Weakly Supervised Learning with Multi-Stream CNN-LSTM-HMMs to Discover Sequential Parallelism in Sign Language Videos

Oscar Koller, Necati Cihan Camgoz, Hermann Ney, *Fellow, IEEE*, and Richard Bowden, *Senior Member, IEEE*

Abstract—In this work we present a new approach to the field of weakly supervised learning in the video domain. Our method is relevant to sequence learning problems which can be split up into sub-problems that occur in parallel. Here, we experiment with sign language data. The approach exploits sequence constraints within each independent stream and combines them by explicitly imposing synchronisation points to make use of parallelism that all sub-problems share. We do this with multi-stream HMMs while adding intermediate synchronisation constraints among the streams. We embed powerful CNN-LSTM models in each HMM stream following the hybrid approach. This allows the discovery of attributes which on their own lack sufficient discriminative power to be identified. We apply the approach to the domain of sign language recognition exploiting the sequential parallelism to learn sign language, mouth shape and hand shape classifiers. We evaluate the classifiers on three publicly available benchmark data sets featuring challenging real-life sign language with over 1000 classes, full sentence based lip-reading and articulated hand shape recognition on a fine-grained hand shape taxonomy featuring over 60 different hand shapes. We clearly outperform the state-of-the-art on all data sets and observe significantly faster convergence using the parallel alignment approach.

Index Terms—Weakly supervised learning, hybrid CNN-LSTM-HMMs, continuous sign language recognition, lip reading, hand shape recognition

1 Introduction

In this manuscript we propose a new solution to the problem of weakly supervised learning in the field of videos. Our approach exploits parallelism in the visual domain and trains multiple strong classifiers based on weak labels that occur in the image sequences. The approach is relevant to problems which can be split into sub-tasks that occur in parallel. In this work, we only consider the use case of sign language recognition. We model continuous sign language as a sequence of signs that are represented as co-occurring mouth and hand shape patterns. Our algorithm learns strong CNN-LSTM classifiers end-to-end based on weak and noisy labels and embeds them into a multi-stream Hidden-Markov-Model (HMM). We jointly align these modalities by introducing intermediate synchronisation constraints in the HMM that represent the parallel nature of the streams. As such, we boost the learning of mouth shapes and hand shapes from weak labels.

This paper is a novel contribution that can be best understood in the context of our previous work in the field of weakly supervised learning for the labelling of sequence data. In our early work on lip-reading [1], [2], we have shown that HMMs can be used to discover high-quality labels describing mouth shapes present in natural continuous sign language sequences. In [3] we proposed an end-to-end embedding of Convolutional Neural Networks (CNNs) into HMMs and showed its superior performance by exploiting over 1 million weakly labelled articulated hand images. Both applications solved each weak learning problem separately from the other. However, hand motion, shape and mouth gestures are sub-problems of the more complex sign language recognition task and actually occur simultaneously. This can be exploited by jointly defining the weakly supervised learning solution. Actually, many sequence learning problems in the visual domain can be split up into a set of parallel sub-problems. To address this, in this work we make the following contributions:

- To the best of our knowledge, we are the first to tackle weakly supervised learning with intermediate synchronisation constraints among multiple streams. This allows the discovery of groups of attributes which on their own lack sufficient discriminative power to be identified.
- Rather than constraining the input of expert networks by error prone preprocessing (e.g. tracking and cropping the mouth for lip reading), we propose to add multiple loss functions with weakly learnt labels. As such, we dispense with preprocessing and learn powerful mouth and hand shape classifier directly from full images. We compare this to training a system on cropped hands and faces.
- We evaluate our approach on three challenging publicly available data sets for continuous sign language recognition, mouth shape detection and hand shape classification, where we clearly outperform the state-

- *O. Koller, currently with Microsoft, was and H. Ney is with the Human Language Technology and Pattern Recognition Group, RWTH Aachen University, Germany.*
  E-mail: {surname}@cs.rwth-aachen.de
- *C. Camgoz and R. Bowden are with CVSSP, University of Surrey, United Kingdom.*

Manuscript received January 15, 2018.

of-the-art. On PHOENIX 2014, we reduce the word error rate (WER) to 26.0% on both the dev and test set.
- Our proposed hybrid multi-stream CNN-LSTM-HMM achieves significantly faster convergence as opposed to standard single stream methods.

2 Related Work

This work deals with the problem of weakly supervised learning from sequence labels. To tackle the problem, we exploit weak labels covering three modalities, namely gesture, mouth shape and hand shape and exploit the fact that all three contain sequential information with loose time synchronisation with respect to each other. We extend our previous work on hybrid HMM modelling for sign language recognition [3] [4] [5] by adding multi-stream HMMs with synchronisation constraints. The hybrid HMM modelling has shown to outperform other sequence learning approaches on sign language recognition data sets while requiring less memory and allowing for deeper architectures [4]. In this section, we therefore look at related work in the domain of multi-stream modelling, but we also consider literature tackling weakly supervised learning in general.

In 1966, HMMs were first introduced by Baum et al. calling them "Probabilistic Functions of Markov Chains" [6]. Speech recognition soon became an important application for the HMM and the three main problems (probability calculation, state sequence estimation and parameter estimation) had been solved by the time Lawrence Rabiner published his well-known tutorial paper [7] in 1989. One drawback for applications in the field of computer vision, was the fact that standard HMMs are only able to model a single sequential process. However, in the late 90s and early 2000s simultaneously in several fields, the focus shifted towards HMMs that are able to model observations originating from multiple processes: This happened in speech recognition [8] [9], audio-visual speech recognition [10] [11], action recognition [12] and sign language recognition [13] [14] [15] [16].

The different approaches of modelling multiple processes can be divided into the broad categories of feature fusion and decision fusion methods [17]. Where the former stack the features of multiple processes together and model them as a single vector with a multivariate function. However, processes that do not evolve synchronously cannot be appropriately modelled like this: a large disadvantage for the application to the visual domain. The latter group of decision fusion builds the classifier into the fusion process, where approaches may be based on early, intermediate or late fusion.

Early fusion behaves similar to the feature fusion. Multiple HMM streams that have to be characterised by the same number of states evolve in lock-step. All streams always reside in the same pair of locked states. This approach is therefore also known as a state synchronous multi-stream HMM and is more flexible than the feature fusion as the streams can be differently weighted. It assumes, however, that the different processes have absolute synchronicity. An assumption which is often false.

Late fusion trains independent models and combines their final model outputs after processing the inputs separately. While this allows for complete independence and asynchronism of the processes to be modelled, it does not exploit interaction and partial synchronisation, which is shown by our work to be essential to boosting weakly supervised learning performance.

Intermediate fusion combines model outputs while processing the streams. We can either think of it as modelling conditional probabilities where a state in one stream depends on states in the other streams or it may involve completely different independent streams that have specific mechanisms of synchronisation. This is the category our work belongs to and it is particularly well suited for weakly supervised learning. A variety of different schemes in this group have been developed. A good overview can be found in [17]. The most general way of combining several streams that originate from multiple processes are product HMMs [18] [10], [19]. These allow full temporal asynchronicity between the streams and also different topologies (e.g. number of states) per stream. Product HMMs express each possible combination of states as a new composite state, which are then typically modelled with an independent Gaussian Mixture Model (GMM). If in addition to the combination of independent asynchronous streams, we expect them to have specific temporal points of intermediate synchronisation, then the concept of multi-stream HMMs with synchronisation points introduced by Bourlard et al. [8] constitutes a powerful option. Bourlard et al. applied the idea to noise robust sub-band speech recognition, where each stream is trained separately on a different frequency band and temporal synchronisation among the multiple streams is imposed at phoneme, syllable or word boundaries. In his work, the asynchronous parts of the multi-stream HMMs (e.g. inside a syllable) are implicitly represented as a product HMM.

There are several other multi-stream HMM models, such as the factorial HMM [20], the coupled HMM [21] and the asynchronous HMM [11] [22]. Factorial HMMs constitute completely asynchronous streams without any synchronisation until the end of the sequence. They were developed to allow a distributed and therefore more efficient state representation (when multiple sources are involved). Ghahramani presents an algorithm based on mean field approximation that allows for $O(TN^2M)$ complexity, where T is the sequence length, N is the maximum number of states and M the number of streams [20]. In coupled HMMs the states and particularly their emission probabilities depend on all streams. Brand proposes an algorithm ("N-head dynamic programming") that reaches $O(T(NM)^2)$ complexity [21]. The asynchronous HMM requires both streams to share the same topology [23]. This is imposed due to the assumption that there is a single underlying hidden process that has multiple distinct probabilities to emit an observation on all or just on a single stream. Bengio presents an algorithm with $O(T^MN^2)$ complexity if all streams have the same observation length T [22].

Grouping approaches by application domain, in terms of **sign language recognition** there have been two published works that included multi-stream HMMs with synchronisation constraints in a meaningful way [13] [24]. However, both works deal with a limited vocabulary size and

with weak GMM-HMM models, instead of strong state-of-the-art hybrid neural network-based or even CNN-LSTM-HMM models. Moreover, these approaches do not train their models with synchronisation constraints (the models are trained independently and the temporal constraints are only applied during testing). Specifically, in 1999, Vogler and Metaxas [13] used multi-stream GMM-HMMs with synchronisation constraints on sign ends (which they call parallel HMMs) and apply them to continuous American Sign Language recognition using cyber-gloves for feature extraction of the right and the left hand. They report an improvement from 6.7% to 5.8% WER for their 22 sign vocabulary task using 400 training and 99 test sentences. Moreover, their single-stream HMM is trained on right hand input only, while the multi-stream HMM has information from both hands. How much of the improvement is actually due to the multi-stream HMM is left unclear. While they perform recognition using the multi-stream scheme, they train each stream independently. In 2008, von Agris et al. [24], report continuous sign language recognition results of German sign language (DGS) distinguishing a vocabulary of 100 signs, while looking at the hands only. They report an accuracy of 87.8% but do not provide a comparison to using less streams. Moreover, they combine the streams in an unweighted fashion with no normalisation. Several other works which claim to use parallel HMMs [14] [25] [26] for sign language exist. However, they all deal with signs in isolation (not in a continuous sentence sequence), which essentially turns the multi-stream HMM with synchronisation at the end of signs into a standard late fusion approach and is therefore not comparable to the approach analysed in this paper. In 2013, Forster et al. [23] compared different modality combination techniques to recognise continuous DGS with a vocabulary of up to 455 signs. They use multi-stream HMMs with synchronisation constraints at the word-level. However, they only employ the combination during recognition, training each stream separately. As such, they report the performance between a single stream HMM and multi-stream HMM to be 45.6% and 41.9% WER. There exist more recent approaches relying on recurrent neural networks (RNNs) (without HMMs) for sequence modelling that perform intermediate fusion of multiple channels [27]. However, besides using a small inventory of 10 gestures, the authors exclude the problem of temporal segmentation (weak learning) and rely on frame labels instead, which precludes application to more realistic scenarios.

Audio-visual speech recognition constitutes a perfect application to analyse the modelling of sequential parallelism. Both streams (audio and visual stream) are not perfectly synchronous while maintaining synchronisation at least on the word or even at the phone level. In 1996, Tomlinson et al. [19] compared multi-stream HMMs to standard single-stream HMMs on an audio-visual speech recognition task. They model tri-phone sub-word units both for auditory and visual features and apply synchronisation constraints at the phone level. The multi-stream HMM clearly outperforms a single-stream HMM when the audio data is noisy with WERs of 20.3% against 25.7%. Similarly, Neti et al. [28] compare both variants and report 35.2% WER against 37.0% with a multi-stream HMM architecture (synchronised at the phone level) and a single-stream HMM on noisy data. Due to the advances in deep learning, more recent publications often tackle the audio-visual speech recognition (or lip reading only) task with RNN-based encoder-decoder [29] networks. A few of these works also consider learning from multiple streams. Petridis et al. [30] take a variety of profile view angles as different streams. Their pretraining of the single streams prior to training the multi-stream architecture helps in the absence of specific synchronisation constraints. Chung et al. [31] use an encoder-decoder scheme with attention [32]. More exactly, they have a separate attention mechanism for the audio and the video streams. Their approach to adding implicit (word-level) synchronisation constraints starts by training on single word segments first and later evolving to truncated and full sentences. Even though they do not train the underlying CNN feature extractor, the authors run into memory problems, which is often a problem for the encoder-decoder architectures, where full sequences do not fit in the GPU memory. Hybrid CNN-LSTM-HMM models learnt with an Expectation Maximization (EM) algorithm alleviate such problems [4].

Speech recognition was the domain where the multi-stream approach with synchronisation was first applied [33] [8]. Variations with slightly different synchronisation constraints and databases appeared in the following years [34] [35].

There are many approaches to **learning from ambiguous labels** or **weakly supervised learning** (see [36] for an overview). A common approach is to employ multiple instance learning (MIL), treating a video sequence as a bag which is only labelled positive if it contains at least one true positive instance. MIL iteratively estimates the instance labels measuring a predefined loss. Buehler et al. [37] and similarly Kelly et al. [38] apply MIL to learning sign categories from TV subtitles, circumventing the translation problem by performing sign spotting. However, Farhadi and Forsyth [39] were the first to approach the subtitle-sign-alignment problem. They used a HMM to find sign boundaries. Cooper and Bowden [40] solved the same problem by applying efficient data mining methods, an idea that was introduced to the vision community by Quack et al. [41]. Other works use EM [42] to link text and image regions [43]. Wu et al. [44] introduced a non-linear kernel discriminant analysis step in between the expectation and maximisation step. This helped to map the features to a lower dimensional space and allowed the subsequent generative model to better separate the classes. In the field of Automatic Speech Recognition (ASR) we encounter the use of a discriminative classifier with EM [45]. Closely related is also the clustering of spatio-temporal motion patterns for action recognition [46] and Nayak's work on iterated conditional modes [47] to extract signs from continuous sentences. Learning from weak video level annotations is an under exploited approach in the vision community and the previous literature has several shortcomings that we address with this work:

1) Weak learning with specific temporal synchronisation constraints has not been tackled so far.
2) Multi-stream hybrid HMMs with state of the art CNNs and long short term memorys (LSTMs) have not been investigated before

3) Sign language has used multi-stream HMMs but only on very small tasks with weak models and never applied during training.

3 Weakly Supervised CNN-LSTM Training With Multi-Stream HMMs

The proposed algorithm constitutes a powerful solution to the problem of weakly supervised learning from noisy sequence labels to obtain strong and reliable frame labels. It significantly outperforms previous algorithms by adding synchronisation constraints between multiple streams of sequential classifiers learnt on weak labels. Figure 1 shows an example segment from the employed data set with three streams of annotations representing different modalities. It can be seen that each modality on its own has sequential nature, while all three considered jointly evolve in parallel with respect to the synchronisation points (vertical bars in Figure 1).

To better understand the proposed weak learning approach, let us focus on a single stream first. The way we process each stream in isolation is represented in Figure 2, which shows the overall pipeline using only a single modality and a single stream of labels. We assume that the video data is organised into segments of variable length with weak annotation labels available per segment. We consider these labels to be weak in the sense that we have multiple options of label sequences per segment without knowing which one is the correct one. The task is to find the right label sequence or discard all available sequences. Moreover, it is part of the task to find the exact beginning and end video frame of each symbol in the correct label sequence, i.e. align the annotation to the video. The process works as follows: To start with, the algorithm first considers a random weak label option (or the most likely, if prior data exists) to be the correct one. As an initial alignment guess, it linearly segments the available video with respect to the number of symbols in the chosen label sequence. Then we use this frame labelling to learn a CNN-LSTM model from it, which we term the maximisation step. We can then employ the model in a hybrid CNN-LSTM-HMM force alignment framework, which re-estimates the frame alignment. This is the expectation step. The important part is that we use the learnt model to choose the most likely weak label option by estimating the single best viterbi alignment. The algorithm then iterates between expectation and maximisation step and iteratively improves the frame labelling using the EM [42] method. After several re-alignment iterations, the algorithm converges and finally yields a sequence of labels which match the video frames well and a strong CNN-LSTM model, distinguishing the label classes.

3.1 From Single to Multi-Stream

The visual domain possesses highly parallel properties. However, the single-stream weakly learning approach presented in the previous section only makes use of sequential information. We want to exploit the parallelism in the visual medium and extend the approach to multiple streams. Our idea is to incorporate synchronisation points between independently evolving streams. Each stream models the sequential aspects of its modality while the synchronisation ensures that the streams evolve in parallel as shown in Figure 1.

Our proposed multi-stream approach is depicted in Figure 3. It shows how we incorporate sequential parallelism in the learning. To achieve this, we modify the expectation step and incorporate synchronisation constraints in the HMM that estimates the viterbi alignment. We are inspired by multi-stream HMMs introduced by Bourlard et al. [33] in the late ninetees, who used such kind of models for multi-band speech recognition (cf. Section 2). In our proposition, each stream is modelled in a hybrid fashion [48] where a CNN-LSTM estimates the HMM emission probabilities of its stream symbols. The HMM has independent streams that can evolve freely. But we introduce synchronisation points between the streams, which can only be reached by all streams at the same time. They do not resemble standard HMM states as they do not emit any symbols, but they recombine the posterior of all independent streams into a single posterior probability. The exact way this recombination is implemented is a design choice and will be a weighted sum in our case. To sum up, the multi-stream weakly learning approach represents an alteration of the alignment phase (expectation step). Each stream is a separate CNN-LSTM and during modelling (maximisation step) all streams have access to the input images which can be the same or different for each stream. We show that it is feasible to dismiss any preprocessing such as tracking as the neural networks are able to focus on the important parts of the input images based on the different stream labels describing each of the desired modalities.

3.2 What Problems Can We Apply this to?

As long as you can break a weakly supervised learning problem down into a number of parallel sub-problems where each on its own has sequential nature, our proposed joint multi-stream weak learning, can be applied. As an example, think of gesture recognition, where you have access to gesture labels. We are able to reformulate the problem as a combination of hand shape sequences, hand orientations and movements. Each such modality on its own occurs in sequence, but to form a specific gesture the occurrence in parallel is important. In a similar way, the task of activity recognition can be broken down into parallel sub-problems. Activities are composed of human actions with interacting objects. A specific activity requires actions and interacting objects to be present in parallel.

For this manuscript, we use the field of sign language recognition which serves as perfect test bed. It has firm properties of sequential parallelism. Sign linguists represent sign language as parallel co-execution of subunits across multiple modalities. Sign language theory defines four so-called manual parameters which consist of the hand shape, orientation, place of articulation and movement. Additionally, there are non-manual parameters such as mouth shapes, facial expression, head and upper body orientation. Each type of subunit is characterised by a limited number of units. However, combined they can represent an arbitrary number of signs. If we look at how users of sign language actually combine subunits, we notice that

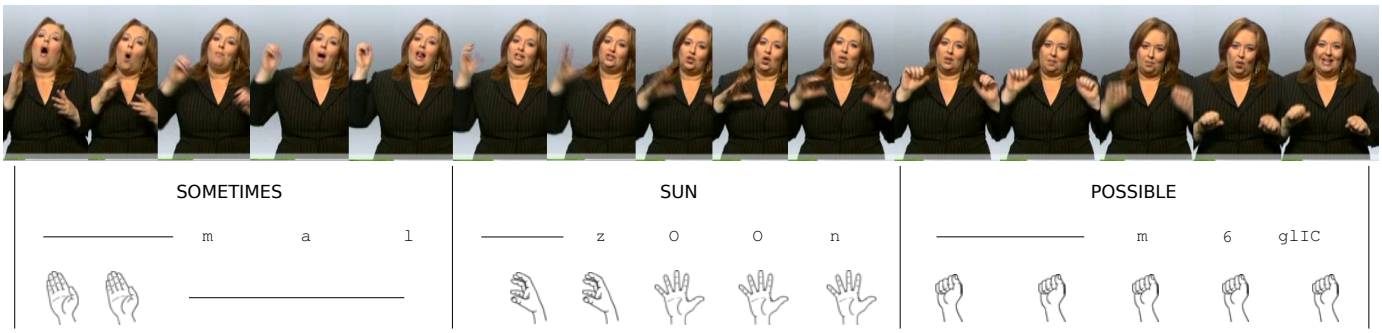

Fig. 1. Example showing from top to bottom: the a video segment of continuous sign language and the three aligned streams: the sign glosses, the mouth shapes described by phonemes and the hand shapes. Vertical bars illustrate the synchronisation constraints across all streams, horizontal bars represent the garbage class.

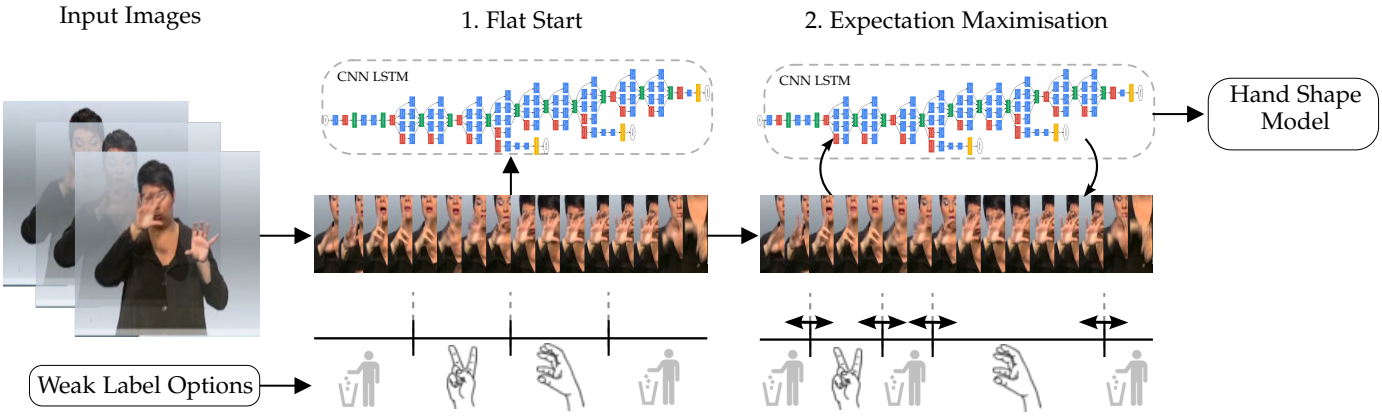

Fig. 2. Single CNN-HMM Stream. Showing initialisation and iterative label and temporal segmentation refinement in an expectation maximisation fashion. We first linearly partition the input stream (1. Flat Start), train a CNN-LSTM model and use this model to re-estimate a new segmentation.

the different modalities occur in parallel, e.g. a sign has a specific configuration of hand shape, movement and mouth shape. Nevertheless, within a single modality we often have sequential concatenations of different subunits, such as sequentially changing hand or mouth shapes. This can be verified in the example shown in Figure 1.

3.3 Problem Formulation

This paper tackles the problem of weakly supervised learning with application to continuous sign language recognition. Given a sequence of input images $x_1^T = x_1, \dots, x_t, \dots, x_T$ and weak annotations for each of the M modalities, the task of weakly supervised learning consists of attributing a matching frame label per modality to each x_t. Weak labels refer to annotations that are not accurate. They typically cover multiple annotation options per sample. One of those annotation options may match the data, or they may be wrong. Typically weak labels originate from automatic processes (automatic translation) or transferred annotations that have not been annotated to match the actual target data. Note, that we use $Pr(\cdot)$ to indicate true probability distributions while $p(\cdot)$ indicates model assumptions. The modality-specific labels are drawn from an inventory $I_m = \{c_{1m}, \dots, c_{cm}, \dots, c_{Cm}, \varnothing_m\}$ of $C_m + 1$ class symbols. Each inventory contains a separate garbage model $\varnothing_m$ to address cases when none of the weak labels match the data. We refer to these models as the garbage class. The symbols are represented by single state HMMs. In this work, we consider sign glosses, mouth shapes and hand shapes as modalities. Hand shapes, for example, encompass symbol classes such as a flat hand or a fist. We break sign language modelling down into mouth and hand shape modelling. We have no manually annotated labels for these modalities available. But we can make certain assumptions about what hand and mouth shapes may be valid candidates to occur with a certain sign. These assumptions are the basis of the weak labels employed in this work. We use these assumptions to generate a finite set of possible sign-mouth-hand combinations, which we call weak label options.

Let us formalise this: An input video segment x_1^T is known to contain the sign word classes $w_1^N = w_1, \dots, w_n, \dots, w_N$. The weak label options are stored in our modality-specific lexicon ψ_m which contains a variety of mappings from $w \rightarrow \tilde{c}$, where each $\tilde{c}$ is a sequence of L concatenated class symbols c_m,

$$\psi_m = \{w : \tilde{c}_1^L \mid \tilde{c} \in I_m\} \tag{1}$$

By combining the limited number of class symbols c_m we can express an unlimited number of sign words w. Each sign word w can map to multiple symbol sequences (which is important as $\tilde{c}$ is ambiguous and a one-to-one mapping would not be sufficient). In terms of sequence constraints, we require each symbol to span an arbitrary length of subsequent images as we assume that symbols are stationary

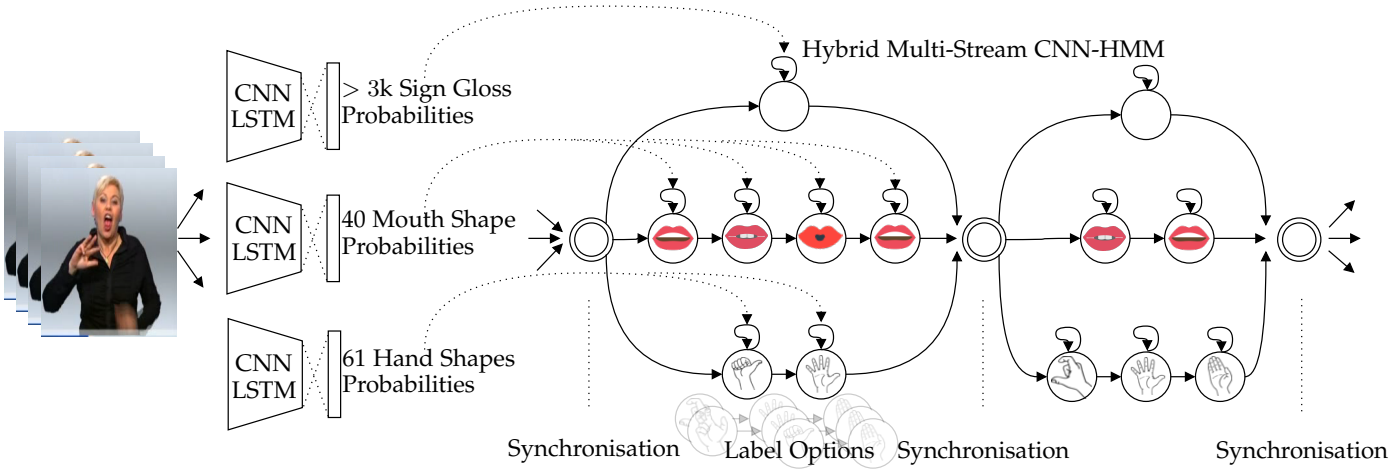

Fig. 3. Multi-stream (3-stream) CNN-HMM with synchronisation at the sign end. Three independent CNN-LSTM models are trained on the same full frame input, while having different loss functions yielding classifiers for sign-gloss, mouth & hand shape modalities. In a hybrid multi-stream HMM framework the networks model HMM emission probabilities. All streams can evolve different in time, but have to recombine at the sign ends which have been chosen as synchronisation points. The HMM is used to re-estimate the frame labelling, improving the modelling in several EM iterations.

and do not instantly disappear or appear. We maximise the m symbol sequences c jointly.

$$\left[c_1^M\right]_{\text{opt}} = \underset{c_1^M, c_m \in \psi_m}{\text{argmax}} \left\{ Pr(w_1^N, c_1^M | x_1^T) \right\} \tag{2}$$

Expressing this as a HMM and representing the words and symbols by modality-specific hidden states s_m, where each hidden state is mapped to a single input image, reveals the weakly supervised learning problem: It is exactly the hidden state sequence of each modality we are interested in. We synchronise each modalities hidden states only with respect to the N sign words. We only consider the best path (viterbi) and finally optimise:

$$\left[s_1^T\right]_{\text{opt}} = \underset{s_1^T}{\text{argmax}} \left\{ \prod_{n=1}^{N} \prod_{m=1}^{M} \max_{s_{t_{n-1}+1,m}^{t_n}} \prod_{t_{n-1}+1}^{t_n} p(x_t, s_{t,m} | s_{t-1,m}, w_n)^{\gamma_m} \right\} \tag{3}$$

Where t_n represents the end time of w_1 and $t_0 + 1$ points to the first image of the segment.

$$\sum_{m=1}^{M} \gamma_m = 1 \tag{4}$$

assures that the probabilities sum up to one, where γ_m is the stream-weight hyperparameter in the optimisation. Due to the discriminatory capabilities of CNN-LSTMs, we solve the problem in an iterative fashion with the EM algorithm [42] embedded in an HMM and use independent CNN-LSTMs for modelling $p(s_m|x)$. Following [48], we convert the CNN-LSTM posteriors into likelihoods by normalising with the priors of each stream. The priors are scaled by the prior scale hyperparameter β:

$$p(x_t, s_{t,m} | s_{t-1,m}, w_n) = \frac{p\left(s_{t,m} | x_t\right)}{p\left(s_m\right)^{\beta}} \cdot p\left(s_{t,m} | s_{t-1,m}, w_1^N\right) \tag{5}$$

$p\left(s_{t,m} | s_{t-1,m}, w_1^N\right)$ are transition probabilities, which we fix and pool across all classes and streams with the exception of the garbage classes that have separately pooled transition probabilities. We intuitively define the synchronisation points at the end of each sign word. This allows sequences of subunits to be found in the data with very weak labels which on their own would not contain enough discriminative power to be successfully identified in the data.

4 Data Sets

We work with several different data sets to validate the approach. All data sets are or will be made publicly available as of publishing this work. The well known challenging real-life continuous sign language data set RWTH-PHOENIX-Weather 2014 [49] [50] [51] constitutes the basis of our work. It covers unconstrained sign language of 9 different signers with a vocabulary of 1081 different signs. The corpus features sign language interpreters and has been recorded from broadcast news. It has been annotated using sign-glosses by deaf specialists. However, the corpus does not have mouth shape annotations. Therefore, following our previous work in [1] and [2], we exploit the correlation between spoken German and mouth shape sequences visible on the signers mouth. We therefore created RWTH-PHOENIX-Weather 2014 T [52], an extension of the previous RWTH-PHOENIX-Weather 2014. It constitutes a parallel corpus including sign language videos, sign-gloss annotations and also German translations (spoken by the news anchor), which are all segmented into parallel sentences. These segmentations originate mostly from PHOENIX 2014, however, several segments differ in length and number due to the different sentence structure of the German translations. Wherever new sentence boundaries were required, we used the PHOENIX 2014 models in [3] to estimate new boundaries based on a forced alignment on the data. Care has been taken to assure that no test and no dev segments from PHOENIX 2014 can be found in the PHOENIX 2014 T train set and also no test and no dev segments from PHOENIX 2014 T can be found in the PHOENIX 2014 train

set. Moreover, we ensured no segments that have been annotated for mouthing sequences in [1] to be present in any sets. Therefore, one can safely evaluate trained PHOENIX 2014 T models on the mouth shapes evaluation from [1] and also on the hand shape evaluation from [3]. The WERs of PHOENIX 2014 and PHOENIX 2014 T are similar but not exactly comparable, as they are calculated on a set with slightly different boundaries and number of segments. Most of our experiments are done on RWTH-PHOENIX-Weather 2014 T, as this allows evaluation on three modalities, namely glosses, mouth shapes and hand shapes. We repeated the best working setup on RWTH-PHOENIX-Weather 2014 to compare against the state-of-the-art. In order to evaluate the mouth shape and hand shape recognition performance, we compete in the mouth shape weakly learning task [1] and the 1 million hands weakly learning task [3]. The former consists of 3687 labelled mouth shapes, annotated from continuous sign language sequences with 11 viseme class labels or a garbage label, while the latter contains 3361 hand shape images that are annotated with one out of 45 orientation independent hand shape class labels. To sum up, this work employs the following data sets:

- RWTH-PHOENIX-Weather 2014 T [52]
- RWTH-PHOENIX-Weather 2014 [51]
- mouthing weakly learning task [1]
- 1 million hands weakly learning task [3]

Statistics of the RWTH-PHOENIX-Weather 2014 T data set is given in Table 1. Figure 6 and Table 2 show the statistics of the 1 million hands data set and mouthing data set, respectively.

TABLE 1
Statistics of the PHOENIX 2014 T recognition ("Gloss") & translation ("German") set.

	PHOENIX 2014 T Gloss			PHOENIX 2014 T German		
	Train	Dev	Test	Train	Dev	Test
# segments	7,096	519	642	7,096	519	642
frames	827,354	55,775	64,627	827,354	55,775	64,627
tot. words	67,781	3,745	4,257	99,081	6,820	7,816
vocabulary	1,085	393	411	2,887	951	1,001
OOV [%]	-	0.5	0.5	-	0.8	0.7
singletons	337	-	-	1,077	-	-

4.1 Creating Weak Mouth Shape Labels

Sign languages and their spoken counterparts do not share the same word order, nor does one word always translate to exactly one sign. Spoken German typically follows the 'subject (S), verb (V), object (O)' structure, while DGS quite strictly uses 'SOV'. We want to exploit the fact that mouthings in sign language often originate from contact with speech. In our corpus PHOENIX 2014 T, for each video segment we have two annotations comprising a sequence of sign-glosses which are in the order of the signs in the video and a translated sentence of spoken German words. Following [1], we employ a forced alignment technique from statistical machine translation presented in [53], which maximises the alignment likelihood on a training corpus of sentence pairs each with a pair of sequences of German words $\mathbf{w} = w_1^J := w_1, \ldots, w_J$ and DGS glosses $\mathbf{g} = g_1^I := g_1, \ldots, g_I$ $(\mathbf{w}, \mathbf{g})$. The alignment variable $\mathbf{a} = a_1^J$ describes the mapping from a source position j to a target position a_j for each sentence pair. The applied approach finds the best Viterbi alignment by maximising the statistical alignment model p_θ, which depends on a set of unknown parameters θ that is learnt from the data:

$$\hat{a}_1^J = \arg\max_{a_1^J} p_\theta(w_1^J, a_1^J | g_1^I) \tag{6}$$

The technique includes the so called IBM Models as alignment models, which account for lexical translation and reordering. For more details refer to [53]. We use the single best alignment words. To cope with noise in those alignments, we apply filtering to the generated (w, g) pairs constituting a mapping $M : \mathcal{G} \rightarrow \mathcal{P}(\mathcal{W})$, where $w \in \mathcal{W} = \{\text{all spoken words}\}$ and $g \in \mathcal{G} = \{\text{all sign glosses}\}$. As we expect our CNN-LSTM-HMM modelling scheme to be much stronger than the GMM-HMMs employed in [1], we use a softer filtering to remove noisy annotations. We apply only an absolute filtering ϑ_A threshold and no relative filtering, such that

$$M(g)' = \{w \in M(g) \,|\, c(w, g) > \vartheta_A\}, \tag{7}$$

where $c(w, g)$ counts the number of occurring pairs (w, g) and ϑ_A is set to 2 instead of 4 as in [1].

Based on the German words that each gloss now maps to, we can build a pronunciation lexicon, which defines the finite set of possible pronunciations that occur with a sign. For this purpose we use a standard automatic speech recognition (ASR) word to phoneme lexicon which has been generated with the publicly available Sequitur Grapheme-to-Phoneme converter [54] trained on a standard ASR task.

However, the mouthings produced by signers often do not constitute fully pronounced words, but rather short fragments of words. Thus, for each full pronunciation we add multiple shorter pronunciations to our lexicon ψ by truncating the word w which consists of a sequence of phonemes $s_1^N = s_1, \ldots, s_N$, such that

$$\psi = \{w' : s_1^{N-\phi} | \phi \in \{0, \ldots, \phi_{trunc}\} \wedge N - \phi \geq \phi_{min}\} \tag{8}$$

Moreover a garbage or 'Non-mouthing' option is added to the lexicon for each entry. See an exemplar entry for the sign-gloss SONNE (english: SUN) which has been aligned to the German adjective sonnig (sunny), with three differently truncated phonetic pronunciations (in SAMPA) and a garbage option as they are added to the lexicon:

```
SONNE#sonnig :
{ /z O n I C/, /z O n I/, /z O n/, GARBAGE }
```

The phonetic inventory consists of 39 phoneme classes and a garbage class.

TABLE 2
Annotation statistics for the RWTH-PHOENIX-Weather continuous mouthing challenge [1] Annotation fraction [%] for each of the employed 11 visemes. 'gb' denotes non-mouthings.

Total frames	A	E	F	I	L	O	Q	P	S	U	T	gb
3687	12.4	8.2	8.8	10.6	4.4	11.9	13.0	9.4	4.3	7.8	22.6	48.6 %

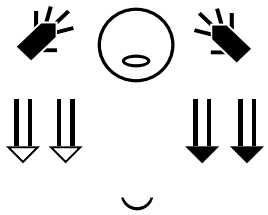

Fig. 4. A SignWriting entry describing the sign RAIN in DGS.

4.2 Creating Weak Hand Shape Labels

There are no hand shape annotations available for the RWTH-PHOENIX-Weather 2014 data. Therefore, following [3], we employ a publicly available user-edited sign language dictionary called **SignWriting** [55]. It constitutes an open online resource, where people can create entries translating from written language to sign language using a pictorial notation form called SignWriting. The sign writing entries also contain hand information. At the time, when processing the German SignWriting lexicon it comprised 24.293 entries. Inspired by [56], we parsed[1] all entries to create the mapping ψ from sign annotations to possible hand shape sequences, where we removed all hand pose related information (such as rotations) from the hand annotations. SignWriting is a universal notation for sign languages developed by Valery Sutton in 1974. It uses the International SignWriting Alphabet 2010, which represents manual and non-manual parts of signs by a set of visual symbols classified in a hierarchical system comprising a total of 652 icon bases. Each base has several degrees of freedom when used in writing a sign: It can be rotated, mirrored and put in context with other parts of the sign (i.e. a right hand). The SignWriting subunit nomenclature consists of a starting "S" and five following digits. The first three digits specify the base symbol, whereas the last two represent its degree of rotation and its state of being mirrored or not. SignWriting bears, due to its stylised nature, little resemblance to continuous signing, but has been used for 3D avatar animation [57] before. Furthermore, SignWriting has redundancy. The same signs can be written in a variety of ways. The SignWriting dictionary is user-edited, published under Creative Commons license and can be freely downloaded in XML format. Each dictionary entry is encoded as a Formal and Regular SignWriting (FSW) code and contains the symbols and their position used to write specific signs. The dictionary is available for over 80 different sign languages, but within the context of this work, only the German Sign Language database is considered. Figure 4 shows the entry of a signing variant for RAIN. Despite the large number of entries in the database, only those entries matching the inventory of the RWTH-PHOENIX-Weather corpus are of interest. Similar to the weak mouth shape labels in Section 4.1, we create a lexicon with mappings from gloss-signs g to possible hand shape sequences h also including the garbage hand shape class $\varnothing$.

$$\psi_{hand} = \{g' : h | g' \in \mathcal{G}, h \in \{h_1 \dots h_H, \varnothing\}\} \tag{9}$$

1. Parser available at: www.hltpr.rwth-aachen.de/~koller/

Fig. 5. 12 exemplary manually annotated hand shape classes are shown. Three labelled frames per class demonstrate intra-class variance and inter-class similarities. Hand-Icons from [58].

It is clear that this kind of mapping from sign-glosses to hand shapes has to be considered a very weak annotation as it has not been created or manually refined for the data set it is finally applied to.

Construction of Lexicon. The next step is to construct the lexicon ψ, given the hand shape annotations. If a sequence of more than one hand shape annotation is available for a given video, we add the whole sequence and each of the hand shapes on its own to the lexicon ψ. This results in multiple hand shape annotations per video, all of which we add to the lexicon ψ. Within the lexicon definition, we also allow the garbage class to be able to account for frames before and after any hand shape.

Throughout this work we follow a hand shape taxonomy by the danish sign language association, which amounts to over 60 different hand shapes, often with very subtle differences such as a flexed versus straight thumb.

Evaluation of hand shapes. To evaluate the final CNN-LSTM hand shape classifier, we chose the challenging articulated hand shape evaluation from [3]. It consists of 3361 manually labelled images from the RWTH-PHOENIX-Weather 2014 Development set. Some of the 45 encountered pose-independent hand shape classes are depicted in Figure 5. They show the large intra-class variance and the strong similarity between several classes. The hand shapes occur with different frequency in the data. The distribution of counts per class can be verified in Figure 6 showing that the top 14 hand shapes explain 90% of the annotated samples.

5 Experimental Evaluation

The aim of this section is to understand the effects of the proposed multi-stream CNN-LSTM-HMM architecture for weakly supervised learning. We consider the case of a single-stream HMM, a 2-stream HMM and a 3-stream HMM. As stated in Section 4, we evaluate the presented approach on three different tasks and data sets, namely continuous sign language recognition, articulated mouth shape and hand shape recognition. We test on the real-life continuous sign language recognition corpus RWTH-PHOENIX-Weather 2014 T [52], the continuous sign language mouth shape data set [1] and the 1-million-hands articulated hand shapes data set [3]. Our focus is on weakly supervised learning, e.g. discovering the articulated signs/mouth

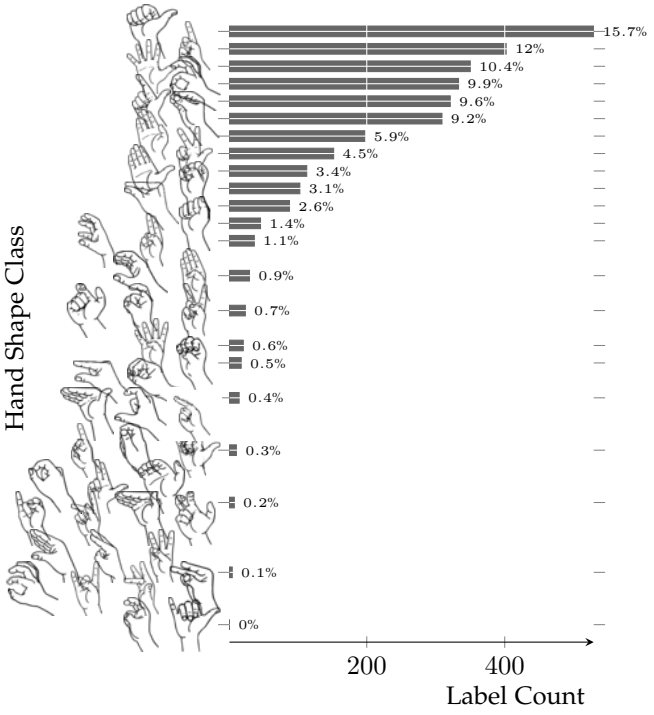

Fig. 6. Ground truth hand shape label count of all 3361 annotations. 45 out of 60 classes are present in the data & could be labelled. If several hand shapes appear close to one label counting bar, each hand shape alone amounts to the mentioned fraction of labels. Hand-Icons from [58].

shapes/hands in the data. Therefore, we model each stream and each classifier's probabilities independently from the others. The main contribution lies in the multi-stream architecture and the joint multi-stream alignment with a state-of-the-art CNN-LSTM-HMM as described in detail in Section 3.

Data preparation. In this work, we show that the proposed approach does not require tracking of the hands, face or mouth regions to learn the corresponding classifiers. However, we still want to evaluate the effect of this design decision and therefore additionally create experiments with tracked inputs. As general preprocessing, we perform a pixelwise mean subtraction based on the statistics of the training set. The main gloss stream always gets the full body as input. The mouth and hand streams either get the same input or learn from tracked hands and face regions. For tracking hands, we employ a model-free dynamic programming tracker and follow our previous work [3]. For the mouth stream, we track facial landmarks with an Active Appearance Model (AAM) as in [2] and crop the image based on the smallest rectangular region covering all AAM tracked points representing the whole face. All streams are scaled to 256x256 pixels and randomly cropped to 224x224 pixels during training.

Lexicons that represent weak labels are necessary for our weak learning scheme. The creation of the mouth shape lexicon ψ_{mouth} has been specified in Section 4.1 and the hand shape lexicon ψ_{hand} in Section 4.2 respectively. The main sign language data set PHOENIX 2014 T has been annotated with a sequence of sign-gloss annotations without explicit start and end times. These annotations provide the strongest supervision. The mouth and hand shape annotations are represented by mappings from these sign-glosses to sequences of mouth and hand shapes. This is represented as one to many mapping, where the entries for a specific sign-gloss also comprise a garbage class to discard this annotation for a given modality. We also allow the insertion of garbage models at the beginning and the end of each modality's mappings to account for temporally different starting and ending points of the mouth shapes with respect to the hand shapes.

Initialisation of the algorithm. The input videos are linearly partitioned (e.g. a flat start) based on a single specific mouth and hand shape label sequence from the lexicons ψ_{mouth} and ψ_{hand}, considering the beginning and end of each segment as garbage class.

HMM settings. We base the HMM part of this work on the freely available state-of-the-art open source speech recognition system RASR [59] for which we have implemented a multi-stream alignment procedure. The 1200 modelled gloss classes and the 60 hand shape classes are represented by a three repeated states where two consecutive states share their probability distributions, whereas the stream-dependent garbage class is always represented by a single state for higher flexibility. For the 39 mouth shapes we employ a single repeated state to accommodate the high speed of mouth shape changes. This results in 3601, 40 and 181 softmax outputs for the gloss, mouth and hand stream, respectively. For forced alignment, we use fixed, non-optimised transition penalties being '1-0-1.5-2' for 'loop-forward-skip-exit' for all HMM states in all streams and '1.5-0-2' for the garbage 'loop-forward-exit' penalties. The usage of fixed, pooled transition penalties has become standard in ASR and Automatic Sign Language Recognition (ASLR). The exit penalties are applied to the hypothesis score when the last state in a mapping has been reached and the stream arrived at a synchronisation point. The prior scaling factor β is set to 0.6 in our experiments for all streams. As already pointed out by [60], we also observe a strong bias in the distribution of hand shape classes in our data, but we decided to maintain it. We expect the Bayesian conversion from posteriors to scaled likelihoods to account for this fact, as described in Section 3.3. We apply different stream-weights λ, which have been optimised on the held out data with a simple grid search. For 2-streams we obtain best results with $\lambda_{\text{sign-gloss}} = 0.8$ and $\lambda_{\text{hand/mouth}} = 0.2$, while in the three stream case we get best results with $\lambda_{\text{sign-gloss}} = 0.7$, $\lambda_{\text{mouth}} = 0.2$ and $\lambda_{\text{hand}} = 0.1$.

CNN-LSTM training. For the neural network training, we extend a powerful and deep CNN with two bi-directional LSTM layers [61], [62]. In order to train the full network end-to-end, the CNN architecture of choice should have a low memory footprint, while still being deep. After comparing different CNN architectures [63], [64], [65], we opted for the 22 layer deep GoogleNet [65] architecture, which we initially pre-train on the 1.4M images from the ILSVRC-2012 [66]. The main building blocks of this architecture are inception modules which are the fusion of multiple convolutional layers with different receptive fields applied to the output of a 1x1 convolution layer which serves as a dimensionality reduction tool. Finally, in addition to the

last classifier, GoogLeNet also makes use of two auxiliary classifiers at lower layers which are added to the final loss with a weight of 0.3. The pre-trained standalone CNN achieves a top-1 accuracy of 68.7% and a top-5 accuracy of 88.9% in the ILSVRC. The network uses ReLUs as non-linearity in its convolutional layers and dropout on the fully connected layers preceding the softmax layers. Dropout ratio is set to 70% on the auxiliary classifiers and 40% on the final classifier to prevent over-fitting.

LSTMs are RNN variants that were developed to overcome the vanishing gradient problem [67] and as such can learn long time dependencies much better than vanilla RNNs. As the gradients are fully differentiable, we can train the recurrent network with Back Propagation Through Time (BPTT) [68]. We use stochastic gradient descent with an initial learning rate $\lambda_0 = 0.001$ for CNN-LSTM architectures and $\lambda_0 = 0.01$ for CNN networks. We employ a polynomial scheme to decrease the learning rate λ_i for iteration i as the training advances while reaching $\lambda_{i_{max}} = 0$ for the maximum number of iterations $i_{max} \mathrel{\widehat{=}} 4$ epochs in our experiments.

$$\lambda_i = \lambda_0 \cdot \left(1 - \frac{i}{i_{max}}\right)^{0.5} \tag{10}$$

Training scheme and scrambled start.

Comparable to [4] and with a similar effect as 'curriculum learning' [69], we observe fastest convergence if we first train a CNN-HMM model with randomly shuffled input images for 4 EM re-alignment iterations starting from a flat start in each stream. Each re-alignment iteration finetunes from the previous one. This yields a sensible temporal alignment. However, remember that for the flat start we had to choose a specific weak label sequence from the set of options. We noticed a skew towards the selected label option. To overcome this, we found it crucial to restart the neural network weight learning from scratch (e.g. from pretrained GoogleNet), but take the new temporal segmentation start and end boundaries for each class in an utterance as intialisation points for a 'scrambled start'. In the lexicons ψ_{mouth} and ψ_{hand} we find multiple weak label options per sign. For a 'scrambled start', we use all of them for initialisation. Each label option allows us to generate one initialisation option with pseudo labels. We do this by linearly distributing the class states to the image frames within the temporal boundaries (flat start with temporal boundaries). We then unify all initialisation options by scrambling and mixing them together. This means we go over each frame of the video. On the first frame we take a pseudo-label from the first weak label initialisation option. On the next frame we take it from the second option and continue accordingly. We train from the given pseudo-label initialisation for 3 EM re-alignment iterations.

After that, we add 2 bi-directional long short term memory (BLSTM) layers to the final classifier (maintaining the first two auxiliary classifiers with standard feed forward structure and a weight of 0.3) and finetune with ordered samples and a batch size of 32 for 5 EM re-alignment iterations.

We perform multi-stream alignment on the dev data as well and use it as an automatic validation measure to verify the learning and choose the best epoch (which is usually the last). In previous experiments [3], we found that the automatic and manual validation converged in a similar fashion, which is why we only rely on automatic validation in this work.

In terms of **run time**, the CNN-LSTM trains with over 170 frames per second (fps) on a NVIDIA Titan X with batch sizes of 32 images. The multi-stream alignment with minor pruning takes roughly 2 seconds in the 2-stream case and 248 seconds in the 3-stream case per sentence on a single core AMD Opteron 6176 Processor, where each sentence contains on average 10 sign-glosses.

Please note, that even though we trained a multi-stream system and used that for efficient weakly supervised learning, we discard the non-related streams when evaluating. Modelling the stream posteriors conditioned on each other can have a positive influence as initial tests revealed. However, a thorough analysis is beyond the scope of this manuscript.

5.1 Task 1: Continuous Sign Language Recognition

Sign language recognition (SLR) is a very challenging test case for weakly supervised sequence learning algorithms. It is a difficult but well defined problem offering real-life challenges (w.r.t. occlusion, motion blur, variety of hand shapes and other articulators). We provide here experiments on the RWTH-PHOENIX-Weather 2014 T data set, as this allows us to evaluate 3-stream modelling. But in order to compare with the state-of-the-art, we also provide results on the RWTH-PHOENIX-Weather 2014 corpus. There, we use exactly the same corpus partitioning as in [51] to ensure comparability to previously published results. We perform hybrid CNN-LSTM-HMM recognition. We measure the error in WER:

$$\text{WER} = \frac{\#\text{deletions} + \#\text{insertions} + \#\text{substitutions}}{\#\text{number of reference observations}} \tag{11}$$

Figure 7 shows the WER measured on the PHOENIX 2014 T dev corpus as a function of the EM re-alignment iterations. The graph shows 1-stream (using sign-glosses during training only), 2-stream (using sign-glosses and mouth shape information during training) and 3-stream setups (using sign-glosses, mouth shape and hand shape information during training). One can clearly see that multi-stream networks result in a much faster convergence, which is due to the improved alignment. Furthermore, we see that each additional stream helps convergence and also results in a lower final error rate. As each re-alignment takes around 6 hours for CNN-LSTM training. Table 3 provides the best achieved WERs with each of the architectures. The 3-stream architecture clearly outperforms the single stream approach on the dev set with 22.1% WER (the lower the better) as opposed to a WER of 24.5% for the single-stream. On the test set, it reaches 24.1% as opposed to 26.5%. The rise in WER at iteration 5 is due to the scrambled start which uses the temporal segmentation and subsequent training from scratch of the neural network parameters. The improvements with respect to the state-of-the-art can be explained by better initialisation using the scrambled start and stronger alignments during training due to the multi-stream architecture.

We have tested the results for statistical significance using the Matched Pair Sentence-Segment Word Error Rate (MAPSSWE) test. The MAPSSWE test was suggested for speech recognition evaluations by Gillick [70]. We used the implementation [71] as part of the NIST evalutation tools. Testing the 1-stream baseline against the 3-stream case, finds strong statistical significance at the level of $p < 0.001$. Testing the baseline against the 2-stream case also finds statistical significance with a level of $p = 0.021$.

We further evaluate the effect of tracking hands and faces for the respective streams and find that tracking helps. The WER drops from 23.1% on dev and 24.8% on test without tracking to 22.1% and 24.1% with tracking. Hand tracking is done using a dynamic programming based model-free tracker as in [51]. Facial landmarks are tracked based on AAMs as presented in [72]. We crop the smallest rectangle containing all landmarks. Note, that the tracked input is only used for the hand and the face stream and the alignments. The gloss stream, which we evaluate on, operates on full images.

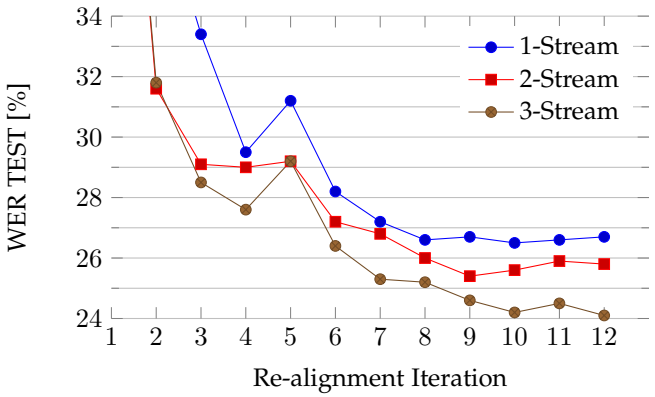

Fig. 7. Dev set WER [%] on glosses with cropped hand and face inputs: compare 1stream against 2stream (gloss, mouth) and 3stream (gloss, mouth, hand)

TABLE 3
Achieved word error rates on RWTH-PHOENIX-Weather 2014 T continuous sign recognition corpus with cropped hand and face inputs

	PHOENIX 2014 T	
Setup	Dev	Test
1 stream CNN-LSTM-HMM	24.5	26.5
2 stream CNN-LSTM-HMM	24.5	25.4
3 stream CNN-LSTM-HMM	22.1	24.1

Table 4 shows the results on the PHOENIX 2014 corpus. It allows a direct comparison to recently published state-of-the-art approaches. The best previously published state of the art [73] achieved 39.4% WER on the dev set and 38.7% on the test set. In [4], we achieved 27.1% WER on the dev set and 26.8% on the test set using a single gloss stream. However, this result was generated relying on an external GMM-HMM alignment to start. For comparison, we provide the WER achieved on a single-stream system, which represents the same setup as in [4]. If we dismiss any external alignments and instead start the learning from scratch, the hybrid single-stream achieves 27.5% and 28.3% WER on the dev and test corpus respectively. The 2-stream sign-gloss and mouth shape system outperforms both results with 26.0% WER on the dev set and 26.0% on the test set. We cannot report 3-stream results, as PHOENIX 2014 has no translations available that can be used to infer the mouthings.

TABLE 4
Comparison to the state of the art on RWTH-PHOENIX-Weather 2014 continuous sign recognition corpus, showing the 1-stream and the 2-stream systems. Note, that no external alignment has been used to generate the multi-stream results.

	Tracked	External	PHOENIX 2014	
Method	hands	alignment	Dev	Test
CNN-LSTM with CTC [74]	yes	no	40.8	40.7
CNN-HMM [75]	yes	yes	38.3	38.8
CNN-LSTM with CTC [73]	yes	no	39.4	38.7
CNN-HMM 1-stream [5]	yes	no	33.6	34.6
CNN-HMM 1-stream [4]	no	yes	29.0	29.4
CNN-LSTM-HMM 1-stream [4]	no	yes	27.1	26.8
CNN-LSTM-HMM 1-stream	no	no	27.5	28.3
CNN-LSTM-HMM 2-stream	no	no	26.0	26.0

5.2 Task 2: Sign Language Mouth Shape Detection

Figure 8 demonstrates the benefit of the multi-stream HMM for weakly supervised learning of mouth shapes. It shows the precision and recall (on the equal error point) for a 1-stream model trained with mouth shape information only, a 2-stream model additionally using sign-gloss annotations and a 3-stream system that also adds the hand shape stream. Note, that the sign-gloss stream provides the strongest supervision, while the hand shapes constitute the signal with the weakest supervision. They originate from a completely unrelated data set as described in Section 4. Figure 8 shows, how the mouthing stream alone ('1-Stream') is strongly outperformed by the 2-stream system. This is due to the fact, that mouth shapes on their own are difficult to align to the data, particularly if we have only access to weak labels originating from the spoken German translations and no labels that actually represent what can be read on the signers lips. The 2-stream system yields 56.8% precision and recall, while the 1-stream system only achieves 47.5%. The third stream (hand shapes) only marginally improves the 2-stream result with 57.0%. This is due to the low degree of supervision this stream provides. Table 5 shows a comparison against previously published results on the continuous mouth shape data set. It shows two previously published results, that relied on an unpublished corpus for training the approach, which contained better quality labels. Additionally, the previously published results relied on heavy pre-processing by an AAM based face tracker that allowed extraction of the mouth region. Our proposed 3-stream approach clearly outperforms the previous approaches, without requiring any preprocessing (we do not use tracking).

5.3 Task 3: Articulated Hand Shape Recognition

In this subsection we present the validation of our method measured on the 1 million hand shape data set from [3].

TABLE 5
Previously published results on the mouthing sequences. Clean Set comprises no frames labelled as garbage.

Method	Corpus for training	AAM facial tracker	Tuned directly on test set	Clean Set		All Set	
				Precision	Recall	Precision	Recall
GMM-HMM [1]	unpublished	yes	yes	47.1	48.2	31.3	43.2
GMM-HMM, then CNN-HMM [2]	unpublished	yes	yes	55.7	55.6	–	–
1-stream CNN-LSTM-HMM	PHOENIX 2014 T	no	no	47.5	47.5	36.6	50.9
2-stream CNN-LSTM-HMM	PHOENIX 2014 T	no	no	56.8	56.8	42.8	59.2
3-stream CNN-LSTM-HMM	PHOENIX 2014 T	no	no	57.0	57.0	42.9	59.5

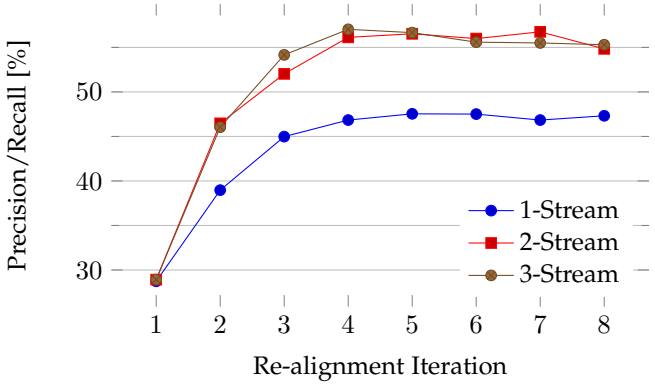

Fig. 8. mouthings (precision/recall), clean conditions: compare 1stream against 2stream (gloss and mouth) and 3stream (gloss, mouth, hand)

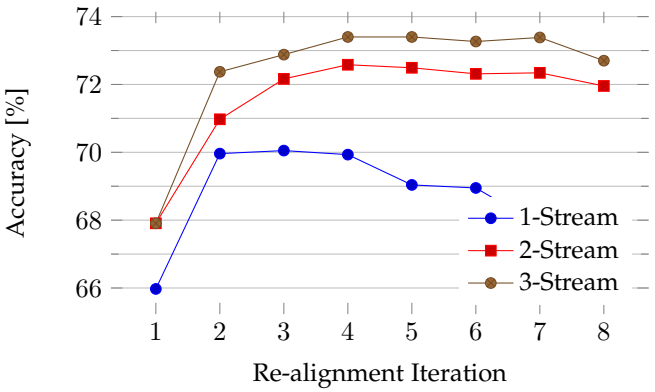

Fig. 9. hands (accuracy): compare 1stream against 2stream (gloss and hand) and 3stream (gloss, mouth, hand)

Figure 9 shows the results across several re-alignment iterations of a hand-only 1-stream system, a 2-stream system additionally trained with sign-glosses and a 3-stream system based on hands, sign-glosses and mouth shapes. We clearly see, how quickly the 1-stream degrades after the peak around the third EM iteration. This is due to drift in the re-alignment process and underlines the weak nature of the labels which originate, as stated in Section 4, from the SignWriting dictionary. As such, the mapping has not been created specifically for the PHOENIX 2014 T data set. When adding the second and the third stream we see, how convergence becomes much more stable. Drift is more difficult (although it still occurs after several iterations) in multi-stream hidden Markov models (HMMs) as the synchronisation constraints prevent it. One can also note that both the sign-gloss stream with stronger supervision, but also the weaker mouth stream significantly improve the final result.

In Table 6, we compare the multi-stream approach against the state-of-the-art on this data set. Note, although [3] used two additional sign language resources for training, namely a Danish and a New Zealand sign language dictionary, we only employ PHOENIX 2014 T. The added sign language dictionaries cover single word signs with weak hand shape annotations. In [3] a curriculum learning strategy is employed that first trains on these samples providing stronger supervision and then starts to train on the much weaker PHOENIX 2014 labels. Nevertheless, the proposed 3-stream approach is able to outperform the previous approach. In Figure 10, we used a gradient-weighted class activation mapping [76] to highlight those image regions that showed the highest activations in the last convolutional layer of the CNN-LSTM when classifying the sequence. In each of the three rows we show the activation maps of a different stream. It illustrates how the sign-gloss model focuses on both hands and face while the mouth model and the hand model focus on their respective modalities even though they were all trained on the same input.

TABLE 6
Comparison of proposed multi-stream CNN-LSTM-HMM against previously published results on the 1-million-hands test data.

Method	Train corpus	Acc. [%]
CNN-HMM [3]	PHOENIX 2014 + Danish + n. Zealand	62.8
Proposed 2-stream	PHOENIX 2014 T	72.6
Proposed 3-stream	PHOENIX 2014 T	73.4

6 Conclusion & Future Work

In this work we presented an approach to the field of weakly supervised learning in the video domain. Our method exploits sequence constraints within a set of independent streams and combines them by explicitly imposing synchronisation points to make use of parallelism present in visual data. We do this by learning multi-stream HMMs while adding intermediate synchronisation constraints among the streams. We embed powerful CNN-LSTM models in each of the HMM streams following the hybrid approach. The hybrid approach allows us to constrain the LSTM context to a feasible length that fits in modern GPUs, while training very deep CNN-BLSTM models end-to-end.

We apply the approach to the domain of sign language recognition exploiting the sequential parallelism to learn a sign language, mouth shape and hand shape classifier. We

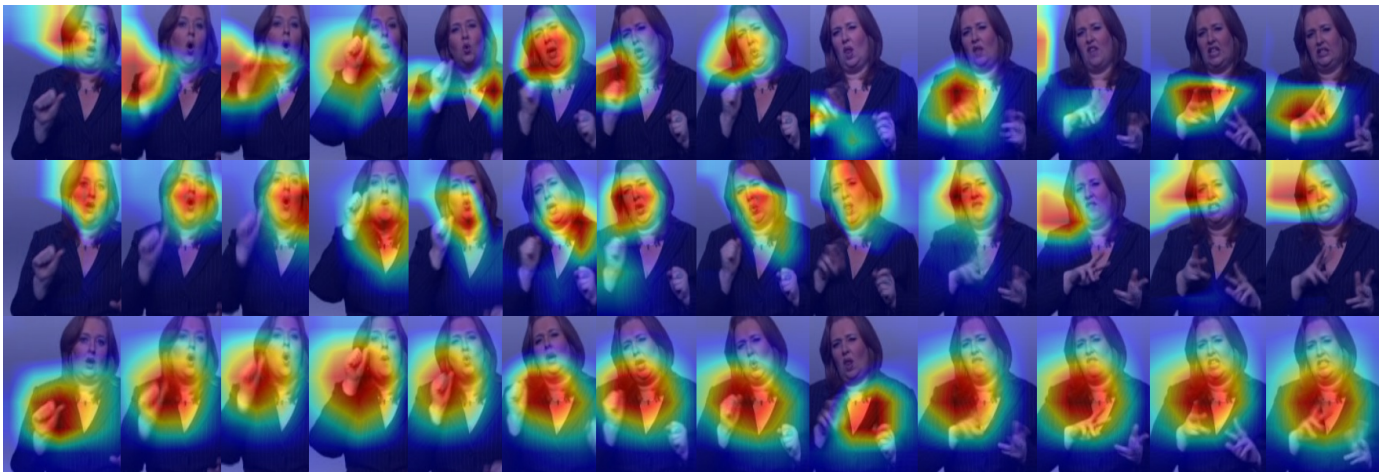

Fig. 10. Saliency maps showing the CNN-LSTMs highest activations. Generated with Grad-Cam [76], from top to bottom: sign-gloss stream, mouth shape stream, hand shape stream. It can be seen that the hand model focuses on the right hand, the mouth model on the mouth and the sign-gloss swaps between both hands and the mouth.

evaluate the classifiers on three publicly available benchmark data sets featuring challenging real-life sign language data with over 1000 classes, full sentence based lip-reading and articulated hand shape recognition on a fine-grained hand shape taxonomy featuring over 60 different hand shapes. We clearly outperform the state-of-the-art on all data sets, improving the best published WER on RWTH-PHOENIX-Weather 2014 to 26.0% and observe significantly faster convergence. We also looked at the activation maps of the learned CNN-LSTMs and can see that each stream focuses on its respective modality even though they were all trained on the same input.

There are several routes that are worth exploring in terms of future work. It would be interesting to extend the approach to the recognition search problem. We could limit the computational complexity by limiting the maximal offset between the streams. Additionally, the proposed model has some limitations: Until now, we do not rely on dynamic weighting of the streams, which would assume that certain streams are experts for certain classes and could further boost results. Moreover, computational complexity of the current implementation makes it difficult to go beyond 4 streams. A parallel GPU implementation seems promising. Also, the multi-stream boundaries (e.g. start and end of the sub-streams) could be more flexible. In sign language we observe mouth shape sequences can overlap with the next sign, which is not supported by our proposed model. Finally, we also noticed that our hand annotations are very weak and noisy. The fact that we did not particularly focus on one hand (e.g. through tracking and cropping) was also responsible for this issue. It might therefore be worth modelling both hands explicitly.

Acknowledgements

This work was supported by the SNSF Sinergia project "Scalable Multimodal Sign Language Technology for Sign Language Learning and Assessment" (SMILE) grant agreement number CRSII2 160811, the European Unions Horizon2020 research and innovation programme under grant agreement no. 762021 (Content4All) and the EPSRC project ExTOL (EP/R03298X/1). We would also like to thank NVIDIA Corporation for their GPU grant.

References

[1] O. Koller, H. Ney, and R. Bowden, "Read My Lips: Continuous Signer Independent Weakly Supervised Viseme Recognition," in *Proceedings of the 13th European Conference on Computer Vision*, Zurich, Switzerland, Sep. 2014, pp. 281–296. 1, 6, 7, 8, 12

[2] ——, "Deep Learning of Mouth Shapes for Sign Language," in *Third Workshop on Assistive Computer Vision and Robotics, ICCV*, Santiago, Chile, Dec. 2015. 1, 6, 9, 12

[3] ——, "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, Jun. 2016, pp. 3793–3802. 1, 2, 6, 7, 8, 9, 10, 11, 12

[4] O. Koller, S. Zargaran, and H. Ney, "Re-Sign: Re-Aligned End-to-End Sequence Modelling with Deep Recurrent CNN-HMMs," in *IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, HI, USA, Jul. 2017. 2, 3, 10, 11

[5] O. Koller, S. Zargaran, H. Ney, and R. Bowden, "Deep Sign: Enabling Robust Statistical Continuous Sign Language Recognition via Hybrid CNN-HMMs," *International Journal of Computer Vision*, vol. 126, no. 12, pp. 1311–1325, Dec. 2018. 2, 11

[6] L. E. Baum and T. Petrie, "Statistical inference for probabilistic functions of finite state Markov chains," *The annals of mathematical statistics*, vol. 37, no. 6, pp. 1554–1563, 1966. 2

[7] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989. 2

[8] H. Bourlard, S. Dupont, and C. Ris, "Multi-stream speech recognition," Idiap, Tech. Rep., 1996. 2, 3

[9] H. J. Nock and S. J. Young, "Modelling asynchrony in automatic speech recognition using loosely coupled hidden Markov models," *Cognitive Science*, vol. 26, no. 3, pp. 283–301, 2002. 2

[10] S. Dupont and J. Luettin, "Audio-visual speech modeling for continuous speech recognition," *IEEE Transactions on Multimedia*, vol. 2, no. 3, pp. 141–151, Sep. 2000. 2

[11] S. Bengio, "An Asynchronous Hidden Markov Model for Audio-Visual Speech Recognition," IDIAP, Tech. Rep. Idiap-RR 02-26, Jul. 2002. 2

[12] M. Brand, *Coupled Hidden Markov Models for Modeling Interacting Processes*. MIT Media Lab Perceptual Computing/Learning and Common Sense Techincal Report 405 (Revised), 1997. 2

[13] C. Vogler and D. Metaxas, "Parallel hidden markov models for american sign language recognition," in *Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference On*, vol. 1. IEEE, 1999, pp. 116–122. 2, 3

[14] J. Deng and H. T. Tsui, "A Two-step Approach based on PaHMM for the Recognition of ASL," in *The Fifth Asian Conference On Computer Vision*, Melbourne, Australia, 2002, pp. 126–131. 2, 3

[15] C. Vogler and D. Metaxas, "Handshapes and movements: Multiple-channel ASL recognition," in *Lecture Notes in Computer Science*. Springer, 2004, pp. 247–258. 2

[16] J. Forster, O. Koller, C. Oberdörfer, Y. Gweth, and H. Ney, "Improving Continuous Sign Language Recognition: Speech Recognition Techniques and System Design," in *Workshop on Speech and Language Processing for Assistive Technologies*, Grenoble, France, Aug. 2013, pp. 41–46, satellite Workshop of INTERSPEECH 2013. 2

[17] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A. Senior, "Recent advances in the automatic recognition of audiovisual speech," *Proceedings of the IEEE*, vol. 91, no. 9, pp. 1306–1326, Sep. 2003. 2

[18] A. Varga and R. K. Moore, "Hidden Markov model decomposition of speech and noise," in *Acoustics, Speech, and Signal Processing, 1990. ICASSP-90., 1990 International Conference On*. IEEE, 1990, pp. 845–848. 2

[19] M. J. Tomlinson, M. J. Russell, and N. M. Brooke, "Integrating audio and visual information to provide highly robust speech recognition," in *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, vol. 2, May 1996, pp. 821–824 vol. 2. 2, 3

[20] Z. Ghahramani and M. I. Jordan, "Factorial hidden Markov models," in *Advances in Neural Information Processing Systems*, 1996, pp. 472–478. 2

[21] M. Brand, N. Oliver, and A. Pentland, "Coupled hidden Markov models for complex action recognition," in *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Jun. 1997, pp. 994–999. 2

[22] S. Bengio, "An asynchronous hidden markov model for audio-visual speech recognition," in *Advances in Neural Information Processing Systems*, 2003, pp. 1213–1220. 2

[23] J. Forster, C. Oberdörfer, O. Koller, and H. Ney, "Modality Combination Techniques for Continuous Sign Language Recognition," in *Iberian Conference on Pattern Recognition and Image Analysis*, ser. Lecture Notes in Computer Science 7887. Madeira, Portugal: Springer, Jun. 2013, pp. 89–99. 2, 3

[24] U. von Agris, J. Zieren, U. Canzler, B. Bauer, and K. F. Kraiss, "Recent developments in visual sign language recognition," *Universal Access in the Information Society*, vol. 6, no. 4, pp. 323–362, 2008. 2, 3

[25] S. Theodorakis, A. Katsamanis, and P. Maragos, "Product-HMMs for automatic sign language recognition," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference On*. IEEE, 2009, pp. 1601–1604. 3

[26] J. Ma, W. Gao, and R. Wang, "A Parallel Multistream Model for Integration of Sign Language Recognition and Lip Motion," in *Advances in Multimodal Interfaces — ICMI 2000*, ser. Lecture Notes in Computer Science, T. Tan, Y. Shi, and W. Gao, Eds. Springer Berlin Heidelberg, 2000, no. 1948, pp. 582–589. 3

[27] N. Nishida and H. Nakayama, "Multimodal Gesture Recognition Using Multi-stream Recurrent Neural Network," in *Image and Video Technology*, ser. Lecture Notes in Computer Science, T. Bräunl, B. McCane, M. Rivera, and X. Yu, Eds. Springer International Publishing, Nov. 2015, no. 9431, pp. 682–694. 3

[28] C. Neti, G. Potamianos, J. Luettin, I. Matthews, H. Glotin, D. Vergyri, J. Sison, A. Mashari, and J. Zhou, "Audio-visual speech recognition," in *Final Workshop 2000 Report*, vol. 764, 2000. 3

[29] I. Sutskever, O. Vinyals, and Q. Le, "Sequence to Sequence Learning with Neural Networks," in *Advances in Neural Information Processing Systems*, 2014, pp. 3104–3112. 3

[30] S. Petridis, Y. Wang, Z. Li, and M. Pantic, "End-to-End Multi-View Lipreading," in *Proceedings of the British Machine Vision Conference (BMVC)*. London, UK: BMVA Press, Sep. 2017. 3

[31] J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Lip Reading Sentences in the Wild," in *IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, HI, USA, Jul. 2017. 3

[32] D. Bahdanau, J. Chorowski, D. Serdyuk, P. Brakel, and Y. Bengio, "End-to-End Attention-based Large Vocabulary Speech Recognition," *arXiv preprint arXiv:1508.04395*, 2015. 3

[33] H. Bourlard and S. Dupont, "A new ASR approach based on independent processing and recombination of partial frequency bands," in *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference On*, vol. 1. IEEE, 1996, pp. 426–429, 00350. 3, 4

[34] C. Wellekens, J. Kangasharju, and C. Milesi, "The use of meta-HMM in multistream HMM training for automatic speech recognition." in *ICSLP*, 1998. 3

[35] S. Dupont and H. Bourlard, "Using multiple time scales in a multi-stream speech recognition system," in *EUROSPEECH" 97*, 1997, pp. 3–6. 3

[36] X. Zhu, "Semi-Supervised Learning Literature Survey," Computer Sciences, University of Wisconsin -Madison, Tech. Rep. 1530, 2008. 3

[37] P. Buehler, A. Zisserman, and M. Everingham, "Learning sign language by watching TV (using weakly aligned subtitles)," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference On*. IEEE, 2009, pp. 2961–2968. 3

[38] D. Kelly, J. McDonald, and C. Markham, "Weakly Supervised Training of a Sign Language Recognition System Using Multiple Instance Learning Density Matrices," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 41, no. 2, pp. 526–541, Apr. 2011. 3

[39] A. Farhadi and D. Forsyth, "Aligning ASL for statistical translation using a discriminative word model," in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference On*, vol. 2. IEEE, 2006, pp. 1471–1476. 3

[40] H. Cooper and R. Bowden, "Learning signs from subtitles: A weakly supervised approach to sign language recognition," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, Miami, FL, Jun. 2009, pp. 2568–2574. 3

[41] T. Quack, V. Ferrari, B. Leibe, and L. V. Gool, "Efficient mining of frequent and distinctive feature configurations," in *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference On*. IEEE, 2007, pp. 1–8. 3

[42] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 39, no. 1, pp. 1–38, 1977. 3, 4, 6

[43] C. Wang, D. Blei, and F.-F. Li, "Simultaneous image classification and annotation," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference On*. IEEE, 2009, pp. 1903–1910. 3

[44] Y. Wu, T. Huang, and K. Toyama, "Self-supervised learning for object recognition based on kernel discriminant-EM algorithm," in *Eighth IEEE International Conference on Computer Vision, 2001. ICCV 2001. Proceedings*, vol. 1, 2001, pp. 275–280 vol.1. 3

[45] A. Senior, G. Heigold, M. Bacchiani, and H. Liao, "GMM-free DNN training," in *Proceedings of ICASSP*, 2014, pp. 5639–5643. 3

[46] Y. Yang, I. Saleemi, and M. Shah, "Discovering Motion Primitives for Unsupervised Grouping and One-Shot Learning of Human Actions, Gestures, and Expressions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 7, pp. 1635–1648, Jul. 2013. 3

[47] S. Nayak, S. Sarkar, and B. Loeding, "Automated extraction of signs from continuous sign language sentences using iterated conditional modes," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference On*. IEEE, 2009, pp. 2583–2590. 3

[48] H. A. Bourlard and N. Morgan, *Connectionist Speech Recognition: A Hybrid Approach*. Norwell, MA: Kluwer Academic Publishers, 1993. 4, 6

[49] J. Forster, C. Schmidt, T. Hoyoux, O. Koller, U. Zelle, J. Piater, and H. Ney, "RWTH-PHOENIX-Weather: A Large Vocabulary Sign Language Recognition and Translation Corpus," in *International Conference on Language Resources and Evaluation*, Istanbul, Turkey, May 2012, pp. 3785–3789. 6

[50] J. Forster, C. Schmidt, O. Koller, M. Bellgardt, and H. Ney, "Extensions of the Sign Language Recognition and Translation Corpus RWTH-PHOENIX-Weather," in *International Conference on Language Resources and Evaluation*, Reykjavik, Island, May 2014, pp. 1911–1916. 6

[51] O. Koller, J. Forster, and H. Ney, "Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers," *Computer Vision and Image Understanding*, vol. 141, pp. 108–125, Dec. 2015. 6, 7, 10, 11

[52] C. Camgoz, S. Hadfield, O. Koller, H. Ney, and R. Bowden, "Neural Sign Language Translation," in *IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, Jun. 2018. 6, 7, 8

[53] F. J. Och and H. Ney, "A Systematic Comparison of Various Statistical Alignment Models," *Comput. Linguist.*, vol. 29, no. 1, pp. 19–51, Mar. 2003. 7

[54] M. Bisani and H. Ney, "Joint-sequence models for grapheme-to-phoneme conversion," *Speech Communication*, vol. 50, no. 5, pp. 434–451, May 2008. 7

[55] V. Sutton and D. A. C. f. S. Writing, *Sign Writing*. Deaf Action Committee (DAC), 2000. 8
[56] O. Koller, H. Ney, and R. Bowden, "May the Force be with you: Force-Aligned SignWriting for Automatic Subunit Annotation of Corpora," in *IEEE International Conference on Automatic Face and Gesture Recognition*, Shanghai, PRC, Apr. 2013, pp. 1–6. 8
[57] Y. Bouzid, M. Jbali, O. El Ghoul, and M. Jemni, "Towards a 3d signing avatar from signwriting notation," in *Proceedings of the 13th International Conference on Computers Helping People with Special Needs - Volume Part II*, ser. ICCHP'12. Berlin, Heidelberg: Springer-Verlag, 2012, pp. 229–236. 8
[58] D. McKee, R. McKee, S. P. Alexander, and L. Pivac, "The Online Dictionary of New Zealand Sign Language," http://nzsl.vuw.ac.nz/, 2015. 8, 9
[59] D. Rybach, S. Hahn, P. Lehnen, D. Nolden, M. Sundermeyer, Z. Tüske, S. Wiesler, R. Schlüter, and H. Ney, "RASR - The RWTH Aachen University Open Source Speech Recognition Toolkit," in *IEEE Automatic Speech Recognition and Understanding Workshop*, Waikoloa, HI, USA, Dec. 2011. 9
[60] H. Cooper, N. Pugeault, and R. Bowden, "Reading the signs: A video based sign dictionary," in *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference On*. IEEE, Nov. 2011, pp. 914–919. 9
[61] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional LSTM and other neural network architectures," *Neural Networks*, vol. 18, no. 5, pp. 602–610, 2005. 9
[62] M. Liwicki, A. Graves, H. Bunke, and J. Schmidhuber, "A novel approach to on-line handwriting recognition based on bidirectional long short-term memory networks," in *In Proceedings of the 9th International Conference on Document Analysis and Recognition, ICDAR 2007*, 2007. 9
[63] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," in *ICLR*, May 2015. 9
[64] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in *Advances in Neural Information Processing Systems*, 2012, pp. 1106–1114. 9
[65] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going Deeper With Convolutions," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, Ma, USA, Jun. 2015, pp. 1–9. 9
[66] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, Dec. 2015. 9
[67] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE transactions on neural networks*, vol. 5, no. 2, pp. 157–166, 1994. 10
[68] R. J. Williams and D. Zipser, "Gradient-based learning algorithms for recurrent networks and their computational complexity," *Backpropagation: Theory, architectures and applications*, vol. 1, pp. 433–486, 1995. 10
[69] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum Learning," in *Proceedings of the 26th Annual International Conference on Machine Learning*, ser. ICML '09. New York, NY, USA: ACM, 2009, pp. 41–48. 10
[70] L. Gillick and S. Cox, "Some Statistical Issues in the Comparison of Speech Recognition Algorithms," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Glasgow, Scotland, 1989, pp. 532–535. 11
[71] D. S. Pallet, W. M. Fisher, and J. G. Fiscus, "Tools for the analysis of benchmark speech recognition tests," in *Acoustics, Speech, and Signal Processing, 1990. ICASSP-90., 1990 International Conference On*, 3, pp. 97 –100 vol.1. 11
[72] C. Schmidt, O. Koller, H. Ney, T. Hoyoux, and J. Piater, "Enhancing Gloss-Based Corpora with Facial Features Using Active Appearance Models," in *International Symposium on Sign Language Translation and Avatar Technology*, vol. 2, Chicago, IL, USA, Oct. 2013. 11
[73] R. Cui, H. Liu, and C. Zhang, "Recurrent Convolutional Neural Networks for Continuous Sign Language Recognition by Staged Optimization," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, Jul. 2017. 11
[74] N. C. Camgoz, S. Hadfield, O. Koller, and R. Bowden, "SubUNets: End-to-end Hand Shape and Continuous Sign Language Recognition," in *IEEE International Conference on Computer Vision*, Oct. 2017, pp. 22–27. 11
[75] O. Koller, S. Zargaran, H. Ney, and R. Bowden, "Deep Sign: Hybrid CNN-HMM for Continuous Sign Language Recognition," in *British Machine Vision Conference*, York, UK, Sep. 2016. 11
[76] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations From Deep Networks via Gradient-Based Localization," in *IEEE International Conference on Computer Vision*, 2017, pp. 618–626. 12, 13

Oscar Koller was a doctoral student researcher in the Human Language Technology & Pattern Recognition Group led by Prof. Ney at RWTH Aachen University, Germany. He joined the group in 2011 & followed a dual supervision by Prof. Bowden & his Cognitive Vision group at University of Surrey, UK, where he spent 12 months as a visiting researcher. He is now with Microsoft AI and Research. His main research interests include sign language & gesture recognition, lip reading & speech recognition.

Necati Cihan Camgoz received his B.Sc. and M.Sc. degrees in Computer Engineering from Yildiz Technical University, Istanbul, Turkey and Bogazici University, Istanbul, Turkey, respectively. Currently, he is pursuing a PhD degree at the Centre for Vision, Speech and Signal Processing in the University of Surrey, United Kingdom. His research interests include human-computer interaction, sign language & gesture recognition, and machine translation.

Hermann Ney is a full professor of computer science at RWTH Aachen University, Germany. Previously, he headed the Speech Recognition Group at Philips Research. His main research interests include the area of statistical methods for pattern recognition and human language technology and their specific applications to speech recognition, machine translation, and image object recognition. In particular, he has worked on dynamic programming for continuous speech recognition, language modeling, and phrase-based approaches to machine translation. He has authored and coauthored more than 600 papers in journals, books, conferences, and workshops. In 2006, he was the recipient of the Technical Achievement Award of the IEEE Signal Processing Society. He was a member of the Speech Technical Committee of the IEEE Signal Processing Society from 1997 to 2000. He is a fellow of the IEEE.

Richard Bowden received a BSc and MSc from the Universities of London and Leeds and a PhD from Brunel University for which was awarded the Sullivan Doctoral Thesis Prize. He is Professor of computer vision and machine learning at the University of Surrey leading the Cognitive Vision Group within the Centre for Vision, Speech and Signal Processing and was awarded a Royal Society Leverhulme Trust Senior Research Fellowship in 2013. His research centres on the use of computer vision to locate, track, and understand humans. He is an associate editor for the journals Image and Vision computing and IEEE TPAMI. He was a member of the British Machine Vision Association (BMVA) executive committee and a company director for seven years. He is a member of the BMVA, a fellow of the Higher Education Academy, a senior member of the IEEE and Fellow of the International Association of Pattern Recognition (IAPR).