Deep Sign: Hybrid CNN-HMM for Continuous Sign Language Recognition

Oscar Koller[1]
koller@cs.rwth-aachen.de
Sepehr Zargaran[1]
sepehr.zargaran@rwth-aachen.de
Hermann Ney[1]
ney@cs.rwth-aachen.de
Richard Bowden[2]
r.bowden@surrey.ac.uk

[1] Human Language Technology and Pattern Recognition Group
RWTH Aachen University
Aachen, Germany

[2] Centre for Vision Speech and Signal Processing
University of Surrey
Guildford, UK

Abstract

This paper introduces the end-to-end embedding of a CNN into a HMM, while interpreting the outputs of the CNN in a Bayesian fashion. The hybrid CNN-HMM combines the strong discriminative abilities of CNNs with the sequence modelling capabilities of HMMs. Most current approaches in the field of gesture and sign language recognition disregard the necessity of dealing with sequence data both for training and evaluation. With our presented end-to-end embedding we are able to improve over the state-of-the-art on three challenging benchmark continuous sign language recognition tasks by between 15% and 38% relative and up to 13.3% absolute.

1 Introduction

Gesture is a key part in human to human communication. However, the role of visual cues in spoken language is not well defined. Sign language on the other hand provides a clear framework with a defined inventory and grammatical rules that govern joint expression by hand (movement, shape, orientation, place of articulation) and by face (eye gaze, eye brows, mouth, head orientation). This makes sign languages, the natural languages of the deaf, a perfect test bed for computer vision and human language modelling algorithms targeting human computer interaction and gesture recognition. Videos represent time series of changing images. The recognition of sign language therefore needs to be able to cope with variable input sequences and execution speed. Different schemes are followed to achieve this ranging from sliding window approaches [20] to temporal normalisations [18] or dynamic time warping [14]. While in the field of automatic speech recognition, HMM dominate the field, they remainx rather unpopular in computer vision related tasks. This may be related to the comparatively poor image modelling capabilities of Gaussian Mixture Models (GMMs), which are traditionally used to model the observation probabilities within such a framework. More recently, deep Convolutional Neural Networks (CNNs) have outperformed other approaches

Pages 136.1-136.12
DOI: https://dx.doi.org/10.5244/C.30.136

in all computer vision tasks. Which is why we focus on integrating CNNs in a Hidden-Markov-Model (HMM) framework, extending an interesting line of work [13, 16, 27], which we will discuss more closely in Section 2.

In the scope of this paper we make several contributions:

1. To the best of our knowledge, we are the first to embed a deep CNN in a HMM framework in the context of sign language and gesture recognition, while treating the outputs of the CNN as true Bayesian posteriors and training the system as a hybrid CNN-HMM in an end-to-end fashion.

2. We present a large relative improvement of over 15% compared to the state-of-the-art on three challenging standard benchmark continuous sign language recognition data sets.

3. We analyse the impact of the alignment quality on the hybrid performance and we experimentally compare the hybrid and tandem approach, which has not been done in the domain of gesture before.

The remainder of this paper is organised as follows: In Section 2 we discuss the related work. Section 3 introduces the proposed approach, while Section 4 shows the experimental validation on three publicly available benchmark continuous sign language recognition data sets. Finally, we summarise the paper in Section 5.

2 Related Work

Following the recent popularity of CNNs [17] in computer vision, several works have made use of it in gesture and sign language recognition [9, 12, 19]. However, in most previous CNN-based approaches the temporal domain is not elegantly taken into consideration. Most approaches use a sliding window or simply evaluate the output in terms of overlap with the ground truth [21]. Moreover, CNNs are usually trained on the frame-level. Therefore, problems also arise during training, unless the data set specifies frame-level labels (*e.g.* [3]). This is usually not the case, especially for sign language footage or other real-life data sets. Available annotation usually consists of sequences of signs without explicit frame-level information. As such, the focus of the field should move more towards approaches that deal with variable length inputs and outputs and do not require explicit frame labeling. Graphical models such as HMMs lend themselves well to this task and combine the best of different worlds when integrated with CNNs.

A few works have joined neural networks and HMMs before in the scope of gesture and sign language recognition. Wu *et al.* [27] use a 3D CNNs to model the observation probabilities in a HMM. However, they interpret the CNN outputs as likelihoods $p(x|k)$ for an image x and a given class k. This is surprising as already in the nineties Richard and Lippmann [22] showed that neural networks outputs are better interpreted as posteriors $p(k|x)$. This is best accounted for in a Bayesian framework. In the field of speech recognition Bayesian hybrid neural network HMMs were introduced in 1994 [1] and became the approach of choice. Le *et al.* [16] followed this line of thoughts, but only employed a shallow legacy neural network that is learnt to distinguish twelve artificial actions. Koller *et al.* [13] employed CNNs in a hybrid Bayesian fashion to perform weakly supervised training, learning a cross- data set hand shape classifier. Their work is closely related to ours, but there are clear differences: 1) we learn the CNN top down using hidden states of sign-words as underlying targets, whereas

they start bottom up with different hand shapes as building blocks for signs. 2) we learn the CNN-HMM in an end to end fashion from video input to gloss output, whereas they learn an intermediate subunit CNN which then serves as feature extractor for an additional GMM-HMM sign model. In the experimental evaluation of this manuscript we will show that our approach clearly outperforms theirs. [28] is also closely related to this work, but they do not interprete the CNN outputs in a Bayesian way, they use different inputs to the CNN (full body RGB and depth, as opposed to us only using a cropped hand patch) and different inputs to the HMM.

3 Theoretical Background

An overview of the proposed algorithm can be found in Figure 1. Given an input video as

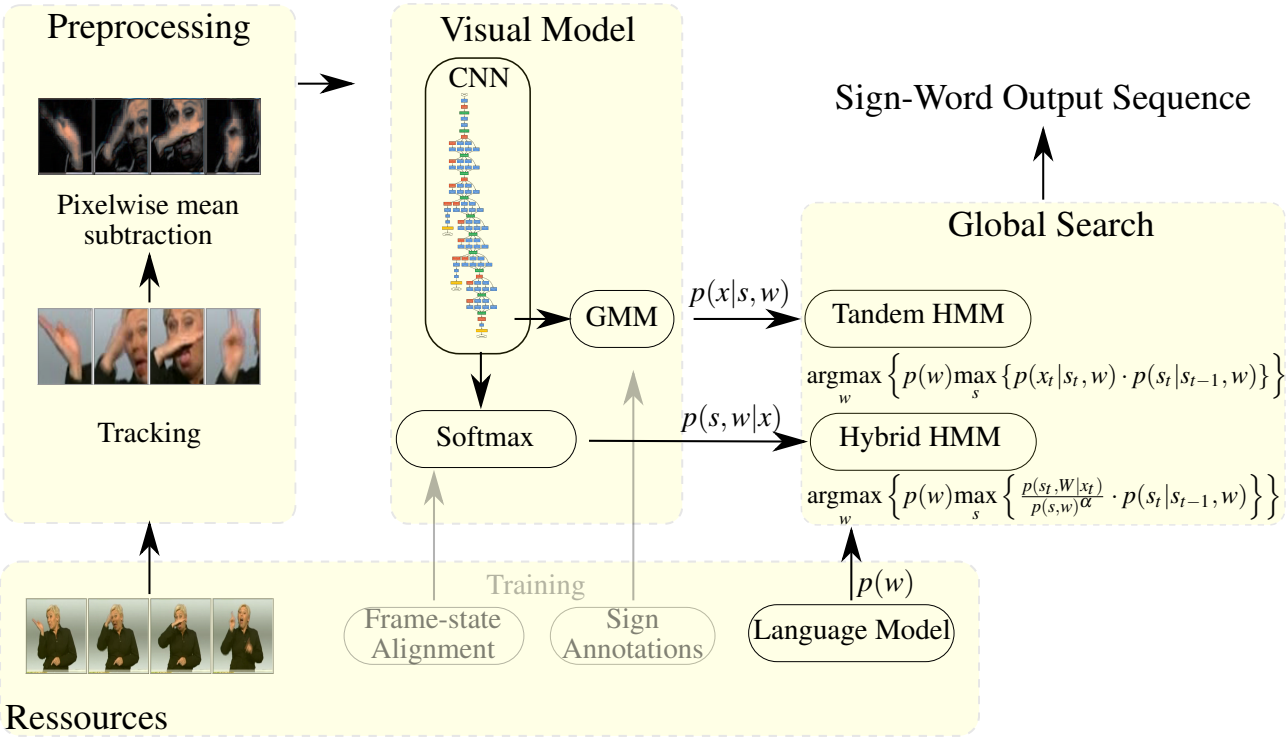

Figure 1: Overview of the proposed CNN-HMM hybrid approach. For clarification the tandem approach is also depicted.

a sequence of images $x_1^T = x_1,\ldots,x_T$ automatic continuous sign language recognition tries to find an unknown sequence of sign-words w_1^N for which x_1^T best fit the learned models. We assume that images and sign-words occur in a monotonous fashion without reordering. Therefore, the problem of sign language recognition constitutes a clearly different case than translating from sign language to spoken language, where re-orderings are necessary. To find the best fitting sequence, we follow a statistical paradigm using Bayes' decision rule and maximise the class posterior probability $Pr(w_1^N|x_1^N)$:

$$x_1^T \rightarrow [w_1^N]_{\text{opt}} = \underset{w_1^N}{\text{argmax}}\left\{Pr(w_1^N|x_1^T)\right\} \tag{1}$$

When modelling the true class posterior probability by a generative model (as usually done in Automatic Speech Recognition (ASR) in a GMM-HMM framework), we decompose

it into the product of two different knowledge sources being the language model $p(w_1^N)$ and the visual model $p(x_1^T|w_1^N)$. To account for the temporal variation of our input, we model the problem using a HMM, a stochastic finite state automaton. In a HMM, each sign-word is modelled by a predefined number of hidden states s. We make a first order Markov assumption and a Viterbi or maximum approximation and therefore maximise following equation:

$$[w_1^N]_{\text{opt}} = \underset{w_1^N}{\text{argmax}} \left\{ p(w_1^N) \max_{s_1^T} \left\{ \prod_t p(x_t|s_t, w_1^N) \cdot p(s_t|s_{t-1}, w_1^N) \right\} \right\}, \tag{2}$$

In the scope of this work we model the emission probability of the HMM $p(x_t|s_t, w_1^N)$ by an embedded CNN, which is known to possess much more powerful image modelling capabilities than generative models such as GMMs. However, as pointed out by [22], the CNN is a discriminative model that models the posterior probability. Inspired by the hybrid approach known from ASR [1], we use the CNN to model the posterior probability $p(s|x)$ for a hidden state s given the input x. We employ a given frame-state-alignment (produced by a baseline GMM-HMM system) as frame labelling. To employ the posteriors in Equation 2, we need to convert them to likelihoods following Bayes' rule as follows:

$$p(x_t|s_t, w_1^N) = p(x_t) \cdot \frac{p(s_t, w_1^N|x_t)}{p(s_t, w_1^N)} \tag{3}$$

where $p(s_t, w_1^N)$ denotes the state counts in our frame-state-alignment. We add the scaling factor α as a hyper-parameter allowing us to control the impact of the state prior. Neglecting the constant prior of the frames $p(x_t)$, Equation 2 then becomes:

$$[w_1^N]_{\text{opt}} = \underset{w_1^N}{\text{argmax}} \left\{ p(w_1^N) \max_{s_1^T} \left\{ \prod_t \frac{p(s_t, w_1^N|x_t)}{p(s_t, w_1^N)^\alpha} \cdot p(s_t|s_{t-1}, w_1^N) \right\} \right\}, \tag{4}$$

We employ a pooled *state transition* model across all sign-words $Pr(s_t|s_{t-1})$ that define the transitions in the HMM in bakis structure (left-to-right structure; forward, loops and skips across at most one state are allowed, where two subsequent states share the same class probabilities). The garbage class is modelled as an ergodic state with separate transition probabilities to add flexibility, such that it can always be inserted between sequences of sign-words.

As discussed in Section 2, most other works either use the CNN's output in a non Bayesian interpretation [28] or employ the CNN as feature extractor rather than as classifier. In the so-called tandem approach [8] the activations of a fully connected layer or the feature maps of a convolutional layer are dumped, post-processed [13] and then modelled in a GMM-HMM framework. In our experiments in the following section, we compare the hybrid to the tandem approach. Besides the significantly higher computational cost for extracting features and retraining a system, we agree with previous findings from speech and handwriting recognition [6] that the hybrid approach shows equal or superior performance compared to the tandem approach.

4 Experiments

In this section we first mention the implementation details in Subsection 4.1 required to reproduce our experiments. Then we conduct experiments in Subsection 4.2 on three publicly available standard benchmark continuous sign language data sets:

1. RWTH-PHOENIX-Weather 2012 [4]
2. RWTH-PHOENIX-Weather Multisigner 2014 [11]
3. SIGNUM single signer [26]

The corpora are described in detail in [11] and constitute challenging real-life and artificial sign language data sets featuring up to nine different signers recorded from broadcast news.

4.1 Implementation Details

Image preprocessing. We use a dynamic programming based tracking approach similar to [2] to track the right hand across a sequence of images. In all employed data sets the right hand corresponds to the signer's dominant hand, which is the hand that plays the principle role in signing. We crop a rectangle of 92x132 pixel around the centre of the hand. However, the original images suffer a constant distortion due to the broadcast nature of the videos, which corresponds to a contraction of the image width by factor 0.7. To compensate for this distortion we enlarge the crops to the quadratic size of 256x256. Thereafter the pixel-wise mean of all images in the train set is subtracted from each image. Finally, for data augmentation we follow an online cropping scheme, which randomly crops out a 224x224 pixel rectangle to match the size of images in our model which was pre-trained on ImageNet.

Convolutional Neural Network Training. We base our CNN implementation on [10], which uses the NVIDIA CUDA Deep Neural Network GPU-accelerated library. We empirically compared recent CNN architectures [15, 24, 25] and opted for the GoogLeNet [25] 22 layers deep CNN architecture with around 5 million parameters. GoogLeNet has shown many times in the past, most notably in the ImageNet2014 Challenge, that it can be quite effective in combining impressive performance with minimal computational resources. Much of the improvements in this architecture compared to others' stems from the inception module which in short combines filters of different sizes after applying dimensionality reduction through a 1x1 Convolutional layer. The employed CNN architecture includes 3 classifying layers, meaning that besides the final classifier the network also includes two intermediary auxiliary classifiers. Those encourage discrimination in lower stages of the network. The loss of these auxiliary classifiers is added to the total loss with a weight of 0.3. All non-linearities are rectified linear units and each classifier layer is preceded by a dropout layer. We use a dropout rate of 0.7 for the auxiliary layers and 0.4 before the final classifier. As mentioned in Section 3, the CNN training scheme requires a given frame-state-alignment. We test different alignments as discussed in Subsection 4.2. For SIGNUM and RWTH-PHOENIX-Weather 2014 we use the best results published in [13] as alignment, whereas for RWTH-PHOENIX-Weather 2012 we use [11]. The alignment is used to generate a training and validation set to evaluate the per-frame accuracy and stop the training at a good point (usually among the last iterations). The training and validation set have a ratio of 10 to 1. Furthermore, since SIGNUM contains a lot of frames without any sign-words (background), we resample the background class to match the second most frequent class. We pretrain the

network on the ImageNet data set. Then we finetune the network for 80,000 iterations with a batch size of 32 images. Training uses stochastic gradient descent with 0.9 momentum and a starting learning rate of 0.01 which decreases using a polynomial scheme down to 0.0005 in the last iterations. After every 1000 iterations we calculate accuracy on the validation set and save the model.

CNN inference. Once the CNN training is finished, we choose the model that yielded the highest accuracy on the automatic validation set. We consider all three classifiers (the main one and the two auxiliary ones) for estimating the best performing iteration. For the proposed hybrid CNN-HMM approach we add a softmax and use the resulting posteriors in our HMM as observation probabilities. In the tandem CNN-HMM approach we employ the activations from the last layer before the softmax that yields the highest accuracy on the validation data. With RWTH-PHOENIX-Weather 2012, this is a fully connected layer of the first auxilary classifier, possibly because the data set does not provide enough data for training an earlier softmax is favourable. For RWTH-PHOENIX-Weather 2014 and SIGNUM it is a pooling layer before the main classifier. for the first before the chosen classifier which yields 1024 values. The tandem system requires feature extraction for both training and test sets, since a GMM-HMM system will be retrained with them. After a global variance normalisation, we apply PCA to reduce the feature dimension to 200.

Continuous Sign Language Recognition. We base the HMM part of this work on the freely available state-of-the-art open source speech recognition system RASR [23]. Following the hybrid approach we use these posterior probabilities, as well as the corresponding class priors (with a scaling factor of 0.3) to approximate likelihoods. As for RWTH-PHOENIX-Weather 2014 and SIGNUM, we model each sign-word with three hidden states. However, inspired by [11], in RWTH-PHOENIX-Weather 2012 we follow a length modelling scheme where sign-words are represented by more or fewer states depending on their average alignment length. In agreement to most sign language recognition literature, we measure the system performance in Word Error Rate (WER). word error rate (WER) is based on the levenshtein alignment between reference and hypothesis sentence and it measures the required numbers of deletion, insertion and substitution operations to transform the recognised hypothesis into the reference sequence.

$$\text{WER} = \frac{\#\text{deletions} + \#\text{insertions} + \#\text{substitutions}}{\#\text{reference observations}} \tag{5}$$

The loop, skip and exit transition penalties are optimised on the dev set in order to minimise the WER. Furthermore, the search space is pruned in order to boost performance and reduce memory consumption. We perform histogram and threshold pruning.

4.2 Results

We are aware of the fact that the hybrid approach requires a good frame-state alignment in order to guarantee a stable estimation of the CNN parameters. In the first experiment we analyse the impact of this training alignment. Results can be seen in Figure 2 where we display the relation between the alignment quality and the WER achieved by the hybrid CNN-HMM system. The alignment quality is also measured in WER and corresponds to the recognition performance of a standard GMM-HMM trained using this alignment. Results for both RWTH-PHOENIX-Weather 2014 development and test set are given. An alignment reaching 60.9% WER is improved by over 30% relative to 41.9% using the proposed hybrid approach. A baseline system with 51.6% experiences over 20% improvement and a baseline

alignment of 47.1% is still improved by 18% relative. We understand that the relative impact of the hybrid modelling decreases with better starting alignments, but we also note that the WER change seems to accelerate with improved alignments.

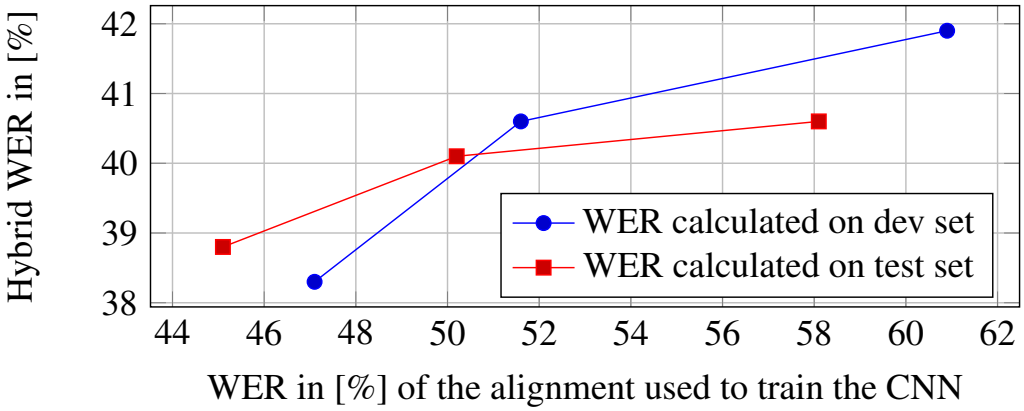

Figure 2: Impact of alignment: continuous sign language recognition results on RWTH-PHOENIX-Weather 2014 Multisigner showing how the initial alignment impacts the CNN learning and subsequent hybrid modelling.

Table 1 and 2 show a detailed comparison to the state-of-the art on the three employed benchmark corpora. Note that the proposed approach currently exploits only a single cropped hand of the signer. Sign language is highly multimodal and makes heavy use of manual components (hand shape, orientation, place of articulation, movement) and also non-manual components (facial expression, eyebrow height, mouth, head orientation, upper body orientation). For a fair comparison in Table 1, we only listed competitors that also just focus on the single hand. The previously best hand only result also relied on CNN models, but did not employ the hybrid approach end-to-end in recognition. It set the benchmark on PHOENIX 2014 Multisigner to 51.6% WER. However, our proposed CNN-HMM achieves a strong result of 38.3% and 38.8% on dev and test respectively. This corresponds to 11.9% WER absolute or over 20% relative improvement. On the single signer corpus RWTH-PHOENIX-Weather 2012 the proposed approach improved the best baseline from 35.5% to 30.0%, still being a relative improvement of over 15%. On SIGNUM we improve the best known word error rates from 35.5% to 30.0% and from 12.0% to 7.4%.

As can be seen in Table 2, the unimodal hybrid CNN-HMM even outperforms the best known multimodal systems by over 18%, which employ very sophisticated features comprising both manual and non-manual components, despite the fact that it uses hand appearance only.

	PHOENIX 2014				PHOENIX 2012		SIGNUM	
	Dev		Test		Test		Test	
	del/ins	WER	del/ins	WER	del/ins	WER	del/ins	WER
[26]	–	–	–	–	–	–	–	19.2
HoG-3D	25.8/4.2	60.9	23.2/4.1	58.1	19.5/4.9	43.5	2.8/2.4	12.5
1-Mio-Hands [13]	19.1/4.1	51.6	17.5/4.5	50.2	15.8/3.3	35.5	1.5/2.5	12.0
CNN-Hybrid	12.6/5.1	**38.3**	11.1/5.7	**38.8**	16.8/2.1	**30.0**	1.4/1.4	**7.4**

Table 1: Continuous sign language recognition results on RWTH-PHOENIX-Weather 2012 and RWTH-PHOENIX-Weather 2014 Multisigner using just dominant hand information.

	Modality			PHOENIX 2014				PHOENIX 2012		SIGNUM	
				Dev		Test		Test		Test	
	r-hand	l-hand	face	del/ins	WER	del/ins	WER	del/ins	WER	del/ins	WER
[26]	X	X	X	–	–	–	–	–	–	–	12.7
[7]	X	X	X	–	–	–	–	–	–	–	11.9
[5]	X	X	X	–	–	–	–	–	41.9	–	10.7
[11]	X	X	X	23.6/4.0	57.3	23.1/4.4	55.6	–	34.3	1.7/1.7	10.0
[11] CMLLR	X	X	X	21.8/3.9	55.0	20.3/4.5	53.0	–	–	–	–
[13]+[11]	X	X	X	16.3/4.6	47.1	15.2/4.6	45.1	–	–	0.9/1.6	7.6
CNN-Hybrid	X			12.6/5.1	**38.3**	11.1/5.7	**38.8**	16.8/2.1	**30.0**	1.4/1.4	**7.4**

Table 2: Multi-modal continuous sign language recognition results on RWTH-PHOENIX-Weather 2014 Multisigner and SIGNUM.

Figure 3 compares the hybrid CNN-HMM modelling against the tandem modelling as presented in Section 3. We can see that the hybrid approach outperforms the tandem approach on both data sets RWTH-PHOENIX-Weather 2012 and RWTH-PHOENIX-Weather 2014. This is consistent with the literature [6].

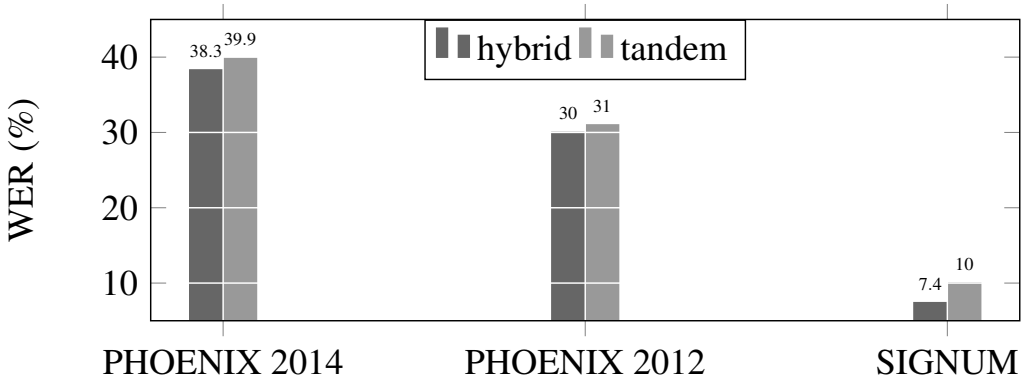

Figure 3: WER of the hybrid and the tandem approach side-by-side.

Qualitive examination of the top confusions made by the approach highlight confused pairs such as “SNOW” with “RAIN” or “SHOWER” with “RAIN”. However, these signs share the same hand configurations, whereas only the mouth shape changes. Given the classification relies purely on the right hand, it is understandable why it cannot distinguish between these signs. The top 30 confusions all relate to this type of error.

Computational requirements. Using a GeForce GTX 980 gpu with 4GB memory, training on the Phoenix-2012 data set is done at the speed of $\sim$ 150 frames per second (fps) and inference at a rate of $\sim$ 450 fps. Using the same hardware on Phoenix-2014 data set yields $\sim$ 35 fps for training and $\sim$ 350 fps for inference. HMM recognition is done at $\sim$ 2 fps for Phoenix-2012 and due to the tighter pruning $\sim$ 25 fps for Phoenix-2014. The HMM parameter optimisation took a total of $\sim$ 38 hours for Phoenix-2012 and $\sim$ 130 hours for Phoenix-2014 using a single core machine with 2GB RAM.

5 Conclusion

In this work we introduced an end-to-end embedding of a CNN into a HMM, while interpreting the outputs of the CNN in a truly Bayesian fashion. Most state-of-the-art approaches in gesture and sign language modelling use a sliding window approach or simply evaluate the output in terms of overlap with the ground truth. While this is sufficient for data sets that provide such training and evaluation characteristics, it is unsuitable for real world use. For the field to move forward more realistic scenarios, such as those imposed by challenging real-life sign language corpora, are required. We presented a hybrid CNN-HMM framework that combines the strong discriminative abilities of CNNs with the sequence modelling capabilities of HMMs, while abiding to Bayesian principles.

To the best of our knowledge, we believe to be the first work to embed a deep CNN in a HMM framework in the context of sign language and gesture recognition, while treating the outputs of the CNN as true Bayesian posteriors and training the system end-to-end as a hybrid CNN-HMM.

We present a large relative improvement compared to the current state-of-the-art on three challenging benchmark continuous sign language recognition data sets. On the two single signer data sets RWTH-PHOENIX-Weather 2012 and SIGNUM we improve the best known word error rates from 35.5% to 30.0% and from 12.0% to 7.4% respectively, while only employing basic hand-patches as input. On the difficult 9 signer $>$ 1000 vocab RWTH-PHOENIX-Weather 2014 Multisigner, we lower the error rates from 51.6% / 50.2% to 38.3% / 38.8% on dev / test.

In terms of future work, we would like to extend our approach to cover all relevant modalities. Moreover, techniques to overcome the necessary initial alignment will also be investigated.

6 Acknowledgments

This work was partially funded by the SNSF Sinergia project "Scalable Multimodal Sign Language Technology for Sign Language Learning and Assessment (SMILE)" grant agreement number CRSII2_160811

References

[1] Herve A. Bourlard and Nelson Morgan. *Connectionist speech recognition: a hybrid approach*, volume 247. Springer Science & Business Media, 1994.

[2] Philippe Dreuw, Thomas Deselaers, David Rybach, Daniel Keysers, and Hermann Ney. Tracking Using Dynamic Programming for Appearance-Based Sign Language Recognition. In *IEEE International Conference Automatic Face and Gesture Recognition*, pages 293–298, Southampton, UK, April 2006. IEEE.

[3] Sergio Escalera, Xavier Baró, Jordi Gonzalez, Miguel A. Bautista, Meysam Madadi, Miguel Reyes, Víctor Ponce-López, Hugo J. Escalante, Jamie Shotton, and Isabelle Guyon. Chalearn looking at people challenge 2014: Dataset and results. In *Computer Vision-ECCV 2014 Workshops*, pages 459–473. Springer, 2014.

[4] Jens Forster, Christoph Schmidt, Thomas Hoyoux, Oscar Koller, Uwe Zelle, Justus Piater, and Hermann Ney. RWTH-PHOENIX-Weather: A Large Vocabulary Sign Language Recognition and Translation Corpus. In *International Conference on Language Resources and Evaluation*, pages 3785–3789, Istanbul, Turkey, May 2012.

[5] Jens Forster, Christian Oberdörfer, Oscar Koller, and Hermann Ney. Modality Combination Techniques for Continuous Sign Language Recognition. In *Iberian Conference on Pattern Recognition and Image Analysis*, Lecture Notes in Computer Science 7887, pages 89–99, Madeira, Portugal, June 2013. Springer.

[6] Pavel Golik, Patrick Doetsch, and Hermann Ney. Cross-entropy vs. squared error training: a theoretical and experimental comparison. In *INTERSPEECH*, pages 1756–1760, 2013.

[7] Yannick Gweth, Christian Plahl, and Hermann Ney. Enhanced Continuous Sign Language Recognition using PCA and Neural Network Features. In *CVPR 2012 Workshop on Gesture Recognition*, pages 55–60, Providence, Rhode Island, USA, June 2012.

[8] Hynek Hermansky, Daniel W. Ellis, and Shantanu Sharma. Tandem connectionist feature extraction for conventional HMM systems. In *Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings. 2000 IEEE International Conference on*, volume 3, pages 1635–1638. IEEE, 2000.

[9] Jie Huang, Wengang Zhou, Houqiang Li, and Weiping Li. Sign Language Recognition using 3D convolutional neural networks. In *2015 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6, June 2015. doi: 10.1109/ICME.2015.7177428.

[10] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional Architecture for Fast Feature Embedding. *arXiv preprint arXiv:1408.5093*, 2014.

[11] Oscar Koller, Jens Forster, and Hermann Ney. Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers. *Computer Vision and Image Understanding*, 141:108–125, December 2015. ISSN 1077-3142. doi: 10.1016/j.cviu.2015.09.013.

[12] Oscar Koller, Hermann Ney, and Richard Bowden. Deep Learning of Mouth Shapes for Sign Language. In *Third Workshop on Assistive Computer Vision and Robotics, ICCV*, Santiago, Chile, December 2015.

[13] Oscar Koller, Hermann Ney, and Richard Bowden. Deep Hand: How to Train a CNN on 1 Million Hand Images When Your Data Is Continuous and Weakly Labelled. In *IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, June 2016.

[14] Ravikiran Krishnan and Sudeep Sarkar. Conditional distance based matching for one-shot gesture recognition. *Pattern Recognition*, 48(4):1298–1310, 2015.

[15] Alex Krizhevsky, Ilya Sutskever, and Geoff Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25*, pages 1106–1114, 2012.

[16] Hai-Son Le, Ngoc-Quan Pham, and Duc-Dung Nguyen. Neural Networks with Hidden Markov Models in Skeleton-Based Gesture Recognition. In Viet-Ha Nguyen, Anh-Cuong Le, and Van-Nam Huynh, editors, *Knowledge and Systems Engineering*, number 326 in Advances in Intelligent Systems and Computing, pages 299–311. Springer International Publishing, 2015. ISBN 978-3-319-11679-2 978-3-319-11680-8. doi: 10.1007/978-3-319-11680-8_24.

[17] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[18] Pavlo Molchanov, Shalini Gupta, Kihwan Kim, and Jan Kautz. Hand Gesture Recognition with 3D Convolutional Neural Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–7, 2015.

[19] Natalia Neverova, Christian Wolf, Graham W. Taylor, and Florian Nebout. Multi-scale Deep Learning for Gesture Detection and Localization. In Lourdes Agapito, Michael M. Bronstein, and Carsten Rother, editors, *Computer Vision - ECCV 2014 Workshops*, Lecture Notes in Computer Science, pages 474–490. Springer International Publishing, September 2014. ISBN 978-3-319-16177-8 978-3-319-16178-5. doi: 10.1007/978-3-319-16178-5_33.

[20] Eng-Jon Ong, Oscar Koller, Nicolas Pugeault, and Richard Bowden. Sign Spotting using Hierarchical Sequential Patterns with Temporal Intervals. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1931–1938, Columbus, OH, USA, June 2014.

[21] Lionel Pigou, Aäron van den Oord, Sander Dieleman, Mieke Van Herreweghe, and Joni Dambre. Beyond Temporal Pooling: Recurrence and Temporal Convolutions for Gesture Recognition in Video. *arXiv:1506.01911 [cs, stat]*, June 2015.

[22] Michael D. Richard and Richard P. Lippmann. Neural Network Classifiers Estimate Bayesian a posteriori Probabilities. *Neural Computation*, 3(4):461–483, 1991.

[23] David Rybach, Stefan Hahn, Patrick Lehnen, David Nolden, Martin Sundermeyer, Zoltán Tüske, Simon Wiesler, Ralf Schlüter, and Hermann Ney. RASR - The RWTH Aachen University Open Source Speech Recognition Toolkit. In *IEEE Automatic Speech Recognition and Understanding Workshop*, Waikoloa, HI, USA, December 2011.

[24] Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv:1409.1556 [cs]*, September 2014.

[25] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going Deeper With Convolutions. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9, Boston, Ma, USA, June 2015.

[26] Ulrich von Agris, Moritz Knorr, and K.-F. Kraiss. The significance of facial features for automatic sign language recognition. In *Automatic Face & Gesture Recognition, 2008. FG'08. 8th IEEE International Conference on*, pages 1–6. IEEE, 2008.

[27] Di Wu and Ling Shao. Deep dynamic neural networks for gesture segmentation and recognition. In *Computer Vision-ECCV 2014 Workshops*, pages 552–571. Springer, 2014.

[28] Di Wu, Lionel Pigou, Pieter-Jan Kindermans, Nam LE, Ling Shao, Joni Dambre, and Jean-Marc Odobez. Deep Dynamic Neural Networks for Multimodal Gesture Segmentation and Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP(99):15, 2016. ISSN 0162-8828.