This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the version available on IEEE Xplore.

Deep Hand: How to Train a CNN on 1 Million Hand Images When Your Data Is Continuous and Weakly Labelled

Oscar Koller, Hermann Ney
Human Language Technology & Pattern Recog.
RWTH Aachen University, Germany
{koller,ney}@cs.rwth-aachen.de

Richard Bowden
Centre for Vision Speech & Signal Processing
University of Surrey, UK
r.bowden@surrey.ac.uk

Abstract

This work presents a new approach to learning a frame-based classifier on weakly labelled sequence data by embedding a CNN within an iterative EM algorithm. This allows the CNN to be trained on a vast number of example images when only loose sequence level information is available for the source videos. Although we demonstrate this in the context of hand shape recognition, the approach has wider application to any video recognition task where frame level labelling is not available. The iterative EM algorithm leverages the discriminative ability of the CNN to iteratively refine the frame level annotation and subsequent training of the CNN. By embedding the classifier within an EM framework the CNN can easily be trained on 1 million hand images. We demonstrate that the final classifier generalises over both individuals and data sets. The algorithm is evaluated on over 3000 manually labelled hand shape images of 60 different classes which will be released to the community. Furthermore, we demonstrate its use in continuous sign language recognition on two publicly available large sign language data sets, where it outperforms the current state-of-the-art by a large margin. To our knowledge no previous work has explored expectation maximization without Gaussian mixture models to exploit weak sequence labels for sign language recognition.

1. Introduction

Convolutional Neural Networks (CNNs) have been demonstrated to provide superior performance in many tasks. But to achieve this they require large amounts of labelled training data which in many areas is a limiting factor. Pose-independent hand shape recognition, crucial to gesture and sign language recognition, suffers from large visual intra-class ambiguity and therefore places further burden on the acquisition of training data. Typically, only small and quite specific labelled data sets exist ([16, 26]) which usually do not provide sufficiently fine-grained hand shape classes suitable for sign language recognition. Recent advances in sign language research have given rise to many publicly available sign language lexicons that allow searching of the videos by the index of hand shapes. These resources constitute noisy but valuable data sources. In this work, we exploit the modelling capabilities of a pre-trained 22 layer deep convolutional neural network and integrate it into a force-aligning algorithm that converts noisy video level annotations into a strong frame level classifier. As such, this manuscript provides the following contributions:

- formulation of an EM-based algorithm integrating CNNs with Hidden-Markov-Models (HMMs) for weak supervision and overcoming the temporal alignment problem in continuous video processing tasks using the strong discriminative capabilities of CNN architectures
- robust fine grained single frame hand shape recognition based on a CNN-model, trained on over 1 million hand shapes and shown to generalise across data sets without retraining
- making an articulated sign language hand shape data set publicly available comprising 3361 manual labelled frames in 45 classes [1]
- and integration of pose-independent hand shape subunits into a continuous sign language recognition pipeline

This paper is organised as follows: after introducing the related literature in Section 2 we give a precise problem formulation and the solution in Section 3. Section 4 introduces the employed data sources. Subsequently, we evaluate the approach in Section 5 in two parts: firstly classifying single frames and secondly in a continuous sign language recognition pipeline. The paper closes with a conclusion in Section 6

[1] Available at: http://www.hltpr.rwth-aachen.de/~koller/1miohands

2. State-of-the-Art

This work deals with the problem of weakly supervised learning from sequence labels applied to the problem of hand shape recognition. We therefore look at the state-of-the-art in both areas related to the domains of gesture and sign language.

Hand shape recognition from a single image may be understood as the hand pose configuration specified by joint positions and angles which to date are mostly estimated based on depth images and pixel-wise hand segmentation [33, 21]. However, in the scope of this work, hand shape recognition is seen as a classification task of a specific number of defined hand shapes. Known approaches fall into three categories: (i) template matching against a large data set of often synthetic gallery images [25] or contour shapes [1, 3]; (ii) generative model fitting approaches [35, 10, 28]; and (iii) discriminative modelling approaches such as Cooper *et al*. [6]. Cooper uses random forests trained on HOG features to distinguish 12 hand shapes, each trained on 1000 training samples. However, they restricted the classifier to work on hands not in motion and applied it only to isolated sign language recognition. There seems to be no previous work exploiting CNNs for hand shape classification other than [40] which only distinguishes 6 classes trained with 7500 images per class. A few recent publications apply CNNs to finger and joint regression based on depth data [38, 24]. Tompson *et al*. [34] present a CNN-based hand pose estimation based on depth data. They generate computationally heavy heat maps for 2D joint locations and infer the 3D hand pose by the depth channel and inverse kinematics.

There are many approaches to **learning from ambiguous labels** or **weakly supervised learning** (see [42] 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*. [4] and similarly Kelly *et al*. [17] apply MIL to learning sign categories from TV subtitles, circumventing the translation problem by performing sign spotting. However, Farhadi and Forsyth [9] were the first to approach the subtitle-sign-alignment problem. They used a HMM to find sign boundaries. Cooper and Bowden [5] solved the same problem by applying efficient data mining methods, an idea that was introduced to the vision community by Quack *et al*. [27]. Another approach uses Expectation Maximisation (EM) [7] to fit a model to data observations. Koller *et al*. [20] used EM to fit a Gaussian Mixture Model (GMM) to Active Appearance Model (AAM) mouth features in order to find and model mouth shape sequences in sign language. Other works use EM to link text and image regions [37]. Wu *et al*. [39] introduced a non-linear kernel discriminant analysis step in between the expectation and maximisation step to map the features to a lower dimensional space which could help 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 [30]. Closely related is also the clustering of spatio-temporal motion patterns for action recognition [41] and Nayak's work on iterated conditional modes [23] to extract signs from continuous sentences. Learning frame labels from video annotations is an underexploited approach in the vision community and the previous literature has several shortcomings that we address with this work:

1. The discriminative capabilities of CNNs have not yet been integrated into a weakly supervised learning scheme able to exploit large ambiguously labelled data sets.

2. No previous work has explicitly worked on posture and pose-independent hand shape classification, which is crucial in real-life sign language footage, as hand shape and posture have been determined as independent information sources by sign linguists.

3. To our knowledge no previous work has exploited the classification power of CNNs with application to sign language hand shape classification.

4. No previous work has trained a classifier on over a million hand shapes of real sign language data.

5. No previous work has dealt with data set independent hand shape classification.

However, there is much to be gained from addressing these shortcomings. If CNNs can be trained using weak video annotation, then we can leverage the power of CNNs to generalise over large data sets.

3. Weakly Supervised CNN Training

The proposed algorithm constitutes a successful solution to the problem of weakly supervised learning from noisy sequence labels to correct frame labels. An overview of the approach is given in Figure 1, which shows the overall pipeline specific to the task of hand shape classification. However, the algorithm could be easily applied to other tasks. The input images are cropped around the tracked hands, which forms the input to our weakly supervised CNN training. The iterative learning algorithm needs an initialisation, which is referred to as 'flat start'. This involves linearly partitioning the input frames to an available initial annotation, usually a single hand shape class preceded and followed by instances of the garbage class (as the hand shape is expected to happen in the middle of the sequence). The algorithm iteratively refines the temporal class

boundaries and trains a CNN that performs single image hand shape recognition. While refining the boundaries, it may drop the label sequence or exchange it for one that better fits the data. The iterative process is similar to a forced alignment procedure, however, rather than using Gaussian mixtures as the probabilistic component we use the outputs of the CNN directly.

3.1. Problem Formulation

Given a sequence of images $x_1^T = x_1, \ldots, x_T$ and an ambiguous class label $\tilde{l}$ for the whole sequence, we want to jointly find the true label l for each frame and train a model such that the class symbol posterior probability $p(k|x)$ over all images and classes is maximised. We assume that a lexicon ψ of possible mappings from $\tilde{l} \rightarrow l$ exists, where l can be interpreted as a sequence of up to L class symbols k,

$$\psi = \left\{ \tilde{l} : l_1^L \mid l \in \{k_1, \ldots, k_N, \varnothing\} \right\} \tag{1}$$

Optionally, l may be an empty symbol corresponding to a garbage class. Each $\tilde{l}$ can map to multiple symbol sequences (which is important as $\tilde{l}$ is ambiguous and a one-to-one mapping would not be sufficient). In terms of sequence constraints, we only require each symbol to span an arbitrary length of subsequent images as we assume that symbols (in our application: hand shapes) are somewhat stationary and do not instantly disappear or appear.

Due to the promising discriminatory capabilities of CNNs, we solve the problem in an iterative fashion with the EM algorithm [7] in a HMM setting and use the CNN for modelling $p(k|x)$.

3.2. Sequential Time-Decoding

The basic idea of EM is to start with a random model initialisation and then iteratively (i) update the assignment of class labels to images (E-Step) and then (i) re-estimate the model parameters to adapt to the change (M-Step).

The E-Step consists of the forward-backward algorithm, which identifies the sequence of class symbols aligned to the images that best fits the learnt model. Using Bayes' decision rule, we maximise the posterior probability over all possible true labels l, corresponding to casting the class symbol model $Pr(x_t|k_t)$ given by the CNN as the marginal over all possible HMM temporal state sequences $s_1^T = s_1, \ldots, s_T$ defined by the symbol sequences in ψ. For an efficient implementation, following [11], we assume a first order Markov dependency and maximum approximation:

$$x_1^T \rightarrow [k_1^T]_{\text{opt}} = \operatorname*{argmax}_{k_1^N} \left\{ Pr(l) \max_{s_1^T} \left\{ Pr(x_t|k_1^N) \cdot Pr(s_t|s_{t-1}) \right\} \right\} \tag{2}$$

where $Pr(l)$ denotes the symbol sequence prior probability and $Pr(x_t|k_1^N)$ is modelled by the CNN. To add robustness, we employ a pooled *state transition* model $Pr(s_t|s_{t-1})$ with globally set transition probabilities. Those form a HMM in bakis structure (left-to-right structure; forward, loops and skips across at most one state are allowed, where two subsequent states share the same class probabilities). The garbage class is modelled as an ergodic state with separate transition probabilities to add flexibility, such that it can always be inserted between sequences of symbols.

Usually, this approach is used jointly with GMMs, which model directly $p(x|k)$ as generative models. However, the CNN models the posterior probability $p(k|x)$. Inspired by the hybrid approach [2] known from ASR we convert the CNN's posterior output to likelihoods given the class counts in our data ($p(k)$) using the Bayes' rule as follows:

$$p(x_t|k) \propto p(k|x_t)/p(k)^{\alpha} \tag{3}$$

This allows us to add symbol sequence prior knowledge from the lexicon ψ. Equation 2 then becomes:

$$\operatorname*{argmax}_{k_1^N} \left\{ p(l) \max_{s_1^T} \left\{ \frac{p(k_t|x_t)}{p(k)^{\alpha}} \cdot p(s_t|s_{t-1}) \right\} \right\}, \tag{4}$$

where the scaling factor α is a hyperparameter allowing us to control the impact of the class prior.

3.3. Convolutional Neural Network Architecture

Knowing the weakly supervised characteristics of our problem, we would like to incorporate as much prior knowledge as possible to guide the search for the true symbol class labels. Pre-trained CNN models constitute such a source of knowledge, which seems reasonable as the pre-trained convolutional filters in the lower layers may capture simple edges and corners, applicable to a wide range of image recognition tasks. We opt for a model previously trained in a supervised fashion for the ImageNet Large-Scale Visual Recognition Challenge (ILSVRC) 2014 . We choose a 22 layer deep network architecture following [32] which achieves a top-1 accuracy of 68.7% and a top-5 accuracy 88.9% in the ILSVRC. The network involves an inception architecture, which helps to reduce the numbers of free parameters while allowing for a very deep structure. Our model has about 6 million free parameters. All convolutional layers and the last fully connected layer use rectified linear units as non-linearity. Additionally, a dropout layer with 70% ratio of dropouts is used to prevent over-fitting. We base our CNN implementation on [15], which is an efficient C++ implementation using the NVIDIA CUDA Deep Neural Network GPU-accelerated library.

We replace the last pre-trained fully connected layers before the output layers with those matching the number of

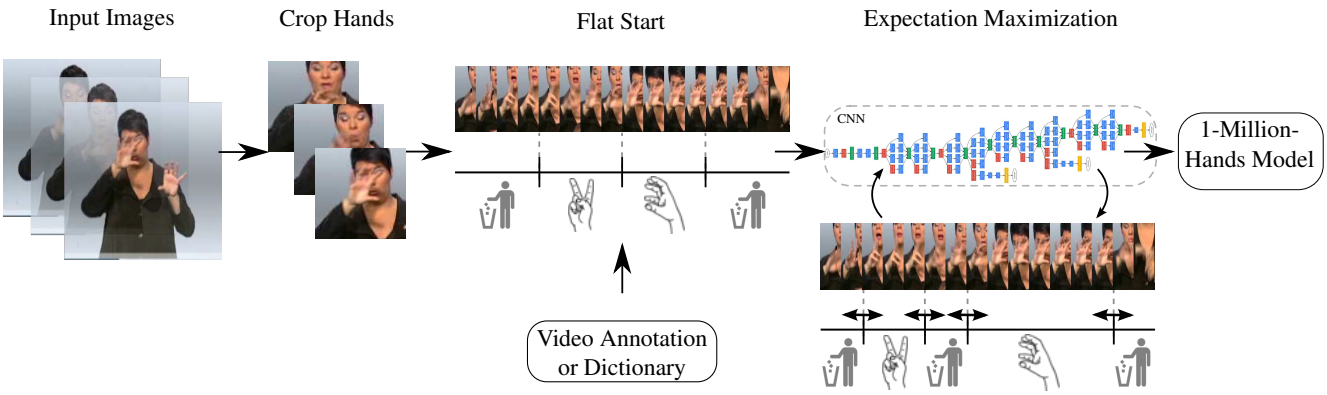

Figure 1. Overview of presented Algorithm.

classes in our problem (plus one garbage class), which we initialise with zeros.

As a preprocessing step, we apply a per pixel mean normalisation to the images prior to fine-tuning the CNN model with Stochastic Gradient Descent (SGD) and a softmax based cross-entropy classification loss E

$$E = -\frac{1}{N}\sum_{n=1}^{N} \log\ p(k|x_n). \tag{5}$$

4. Data Sets

We employ three different data sets for training the hand shape classifier. All data sets feature sign language footage. Two represent video based publicly available sign language lexicons with isolated signs from Danish sign language [14] and New Zealand sign language [22]. The third source represents the training set of RWTH-PHOENIX-Weather 2014 [12], a publicly available continuous sign language data set. Figure 2 shows sample sequences from all three data sets, where it can be seen that the lexicons have single sign data, whereas PHOENIX provides full signed sentences. The Danish data contains hardly any motion blur, whereas there is some motion blur present in the New Zealand data and a large portion of the PHOENIX video frames contain heavy motion blur. The sign language lexica provide linguistic hand shape labels for each of the sign videos that enable a search by hand shape on the lexicon web sites. As for the danish data, we obtained a consolidated version of hand shape annotations directly from the maintainer of the lexicon. However, from a pattern recognition point of view these annotations are extremely ambiguous and noisy. They consist of a single hand shape, sometimes a sequence of two hand shapes, for a whole signed video. As can be seen in Figure 2, the hand shape can be more or less static throughout the video (top example in Figure 2), or it reflects only one temporary portion of a changing hand configuration (middle example in Figure 2). In any

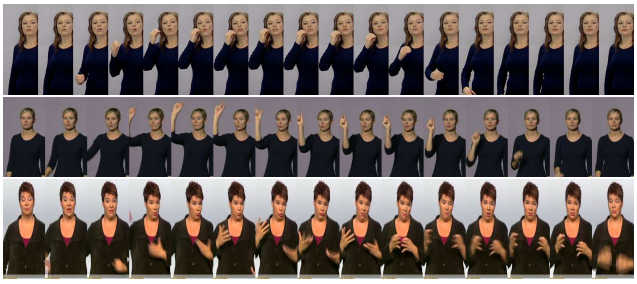

Figure 2. Showing employed data sets for training: Top to bottom, Danish sign language dictionary, New Zealand sign language dictionary and a sentence from RWTH-PHOENIX-Weather corpus.

case, the signer brings his hands from a neutral position to the place of sign execution, while transitioning from a neutral hand shape to the target hand shape composing the sign and to possible subsequent hand shapes. While the sign is performed, it may involve a hand movement, a rotation of the hand and changes in hand shape. The annotation may represent any of these hand shapes or an intermediate configuration that was considered linguistically dominant during the annotation. As there are no hand shape annotations for the RWTH-PHOENIX-Weather data available, we employ a publicly available sign language lexicon called SignWriting [31]. 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 (which contains hand shape information). The German SignWriting lexicon currently comprises 24.293 entries. Inspired by [19], we parsed all entries to create the mapping ψ from sign annotations to possible hand shape sequences, where we remove all hand pose related information (such as rotations) of the hand annotations. This mapping will be made available, in order to make our results reproducible. Throughout this work we follow the hand shape taxonomy by the danish sign language lexicon team, which amounts to over 60 different hand shapes, often with very subtle differences such as a flexed versus straight thumb.

	danish	nz	ph
duration [min]	97	192	532
# frames	145,720	288,593	799,006
# hand shape frames (autom.)	65,088	153,298	786,750
# garbage frames (autom.)	80,632	135,295	12,256
# signed sequences	2,149	4,155	5,672
# signs	2,149	4,155	65,227
# signers	6	8	9

Table 1. Corpus statistics: Danish ('danish'), New Zealand ('nz') and RWTH-PHOENIX-Weather ('ph') sign language data sets used for training the hand shape classifier.

Figure 3. 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 [22].

Statistics of all three data sets are given in Table 1. Garbage and hand shape frames are estimated automatically by our algorithm. All three data sets total to over one million hand shape images produced by 23 individuals.

Some resources have been manually created in the scope of this work. Among them a mapping from the New Zealand and the SignWriting hand shape taxonomy to the employed Danish taxonomy. Some hand shape classes were ambiguous between the two annotation schemes, yielding a one-to-many mapping that could be integrated into ψ, which will also be made available. For evaluating the 1-Million-Hands CNN classifier, we manually labelled 3361 images from the RWTH-PHOENIX-Weather 2014 Development set [2]. Some of the 45 encountered pose-independent hand shape classes are depicted in Figure 3. 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 4 showing that the top 14 hand shapes explain 90% of the annotated samples.

Finally, we evaluate on two publicly available continuous sign language data set benchmarks: (i) RWTH-PHOENIX-Weather 2014 Multisigner corpus [12] , which is a challenging real-life continuous sign language corpus that can be considered to be one of the largest published

[2] Available at: http://www.hltpr.rwth-aachen.de/~koller/1miohands

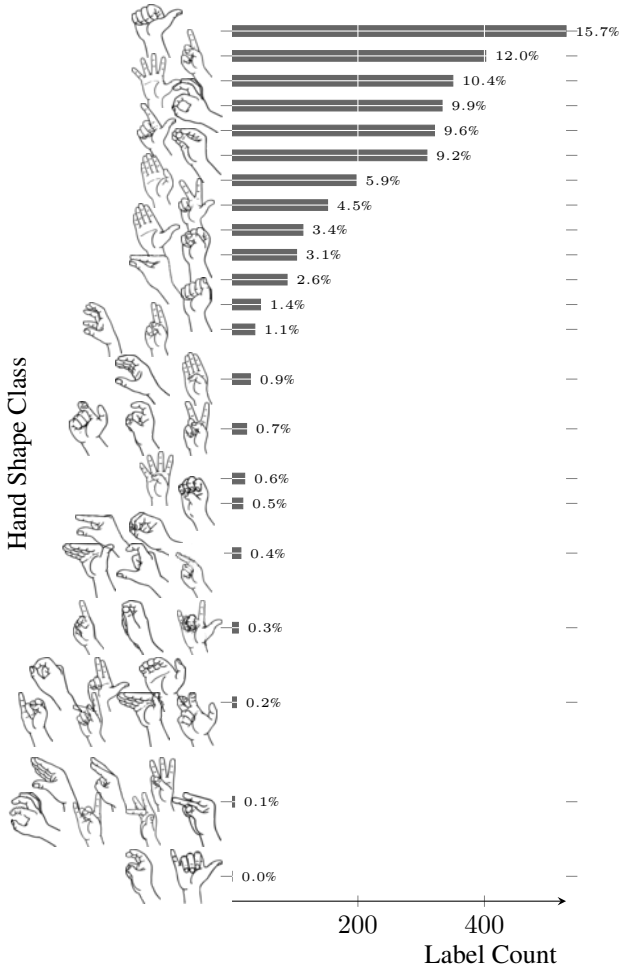

Figure 4. Ground truth hand shape label count of all 3361 annotations. 45 out of 60 classes have been found in the data and could be labelled. If several hand shapes appear close to one label counting bar, each hand shape alone amounts to the mentioned fraction of labels. Hand-Icons from [22].

continuous sign language corpora. It covers unconstrained sign language of 9 different signers with a vocabulary of 1081 different signs. (ii) SIGNUM [36] signer-dependent subset, which has been well established as a benchmark for a significant amount of sign language research. Both data sets are presented in detail in [18].

5. Experiments

In this section we describe the experimental validation of the proposed algorithm with application to learning a robust pose-independent hand shape classifier based on a CNN. In the first two subsections we describe the training parameters and discuss evaluation on the frame-level. Subsection 5.3 applies the learnt 1-Million-Hands model to the challenging

problem of continuous sign language recognition, where it outperforms current state-of-the-art by a large margin.

5.1. Hand Shape Model Training

Data preparation. The data is downloaded and prepared by tracking the hands using a model-free dynamic programming tracker [8]. Being based on dynamic programming, the tracker optimises the tracking decisions over time and traces back the best sequence of tracking decisions at the end of the video. The size of the hand patch is roughly chosen so that it is two to three times the size of a hand. However, the appearance of the hand changes as the signer moves it towards the camera.

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. As described in Section 4, the annotation taxonomy of the New Zealand data does not match the employed Danish taxonomy one to one. This partly 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.

Initialise algorithm. The input videos are linearly partitioned based on a random hand shape label sequence from the lexicon ψ, considering the beginning and end of each video 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 [29]. All 60 hand shape classes are represented by a double state, whereas the garbage class just has a single state. We use fixed, non-optimised transition penalties being '2-0-2' for 'loop-forward-skip' for all hand shape classes and '0-2' for the garbage 'loop-forward'. The scaling factor α is set to 0.3 in our experiments. As already pointed out by [6], we also observe a strong bias in the distribution of hand shape classes in our data, but we decided to maintain it. To speed up CNN training time we randomly sample from the observation sequences of the garbage class. In this way we decrease the amount of garbage frames and match it to the most frequently observed hand shape class.

CNN training. We replace the pre-trained output layers with a 61 dimensional fully connected layer, accounting for 60 hand shape classes and a garbage class. We have empirically noticed that training all layers with an equal learning rate outperforms training just the output layer or weighting the output layer's learning rate. For all experiments we use a fixed learning rate $\text{lr} = 0.0005$ for 3 epochs and finish a last epoch with $\text{lr} = 0.00025$. We select the best training based on the manually annotated evaluation data presented in Section 4, but, as shown in the evaluation, the automatic development data behaves comparably (see Fig. 5).

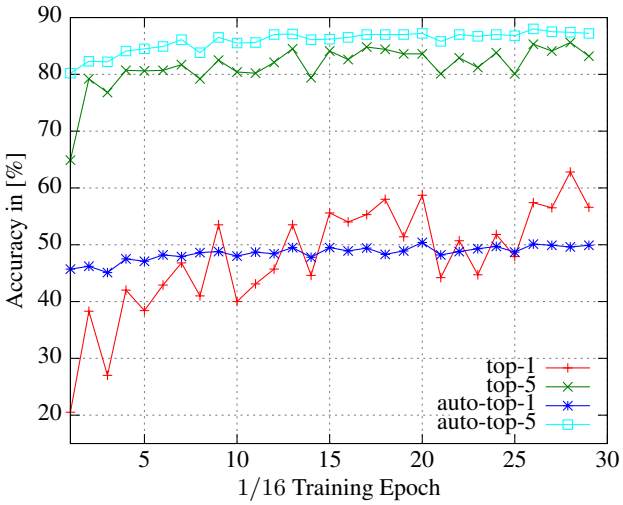

Figure 5. Showing the top-1 and top-5 CNN accuracies for every 16^{th} training epoch measured on the manual annotations ('top-1' 'top-5') and on a development split of the automatically labelled training data ('auto-top-1' 'auto-top-5'). Given is the last iteration of the EM-algorithm yielding a 62.8% top-1 accuracy.

In Figure 5 we show the evolving accuracy during one epoch of CNN training measured 16 times per iteration. Given is both the accuracy on the manually annotated hand shape set, as well the accuracy on a randomly split development set representing the automatic alignment generated by the HMM. It is good to see that both measures converge in a similar fashion, which indicates that using the automatic data for training may be sufficient. To obtain a strong classifier it is good to start with data providing stronger supervision while subsequently adding the remainder.

5.2. Frame-Level Evaluation

In terms of **run time**, the CNN requires 8.24ms in the forward-pass to classify a single image (when supplied in batches of 32 images) on a single GeForce GTX 980 GPU with 4095 MiB. The algorithm can therefore run with over 100fps in a recognition system.

In Table 2 we display the training accuracy of the CNN measured on the manually annotated PHOENIX images across five iterations of the proposed EM algorithm. Three different setups are presented, showing the effect of increased training data. We deploy a system using solely the Danish data, one using the Danish and the New Zealand data and one using all three resources. Note, in the first two cases the CNN successfully classifies handshapes of an unseen data set and is thus independent of the data set (no samples of the evaluation corpus are used for training), as we are measuring the evaluation on the RWTH-PHOENIX-Weather hand shape annotations. We see that the training accuracy increases with each iteration in the first two cases

Iter.	Danish	+nz	+ph	Danish	+nz	+ph
	top-1			top-5		
1	40.3	51.1	51.8	73.0	79.4	79.4
2	47.8	52.1	56.3	77.9	81.6	81.2
3	44.1	54.0	**62.8**	68.3	80.7	**85.6**
4	48.4	59.5	57.7	74.9	84.7	84.2
5	**50.6**	**59.6**	55.3	**76.3**	**86.4**	84.1

Table 2. CNN training accuracies in [%] per EM iteration. 'Danish' stands for the Danish Sign Language Dictionary, where as 'nz' is the New Zealand Sign Language dictionary and 'ph' is the RWTH-PHOENIX-Weather 2014 train set. '+' denotes the aggregation of the current and the data sets to the left.

and then slowly converges. Due to the lower amount of hand shape samples in the Danish case, a single training iteration has less impact on the CNN's weights which results in slower convergence (measured per epoch). We further note that adding PHOENIX data to the train set does not seem to converge to a stable maximum (at least not after a few iterations), but improves to 62.8% top-1 accuracy and then decreases again. This is likely to be due to the fact that the PHOENIX data set covers continuously signed sentences that contain sequences of many different hand shapes. However, the SignWriting annotations used to construct the lexicon ψ are user based, not quality checked and not specifically matching the PHOENIX data set. Therefore the annotations are very noisy, yielding a high variability of the frame alignment produced by the HMM. The best training set yielding 62.8% top-1 and 85.6 top-5 accuracy is used for all subsequent evaluations and henceforth referred to as 1-Million-Hands classifier.

Table 3 shows the per class confusion of the classifier of all 13 classes that were detected. We note that there are six classes with a precision of over 90%, two classes that reach a reasonable 60% or more, three classes that are in the 40% range and the remaining classes achieve a low precision or are not detected at all. This is a very strong result given the fact that the classifier is trained with weak annotations on the video level only and that the hand shape taxonomy understands minor finger angles as different classes. Still, the question remains, why doesn't the approach recognise all hand shapes equally well? Some possible reasons include: (i) Hand shapes in the training set are not equally distributed across the classes. (ii) Hand shapes in the evaluation set are also not equally distributed, leading to a recognition bias. (iii) There may be too few samples for the seldom occurring hand shapes. (iv) There are differences with respect to the hand shape taxonomies used for creating the hand shape labels of the different data sets. We tried to account for these differences when creating a mapping from one taxonomy to another, but there may be errors in this mapping, as we were just looking at the taxonomy description when creating the mapping, not at the data itself.

96.5	0.0	0.0	0.7	0.0	0.0	0.0	0.0	0.0	0.5	0.0	0.0	0.0
2.0	90.1	0.8	0.1	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0
0.0	2.7	94.3	2.3	0.0	0.0	0.0	0.0	0.0	1.9	0.0	0.0	0.0
0.0	0.0	0.0	49.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	18.8	41.7	6.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0
0.0	0.0	0.0	4.1	9.4	81.9	0.0	0.0	0.0	0.5	0.0	0.0	0.0
0.0	0.0	0.0	1.7	0.0	0.0	47.6	2.1	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	1.6	0.0	0.0	0.0	95.5	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	3.9	1.3	0.0	0.0	0.0
0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	64.9	0.0	0.0	0.0
0.0	0.0	0.0	3.5	38.1	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0
0.0	0.0	0.0	0.8	0.0	0.0	0.0	0.0	0.0	0.2	0.0	100.0	0.0
0.4	0.0	0.0	0.6	0.0	0.0	0.0	0.0	0.0	0.8	0.0	0.0	64.3

Table 3. Class confusion of detected classes in [%]. Showing per class precision on the diagonal, true classes on the y-axis and predicted classes on the x-axis of the 62.8% top-1 accuracy. Hand-Icons from [22].

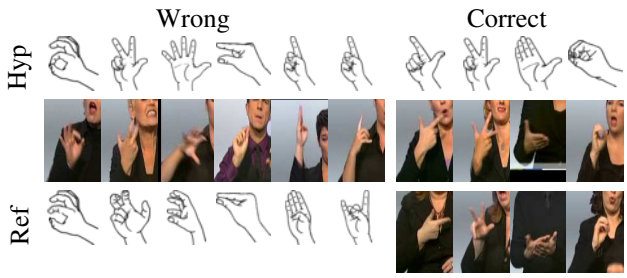

Figure 6. Some examples of correct and wrong classification on the independent evaluation set. "Hyp" refers to the hypothesised class, whereas "Ref" is the reference. Hand-Icons from [22].

Figure 6 shows examples of correct classification as well as failure cases. The figure helps to understand that in several cases (*e.g.* the first four images from the left in Figure 6) the classification is not completely wrong, but does not seem to be able to distinguish minor differences in similar hand shapes (*e.g.* in the first row the index and thumb are recognised as touching, but they are in fact slightly separated). These errors could also happen to untrained humans. The examples in the fifth and sixth column show confusions of visually similar, but for the human clearly distinguishable handshapes (*e.g.* the flat hand seen from the side looks similar to an index finger). However, the examples of correct classification in Figure 6 show us that the 1-Million-Hands model correctly classifies hand shapes independent of the pose and orientation. It also copes well with occlusions as can be seen in column three.

5.3. Continuous Sign Language Recognition

Sign language recognition (SLR) is very suitable to evaluate hand shape classification as it is a difficult but well defined problem offering real-life difficulties (w.r.t. occlusion,

motion blur, variety of hand shapes) hard to find in simple per frame evaluation tasks of current hand shape evaluation data sets. We use the same system as [18] to ensure comparability to previously published results and base the SLR recognition pipeline on [29]. We use the 1024 dimensional feature maps of the last convolutional layer of our CNN, normalise its variance to unity and use PCA to reduce the dimensionality to 200. We evaluate on two publicly available data sets: RWTH-PHOENIX-Weather 2014 Multisigner data set and SIGNUM signer-dependent set presented in Section 4 and measure the error in word error rate (WER):

$$\text{WER} = \frac{\#\text{deletions} + \#\text{insertions} + \#\text{substitutions}}{\#\text{number of reference observations}} \quad (6)$$

We compare the classifier against HoG-3D features, which are succesfully employed as hand shape feature in many state-of-the-art automatic SLR systems (*c.f.* Table 4). On RWTH-PHOENIX-Weather, we see that the 1-Million-Hands model outperforms the standard HoG-3D features by 9.3% absolute WER, being a relative improvement of over 15% from 60.9% down to 51.6%. On SIGNUM the 1-Million-Hands model outperforms the standard HoG-3D features by 0.5% absolute WER, from 12.5% down to 12.0%. On this data set the performance is less as it is more controlled and the tracking is better. This means that the HoG-3D is able to perform better on this easier data than it being a deficiency in the CNN.

We further compare our classifier in a multi-modal setup against the best published recognition results on the employed data sets and perform a stacked fusion with the features proposed by [18] (comprising HoG-3D, right to left hand distance, movement, place of articulation and facial features). Different to [18] we do not perform any sort of speaker or feature adaptation. Table 5 presents the recognition results competing the current state-of-the-art. On RWTH-PHOENIX-Weather, the 1-Million-Hands model adds significant complementary information to the complex state-of-the-art feature vector used by [18] and reduces the WER by 10.2% absolute from 57.3% to 47.1%, being a relative reduction of over 17%. On SIGNUM it reduces the WER by 2.4% absolute from 10.0% to 7.6%, being a relative reduction of 24%. It is suprising that the 1-Million-Hands model generalises so well to the completely unseen SIGNUM data set, particularly w.r.t. large visual differences in background and motion blur.

6. Conclusion

In the course of this work we presented a new approach to learning a frame-based classifier using weakly labelled sequence data by embedding a CNN within an iterative EM algorithm. This allows the labeling of vast amounts of data at the frame level given only noisy video annotation. The iterative EM algorithm leverages the discriminative ability of the CNN to iteratively refine the frame level annotation and subsequent training of the CNN. Using this approach, we trained a fine grained hand shape classifier on over 1 million weakly labelled hand shapes that distinguishes 60 classes and generalises over both individuals and datasets. The classifier achieves 62.8 % recognition accuracy on over 3000 manually labelled hand shape images which will be released to the community. When integrated into a continuous sign language recognition pipeline and evaluated on two standard benchmark corpora, the classifier achieves an absolute improvement of up to 10% word error rate and a relative improvement of over 17% compared to the state-of-the-art. To our knowledge, no previous work has explicitely worked on posture and pose-independent hand shape classification. Moreover, we believe no previous work has exploited the discriminitive power of CNNs with application to hand shape classification in the scope of sign language. Although we demonstrate this in the context of hand shape recognition, the approach has wider application to any video recognition task where frame level labelling is not available.

Acknowledegments: Special thanks to Thomas Troelsgård and Jette H. Kristoffersen, Center for Tegnsprog, Denmark (http://www.tegnsprog.dk) for providing linguistic sign language annotations and videos. We also thank the creators of the online Dictionary of New Zealand Sign Language (http://nzsl.vuw.ac.nz) for sharing their work under CC-license, which allowed us to use the hand shape icons, sign language videos and annotations. This work has been supported by EPRSC grant EP/I011811/1.

	PHOENIX 2014				SIGNUM	
	Dev		Test		Test	
	del/ins	WER	del/ins	WER	del/ins	WER
HoG-3D	25.8/4.2	60.9	23.2/4.1	58.1	2.8/2.4	12.5
1-Mio-H.	19.1/4.1	**51.6**	17.5/4.5	**50.2**	1.5/2.5	**12.0**

Table 4. Hand-only continuous sign language recognition results on RWTH-PHOENIX-Weather 2014 Multisigner and SIGNUM. 1-Mio-H. stands for the presented 1-Million-Hands classifier.

	PHOENIX 2014				SIGNUM	
	Dev		Test		Test	
	del/ins	WER	del/ins	WER	del/ins	WER
[36]	–	–	–	–	–	12.7
[13]	–	–	–	–	–	11.9
[11]	–	–	–	–	–	10.7
[18]	23.6/4.0	57.3	23.1/4.4	55.6	1.7/1.7	10.0
[18] CMLLR	21.8/3.9	55.0	20.3/4.5	53.0	–	–
1-Mio-H.+[18]	16.3/4.6	**47.1**	15.2/4.6	**45.1**	0.9/1.6	**7.6**

Table 5. Multi-modal continuous sign language recognition results on RWTH-PHOENIX-Weather 2014 Multisigner and SIGNUM. 1-Mio-H. stands for the presented 1-Million-Hands classifier.

References

[1] V. Athitsos and S. Sclaroff. Estimating 3D hand pose from a cluttered image. In *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, volume 2, pages II–432. IEEE, 2003. 2

[2] H. A. Bourlard and N. Morgan. *Connectionist speech recognition: a hybrid approach*, volume 247. Springer Science & Business Media, 2012. 3

[3] R. Bowden, D. Windridge, T. Kadir, A. Zisserman, and M. Brady. A Linguistic Feature Vector for the Visual Interpretation of Sign Language. In *Computer Vision-ECCV 2004*, pages 390–401, Czech Republic, Prague, 2004. 2

[4] 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*, pages 2961–2968. IEEE, 2009. 2

[5] 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*, pages 2568–2574, Miami, FL, June 2009. 2

[6] 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*, pages 914–919. IEEE, Nov. 2011. 2, 6

[7] 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)*, pages 1–38, 1977. 2, 3

[8] P. Dreuw, T. Deselaers, D. Rybach, D. Keysers, and H. Ney. Tracking Using Dynamic Programming for Appearance-Based Sign Language Recognition. In *IEEE International Conference Automatic Face and Gesture Recognition*, pages 293–298, Southampton, UK, Apr. 2006. IEEE. 6

[9] 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*, volume 2, pages 1471–1476. IEEE, 2006. 2

[10] H. Fillbrandt, S. Akyol, and K.-F. Kraiss. Extraction of 3D hand shape and posture from image sequences for sign language recognition. In *IEEE International Workshop on Analysis and Modeling of Faces and Gestures, 2003. AMFG 2003*, pages 181–186, Oct. 2003. 2

[11] 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*, Lecture Notes in Computer Science 7887, pages 89–99, Madeira, Portugal, June 2013. Springer. 3, 8

[12] J. Forster, C. Schmidt, O. Koller, M. Bellgardt, and H. Ney. Extensions of the Sign Language Recognition and Translation Corpus RWTH-PHOENIX-Weather. In *Language Resources and Evaluation*, pages 1911–1916, Reykjavik, Island, May 2014. 4, 5

[13] Y. Gweth, C. Plahl, and H. Ney. Enhanced Continuous Sign Language Recognition using PCA and Neural Network Features. In *CVPR 2012 Workshop on Gesture Recognition*, pages 55–60, Providence, Rhode Island, USA, June 2012. 8

[14] Jette H. Kristoffersen, Thomas Troelsgård, Anne Skov Hardell, Bo Hardell, Janne Boye Niemelä, Jørgen Sandholt, and Maja Toft. Ordbog over Dansk Tegnsprog. http://www.tegnsprog.dk/, 2008-2016. 4

[15] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional Architecture for Fast Feature Embedding. *arXiv preprint arXiv:1408.5093*, 2014. 3

[16] M. Kawulok. Fast propagation-based skin regions segmentation in color images. In *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 1–7, Apr. 2013. 1

[17] 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*, 41(2):526–541, Apr. 2011. 2

[18] 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*, 141:108–125, Dec. 2015. 5, 8

[19] 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*, pages 1–6, Shanghai, PRC, Apr. 2013. 4

[20] 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*, pages 281–296, Zurich, Switzerland, Sept. 2014. 2

[21] P. Krejov, A. Gilbert, and R. Bowden. Combining discriminative and model based approaches for hand pose estimation. In *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, pages 1–7. IEEE, 2015. 2

[22] D. McKee, R. McKee, S. P. Alexander, and L. Pivac. The Online Dictionary of New Zealand Sign Language. http://nzsl.vuw.ac.nz/, 2015. 4, 5, 7

[23] 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*, pages 2583–2590. IEEE, 2009. 2

[24] M. Oberweger, P. Wohlhart, and V. Lepetit. Hands Deep in Deep Learning for Hand Pose Estimation. *arXiv:1502.06807 [cs]*, Feb. 2015. 2

[25] M. Potamias and V. Athitsos. Nearest neighbor search methods for handshape recognition. In *Proceedings of the 1st international conference on PErvasive Technologies Related to Assistive Environments*, PETRA '08, pages 30:1–30:8, New York, NY, USA, 2008. ACM. 2

[26] N. Pugeault and R. Bowden. Spelling It Out: Real–Time ASL Fingerspelling Recognition. In *IEEE Workshop on*

Consumer Depth Cameras for Computer Vision, Barcelona, Spain, Proc ICCV, 2011. 1

[27] 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*, pages 1–8. IEEE, 2007. 2

[28] A. Roussos, S. Theodorakis, V. Pitsikalis, and P. Maragos. Dynamic affine-invariant shape-appearance handshape features and classification in sign language videos. *The Journal of Machine Learning Research*, 14(1):1627–1663, 2013. 2

[29] 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. 6, 8

[30] A. Senior, G. Heigold, M. Bacchiani, and H. Liao. GMM-free DNN training. In *Proceedings of ICASSP*, pages 5639–5643, 2014. 2

[31] V. Sutton and D. A. C. f. S. Writing. *Sign writing*. Deaf Action Committee (DAC), 2000. 4

[32] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going Deeper with Convolutions. *arXiv:1409.4842 [cs]*, Sept. 2014. 3

[33] D. Tang, T.-H. Yu, and T.-K. Kim. Real-Time Articulated Hand Pose Estimation Using Semi-supervised Transductive Regression Forests. In *2013 IEEE International Conference on Computer Vision (ICCV)*, pages 3224–3231, Dec. 2013. 2

[34] J. Tompson, M. Stein, Y. Lecun, and K. Perlin. Real-time continuous pose recovery of human hands using convolutional networks. *ACM Transactions on Graphics (TOG)*, 33(5):169, 2014. 2

[35] J. Triesch and C. von der Malsburgb. Classification of hand postures against complex backgrounds using elastic graph matching. *Image and Vision Computing*, 20:937–943, 2002. 2

[36] U. von Agris, M. Knorr, and K.-F. Kraiss. The significance of facial features for automatic sign language recognition. In *Automatic Face & Gesture Recognition, 2008. FG'08. 8th IEEE International Conference on*, pages 1–6. IEEE, 2008. 5, 8

[37] 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*, pages 1903–1910. IEEE, 2009. 2

[38] A. Wetzler, R. Slossberg, and R. Kimmel. Rule Of Thumb: Deep derotation for improved fingertip detection. *arXiv preprint arXiv:1507.05726*, 2015. 2

[39] 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*, volume 1, pages 275–280 vol.1, 2001. 2

[40] T. Yamashita and T. Watasue. Hand posture recognition based on bottom-up structured deep convolutional neural network with curriculum learning. In *IEEE International Conference on Image Processing (ICIP)*, pages 853–857. IEEE, 2014. 2

[41] 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*, 35(7):1635–1648, July 2013. 2

[42] X. Zhu. Semi-Supervised Learning Literature Survey. Technical Report 1530, Computer Sciences, University of Wisconsin -Madison, 2008. 2