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

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 strong discriminative abilities of CNNs with sequence modeling 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% & 38% relative & up to 13.3% absolute.

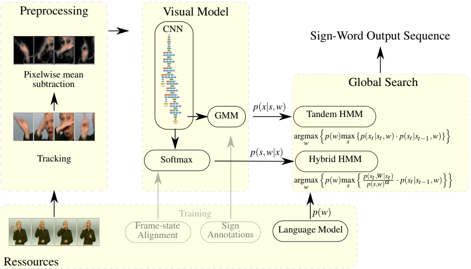
Preprocessing
Pixelwise mean subtraction
Tracking
Visual Model
CNN
GMM
Softmax
$p(x|s,w)$
$p(s,w|x)$
Sign-Word Output Sequence
Global Search
Tandem HMM
$\operatorname{argmax}_w \left\{ p(w) \max_s \{ p(x_t|s_t,w) \cdot p(s_t|s_{t-1},w) \} \right\}$
Hybrid HMM
$\operatorname{argmax}_w \left\{ p(w) \max_s \left\{ \frac{p(s_t,W|x_t)}{p(s,w)^\alpha} \cdot p(s_t|s_{t-1},w) \right\} \right\}$
$p(w)$
Training
Frame-state Alignment
Sign Annotations
Language Model
Ressources

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

Gesture is a key part in human communication. However, it does not have a well defined structure. 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 a perfect test bed for computer vision and human language modeling algorithms targeting human computer interaction and gesture recognition.

Following the recent popularity of CNNs in computer vision, several works have made use of it in gesture and sign language recognition. 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 on the frame level. We present a hybrid modeling scheme that incorporates a CNN into a HMM. Inspired by the hybrid approach known from speech recognition [1], we use the CNN to model the posterior probability $p(s|x)$ for a hidden state s given the input image x. In this way only the CNN needs to be trained. Opposed to previous works combining CNNs with HMMs, we convert the posteriors into scaled likelihoods using Bayes' rule such that they neatly integrate into the HMM-framework.

We make several contributions:

1. 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 data sets.

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

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