Deaf signers, deep neural networks show potential to build upon for SLR. The need for manual feature engineering, specialized hardware or other constraints decreases with more available corpora, advancements in unsupervised learning (learning from data without annotations) and language modeling.

5. Acknowledgments

We would like to thank NVIDIA Corporation for the donation of a GPU used for this research. The research leading to these results has received funding from the Agency for Innovation by Science and Technology in Flanders (IWT).

6. Bibliographical References

Chai, X., Li, G., Lin, Y., Xu, Z., Tang, Y., Chen, X., and Zhou, M. (2013). Sign language recognition and translation with kinect. In *IEEE Conf. on AFGR*.

Charles, J., Pfister, T., Everingham, M., and Zisserman, A. (2013). Automatic and efficient human pose estimation for sign language videos. *International Journal of Computer Vision*, pages 1–21.

Crasborn, O. A. and Zwitserlood, I. (2008). The corpus ngt: an online corpus for professionals and laymen. In *Construction and Exploitation of Sign Language Corpora. 3rd Workshop on the Representation and Processing of Sign Languages (LREC)*, pages 44–49. ELDA.

Crasborn, O., Zwitserlood, I., and Ros, J. (2008). The Corpus NGT. A digital open access corpus of movies and annotations of Sign Language of the Netherlands. *Centre for Language Studies, Radboud Universiteit Nijmegen. http://www.ru.nl/corpusngtukgp/.*

Escalera, S., Baró, X., Gonzàlez, J., Bautista, M. A., Madadi, M., Reyes, M., Ponce, V., Escalante, H. J., Shotton, J., and Guyon, I. (2014). Chalearn looking at people challenge 2014: Dataset and results. In *ECCV workshop*.

Graves, A., Mohamed, A.-r., and Hinton, G. (2013). Speech recognition with deep recurrent neural networks. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 6645–6649. IEEE.

Lecun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11).

Liu, J., Liu, B., Zhang, S., Yang, F., Yang, P., Metaxas, D. N., and Neidle, C. (2014). Non-manual grammatical marker recognition based on multi-scale, spatio-temporal analysis of head pose and facial expressions. *Image and Vision Computing*, 32(10):671–681.

Ong, E.-J. and Bowden, R. (2004). A boosted classifier tree for hand shape detection. In *Automatic Face and Gesture Recognition, 2004. Proceedings. Sixth IEEE International Conference on*, pages 889–894. IEEE.

Oz, C. and Leu, M. C. (2011). American sign language word recognition with a sensory glove using artificial neural networks. *Engineering Applications of Artificial Intelligence*, 24(7):1204–1213.

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

Pigou, L., Oord, A. v. d., Dieleman, S., Van Herreweghe, M., and Dambre, J. (2015). Beyond temporal pooling: Recurrence and temporal convolutions for gesture recognition in video. *arXiv preprint arXiv:1506.01911.*

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556.*

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2014). Going deeper with convolutions. *arXiv preprint arXiv:1409.4842.*

Van Herreweghe, M., Vermeerbergen, M., Demey, E., De Durpel, H., H., N., and Verstraete, S. (2015). Het Corpus VGT. Een digitaal open access corpus van videos and annotaties van Vlaamse Gebarentaal, ontwikkeld aan de Universiteit Gent ism KU Leuven. www.corpusvgt.be.

Wang, R. Y. and Popović, J. (2009). Real-time hand-tracking with a color glove. *ACM transactions on graphics (TOG)*, 28(3):63.