Continuous Sign Language Recognition: Towards Large Vocabulary Statistical Recognition Systems Handling Multiple Signers

Oscar Koller, Jens Forster, Hermann Ney

Human Language Technology and Pattern Recognition - RWTH Aachen University, Germany

Abstract

This work presents a statistical recognition approach performing large vocabulary continuous sign language recognition across different signers. Automatic sign language recognition is currently evolving from artificial lab-generated data to 'real-life' data. To the best of our knowledge, this is the first time system design on a large data set with true focus on real-life applicability is thoroughly presented. Our contributions are in five areas, namely tracking, features, signer dependency, visual modelling and language modelling. We experimentally show the importance of tracking for sign language recognition with respect to the hands and facial landmarks. We further contribute by explicitly enumerating the impact of multimodal sign language features describing hand shape, hand position and movement, inter-hand-relation and detailed facial parameters, as well as temporal derivatives. In terms of visual modelling we evaluate non-gesture-models, length modelling and universal transition models. Signer-dependency is tackled with CMLLR adaptation and we further improve the recognition by employing class language models. We evaluate on two publicly available large vocabulary databases representing lab-data (SIGNUM database: 25 signers, 455 sign vocabulary, 19k sentences) and unconstrained 'real-life' sign language (RWTH-PHOENIX-Weather database: 9 signers, 1081 sign vocabulary, 7k sentences) and achieve up to 10.0% / 16.4% and respectively up to 34.3% / 53.0% word error rate for single signer / multi-signer setups. Finally, this work aims at providing a starting point to newcomers into the field.

Keywords: sign language recognition, statistical modelling, tracking, visual modelling, signer dependency, signer adaptation

1. Introduction

Sign languages (SLs), the natural languages of the Deaf, are known to be as grammatically complete and rich as their spoken language counterparts. Science discovered SLs a few decades ago and research promises new insights into many human language related fields from language acquisition to automatic processing.

SLs are not international and convey meaning by more than just the moving hands. They make use of both 'manual features' (hand shape, position, orientation and movement) and linguistically termed 'non-manual' features consisting of the face (eye gaze, mouthing/mouth gestures and facial expression) and the upper body posture (head nods/shakes and shoulder orientation). All of these language components are used in parallel to complement each other, but depending on the context of an utterance, a specific component may or may not be required to interpret the sign. Sometimes, an individual component plays an integral role within the sign, sometimes modifies the meaning, and sometimes provides spatial or temporal context. Furthermore, the different information channels do not share a fixed temporal alignment, but are rather loosely coupled.

Computer vision methods exist to extract features for these different channels. However, SL constitutes an extremely challenging test bed as it incorporates huge variations inherent to natural languages. High signing speed, motion blur, different lighting and view-point-dependent appearance have to be tackled. Furthermore, ambiguity is inherent to sign languages, as each movement, each change in eye gaze or each appearance of the tongue may or may not have a grammatical or semantic function depending on the context. Thus, learning features and training classifiers that can be applied to SL recognition must cope with a natural variation seldom present in other tasks. At the same time, it constitutes a very well-defined environment for assessing gesture recognition techniques by providing rules and boundaries for naturalness and intelligibility.

Historically, research on automatic sign language recognition (ASLR) had mainly access to small data sets, limited number of signers and a limited recognition vocabulary. Recently, a very exciting era has started. SL research is moving out of the lab into 'real-life' scenarios.

In this paper, we present extensive results and thorough analysis on, to our knowledge, the currently biggest publicly available corpus of continuous sign language (RWTH-PHOENIX-Weather). It covers only 'real-life' signing recorded on public TV broadcast that has been manually labelled by native speakers. To the best of our knowledge, this is the first time, system design on a large data set with true focus on real-life applicability is thoroughly presented. Our contributions are

Email addresses: koller@cs.rwth-aachen.de (Oscar Koller), forster@cs.rwth-aachen.de (Jens Forster), ney@cs.rwth-aachen.de (Hermann Ney)

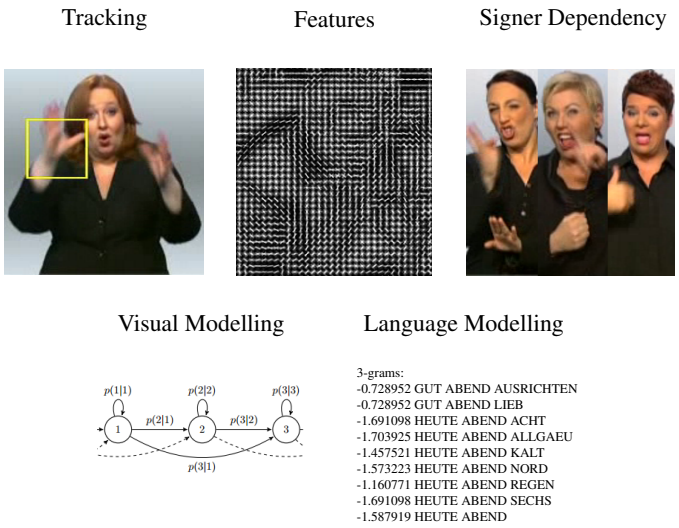

Figure 1: Areas tackled by this paper.

in five areas, namely tracking, features, signer dependency, visual modelling and language modelling.

We experimentally show the importance of tracking for sign language recognition, with respect to the hands and facial landmarks. We further contribute by explicitly enumerating the impact of multimodal sign language features describing hand shape, hand position and movement, inter-hand-relation and detailed facial parameters, as well as temporal derivatives. Among these, the combination of hand gesture features and face features is novel, as well as the definition of the high-level face features.

In terms of visual modelling we evaluate non-gesture-models, length modelling and universal transition models. Signer-dependency is tackled using constrained maximum likelihood linear regression (CMLLR) adaptation. Further, class language models, CMLLR adaptation, as well as non-gesture-models are the new aspects to ASLR.

In Section 2, we introduce the state-of-the-art in the context of sign language recognition and its related sub-fields. In the following two sections, we first present the employed data sets used for evaluating this work (Section 3) and then, in Section 4, the overall recognition system is explained in detail.

The subsequent sections tackle each of the five areas depicted in Figure 1, giving first the technical details and then the experimental evidences. This is meant to open up the field to newcomers, who can estimate the impact of the most important design decisions. In Section 5, the employed tracking techniques are discussed and their impact with respect to the hands and facial landmarks is given. Section 6 presents the employed features covering most important modalities for SL and shows the impact on overall recognition results. Methods improving the visual modelling are presented in Section 7. Our approach to tackling multiple signers is presented in Section 8. The experimental sections end with our contribution to language modelling in Section 9. Finally, the paper closes with a conclusion and discussion of future work in Sections 10 and 11.

2. Related Work

This section describes related work in ASLR and its related disciplines. The field evolved from recognising isolated signs of very limited number, articulated by only a single signer towards more complex settings with continuous natural signing of multiple signers. Thereby, the scientific community advances three tracks simultaneously:

1. The methods to extract relevant information become more sophisticated and precise, moving from expensive glove- and accelerometer-based setups to non-intrusive computer vision techniques.
2. The modelling of SL evolves to accommodate both linguistic and data-driven findings, aiming to fully reflect the complexity of the visual language.
3. The available data sets become more challenging, bigger and closer to real-life signing.

Although more recently ASLR is starting to tackle 'real-life' continuous signing data, the majority of work in the community still focuses on the recognition of isolated signs mostly in artificial settings.

2.1. Sign Language Recognition

Tamura et al. [64] were the first to start exploring the world of ASLR. They built a system to recognise isolated signs of Japanese SL by modelling the shape, movement and location of the hand using a simple colour segmentation. A lot of the early ASLR systems then employed glove-based motion tracking systems to overcome difficulties with vision-based feature extraction and tracking. This allowed to increase the recognition vocabularies while still achieving high accuracy on simpler tasks. In this way, Kadous et al. [34] distinguished 95 Australian Sign Language (AUSLAN) signs with accuracies of around 80% using decision trees as classifier. Two years later Liang et al. [42] proved to recognise a lexicon of 250 different signs of Taiwanese SL with a similar error rate. However, due to high cost of motion capture systems and thus low real-world applicability, coloured gloves and computer vision techniques started emerging [13].

More advanced visual tracking methods allow to design non-intrusive vision-based approaches that do not require the signers to wear any sort of gloves [3, 76]. Starner et al. [60] mounted a camera into a hat and used hidden Markov models (HMMs) to recognise a data set of 40 different American Sign Language (ASL) signs. A good overview of ASLR is given in [52]. Furthermore, a number of researchers consider the problem as sign spotting [54, 20], where the aim is not to recognise whole sentences, but rather single instances of signs within sentences.

Sign Sub-Units. The need to break whole signs up into sub-units in order to use limited training data more efficiently and in order to allow scaling up of the vocabulary has been an unsolved problem since the early work of Waldron et al. [70]. Inspired by grapheme-to-phoneme conversion, Pitsikalis et al. [55] extract sub-unit definitions from linguistic annotation in HamNoSys [31] to improve an HMM-based system recognising isolated sign in Greek sign language. Cooper et al. [7] compare

boosted sequential pattern trees to HMMs using linguistically inspired sub-units and 3D tracking information finding that the trees outperform HMMs for BSL. Koller et al. [39] employ an open SignWriting [63] dictionary to produce and align linguistically meaningful sub-units to signs in German Sign Language (DGS).

Unsupervised approaches. To tackle limited or even no available training data, several works [36, 6, 47, 53] aim at exploiting co-occurrences of weak cues to learn hand-based sign language models.

Handling Co-articulation and Noise. Signs differ based on the preceding and following sign leading to huge visual variability. To overcome these co-articulation issues, background, noise and co-articulation modelling is needed. Lee et al. [41] investigate adaptive thresholding for individual sign HMM and report recognition improvements on a very limited corpus. Another approach to model co-articulation by Yang et al. [74] uses nested dynamic programming to optimise the time sequence of a co-articulation movement separate from the signs. The method was evaluated on a 39 vocabulary continuous sign language corpus and a drastic reduction in error rate by 70% is reported. Kelly et al. [35] use a very similar threshold model approach and report 5.2% improvement on an isolated data set giving testament to the limitations of the corpus used in [41].

Modality Combination. As mentioned in Section 1, SLs consist of multiple parallel information streams, also referred to as modalities. The fusion of these modalities has been an active field of research within the community. One approach is to use parallel HMMs, which are reported to recognise isolated signs in American and Chinese sign language achieving recognition accuracies over 90% [12, 71]. Also the fact that modalities can occur in a time asynchronous way has been considered during modelling [22]. Vogler and Metaxas [68] investigate parallel HMMs (PaHMM) for recognition of continuous ASL using cyber-gloves for feature extraction. They report an improvement from 6.7% to 5.8% word error rate (WER) for 22 signs using 400 training and 99 test sentences. Theodorakis et al. [65] evaluated product HMMs for the recognition of 93 isolated, Greek sign language signs and reported that an asynchronous combination of features outperformed a synchronous combination. Aaran et al. [1] implement a fusion technique for hand shapes and facial expression/shoulder motion that only considers the second feature when the decision based on the first information stream has low confidence. Forster et al. [22] investigate techniques to combine not perfectly synchronous information streams within an HMM-based ASLR system finding that synchronisation just at word boundaries improves the recognition performance. Ong et al. [51] use boosted hierarchical sequential pattern trees to recognise isolated and continuous signs in British Sign Language (BSL), DGS and ASL. Their approach seems promising by allowing to combine partly parallel, not perfectly synchronous features through feature selection by the trees. However, on continuous data the approach faces difficulties.

Recognition and Translation. As SLs represent full languages with their own grammar and syntax, an additional translation step should follow recognition in order to bridge the communication gap between deaf and hearing. Tokuda et al. [66] mention an important problem: the SL word inventory is much smaller than the spoken language counter part. However, this is not due to a limited vocabulary, but rather remains an unsolved problem of neglecting SL concepts (i.e. modifier, classifier, indexing) and non-manual features in recognition. Schmidt et al. [59] address this problem by linking a mouthing recognition to the subsequent translation. Other works looking at recognition and translation include Bauer et al. [2], who perform a HMM-based recognition of 100 DGS signs with an accuracy of over 90% and translate it into German text, and [61, 23].

2.2. Features

Feature extraction is an important step in a recognition system. Over the last decades, different features emerged that proved to be successful in the task of ASLR. This subsection aims at depicting how the task of extracting relevant visual information evolved until now.

Basic Features. Humans can understand SL by looking at a video sequence, thus the information must be present in the images. As a naïve descriptor, the rgb values of the full image, patches of the tracked hand or the face can serve as features [10, 21].

Features are often chosen reflecting the knowledge of sign linguists. It is known that the manual channel (hand shape, orientation, position and movement) conveys a big part of the information in SLs. Tracking the hands and extracting advanced features based on their positions is an important requirement in order to focus feature extraction on relevant video/image regions. Histogram of oriented gradients (HOG) by [11] and other 2D feature point descriptors, such as scale invariant feature transformation (SIFT) [44] are frequently encountered in ASLR approaches [8, 53]. HOG-3D [38], an extension over time of HOG, has also shown to produce state-of-the-art-performance [21]. Often trajectories of each single hand or of the interaction of both hands are also used as features [30]. Gabor responses of the forehead have shown to capture non-manual facial expression [43].

High-Level Features. 3D models of hands [45] or faces [58] are used to find higher-level concepts within signing sequences, such as opening the eyes, raising the eyebrows or turning a hand. Recently, viseme patterns have been proposed [40] to reflect mouthings performed during signing. Finally, inspired by the success of neuronal-network-based features in automatic speech recognition, the same concepts are tested for SL [29].

2.3. Sign Language Databases

Current publicly available video-based sign language corpora can be grouped into one of three categories depending on the scientific community they originated from.

1. lexical data sets for every day use
2. linguistic data sets
3. large data sets for pattern recognition purposes

First, there are corpora intended as video-based lexica for sign languages allowing to track and analyse changes in the

vocabulary of sign languages from a linguistic point-of-view. 'The American Sign Language Lexicon Video Dataset' [48] forms such a lexicon for American sign language (ASL), containing more than 3000 signs in multiple video views. The AUSLAN SignBank project[1] provides annotations on a variety of linguistic levels for 357 videos of Australian sign language.

Second, there are corpora intended for linguistic research on isolated signs and continuous sign language allowing to tackle questions like appearance of dialectic variances, differences in pronunciation and sentence structures. Typically, such corpora are created under lab-conditions focusing on certain aspect of sign languages. Corpus NGT [9] contains 12 hours of signing in upper-body and front view totalling 64 000 annotated glosses. Since 2008 the corpus has been extended by translations into various spoken languages. Rutkowski et al. [56] created a corpus for Polish sign language containing about 300h of video footage of 80 deaf signers performing predefined language tasks. The CopyCat corpus [75] covers ASL spoken by children in 420 phrases formed from a vocabulary of 19 signs. For further reference, the University of Hamburg, Germany, created a summary on available linguistic sign language corpora[2].

Third, there are corpora either explicitly created or adapted for natural language processing and/or computer vision tasks. In contrast to the linguistic resources, these corpora feature smaller vocabularies of a couple of hundred signs instead of thousands, higher type/token ratios and focus on a small number of closed language domains. The overall goal is to provide minimum statistics to allow for robust training of statistical models while refraining from focusing on special concepts of SLs. Dreuw et al. [17] give an overview on such corpora. Included in this survey are the RWTH-BOSTON corpora originally created for linguistic research at Boston University and adapted for pattern recognition purposes by RWTH Aachen University featuring multiple signers and up to 7,768 running glosses with a vocabulary size of 483 glosses. Some works [19, 4] present corpora for isolated and continuous sign language recognition for German, Greek, British and French sign language created in the course of the Dicta-Sign[3] project. The corpora include sign language videos shot in high-definition in frontal and side view under controlled lab-conditions. Similar to Corpus NGT, the Dicta-Sign corpora contain bird's eye views of the signers allowing for the study of hand movements in the signing space with regard to the distance from the upper-body of the respective signer. The SIGNUM corpus [69] has been explicitly created for pattern recognition purposes foregoing linguistic considerations and consists of 25 signers and nearly 14,000 running glosses in DGS. Moreover, there is a recent efforts to develop an isolated sign language data set providing depth information [73].

[1] www.auslan.org.au

[2] www.sign-lang.uni-hamburg.de/dgs-korpus/index.php/sl-corpora.html

[3] www.dictasign.eu

3. Data Sets: Overcoming Artificial and Small Corpora

Statistical approaches to automatic speech recognition (ASR) require large corpora of annotated text respective audio data to learn robust models that generalise well to unseen data. There is a lack of suitable video corpora to develop systems employing statistical methods targeting ASLR. SL corpora are mainly recorded for linguistic research, not providing the type/token ratios needed for statistical modelling. Typically, this kind of data differs significantly from the real language encountered outside the research lab. One concept used particularly in linguistic corpora is the concept of staged communicative events trying to elicit special aspects of SL communication. Staged communication events focus on the interaction between one or more signers. While this makes the language encountered more natural, it raises automatic processing to a difficulty level not yet in focus of the machine learning and pattern recognition community.

The difficulty of the corpora situation is further compounded by the fact that SLs are purely visual languages lacking a writing system. The lack of a normed or at least agreed writing system leads to a variety of different annotation schemes including gloss notation, HamNoSys [31] and SignWriting [63].

In this work, two of the largest publicly available SL video corpora are used to investigate statistical modelling and recognition of SLs.

Both corpora feature DGS and use a gloss annotation scheme. The gloss annotation scheme uses words from the enclosing spoken language, e.g. written English in case of BSL, to describe the meaning of a sign rather than its appearance. The SIGNUM database [69] and the RWTH-PHOENIX-Weather database [24] both come with defined recognition setups for signer dependent ASLR as well as multi-signer setups.

The corpora statistics of the single signer setups for both corpora are subsumed in Table 1, the corpora statistics for the multi-signer setup of SIGNUM and RWTH-PHOENIX-Weather are presented in Table 2.

	SIGNUM		PHOENIX	
	Train	Test	Train	Test
duration [h]	3.85	1.05	0.51	0.075
# frames	416,620	114,230	46,282	6751
# sentences	1809	531	304	47
# running glosses	11,109	2805	3309	487
vocabulary size	455	-	266	-
# singletons	0	0	90	-
out-of-vocabulary [%]	-	0.1	-	1.6

Table 1: Corpus statistics: single signer subsets of SIGNUM and RWTH-PHOENIX-Weather.

The SIGNUM database has been created for pattern recognition purposes and aims at reducing the overall complexity of SL and the associated recognition task. Native signers were asked to sign predefined sentences from the domain of every day life, e.g. going to the cinema, waiting for a bus, and are wearing black long-sleeved clothes while standing in front of an dark blue background. The dataset is carefully controlled

	SIGNUM MS		PHOENIX MS		
	Train	Test	Train	Dev	Test
# signers	25	25	9	9	9
duration[h]	33.5	9.2	10.71	0.84	0.99
# frames	3,618,630	996,270	963,664	75,186	89,472
# sentences	15,075	4,425	5,672	540	629
# unique sentences	603	177	5,672	540	629
# running glosses	92,575	23,350	65,227	6,032	7,089
vocabulary size	455	-	1,081	467	500
# singletons	0	0	329	148	167
out-of-vocabulary [%]	-	0.1	–	0.50	0.54

Table 2: Corpora statistics: multi-signer setups of SIGNUM and RWTH-PHOENIX-Weather.

w.r.t. the signer's position towards the camera, the lighting and the signing speed. In the signer dependent subset of SIGNUM, the signer was asked to sign each of the 603 predefined sentences for training and the 177 test sentences 3 times. In the multi-signer setup there are 25 signers performing the sentences only once. Due to the overall staged character and the controlled conditions, the SIGNUM database must be considered to contain artificial lab data limiting the expressiveness of results obtained on the signer dependent but also on the multi-signer setup. Results obtained on this database are not expected to easily carry over to more challenging data containing unconstrained SL recorded outside the research lab.

Video recordings belonging to the SIGNUM corpus are recorded at 780 × 580 pixels and 30 frames per second. Example frames from the corpus are presented in Figure 2.

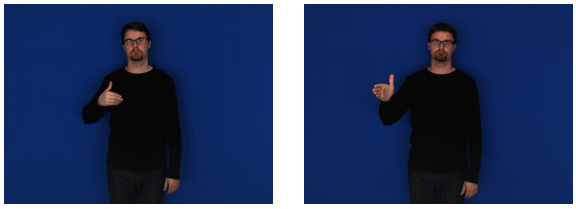

Figure 2: Example frames from SIGNUM corpus.

In contrast to the SIGNUM corpus, the RWTH-PHOENIX-Weather corpus contains SL aired by the German public TV station PHOENIX in the context of weather forecasts as part of the daily news broadcast. Hearing SL interpreters perform live and on-the-fly interpretation of the spoken weather forecast into DGS. This setup leads to SL content that follows the content of the spoken weather forecast and is influenced by the grammatical structure of the spoken weather forecast while featuring speech effects found in unconstrained sign language conversations. Among these speech effects are false starts, hesitations and the use of dialectic pronunciation variants.

Lighting conditions and the positioning of the signer in front of the camera are controlled by the TV studio. All videos have a resolution of 210 × 260 pixel and 25 interlaced frames per second. The low temporal and spatial resolution is due to the broadcast method used by the TV station.

Figure 3 shows an example frame from the original video

Figure 3: RWTH-PHOENIX-Weather Example of original video frame. The sign language interpreter is shown in an overlay on the right of the original video frame.

stream broadcast by PHOENIX. The broadcast of the weather forecast is overlayed with the sign language interpreter leading to the aforementioned spatial resolution of 210 × 260 pixels.

Figure 4 shows the distribution of produced signs per signer in the PHOENIX MS corpus. This also underlines the unconstrained signer coverage in the corpus as Signer 1 features more than 25% of the corpus, while Signer 2 just produces less than 1% of the signs (measured by annotated glosses).

For the RWTH-PHOENIX-Weather single signer setup various annotations are available, such as annotated hand and face tracking positions, annotated hand shapes, manual variant annotation and a manually restricted vocabulary to join visually identical signs. For details refer to [24] and visit our website [4] for download instructions. Methods developed for RWTH-PHOENIX-Weather are expected to carry over to other real-life corpora.

Table 3 shows all previously published results on the RWTH-PHOENIX-Weather and SIGNUM datasets for comparison. Up to now, the best results have been 10.7% WER by [22] and 22.1% by [21] on SIGNUM single and multi-signer and 38.6% WER by [21] on RWTH-PHOENIX-Weather single signer. No results have been published for the recent RWTH-PHOENIX-Weather multi-signer subset.

[4] http://www-i6.informatik.rwth-aachen.de

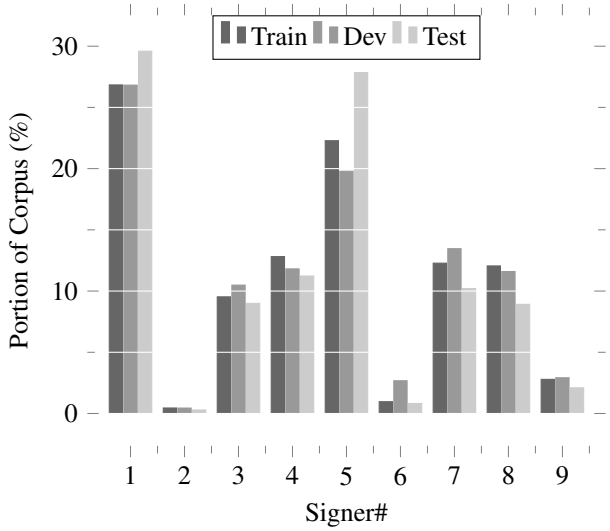

Figure 4: Portion of PHOENIX MS corpus in % per signer based on number of glosses.

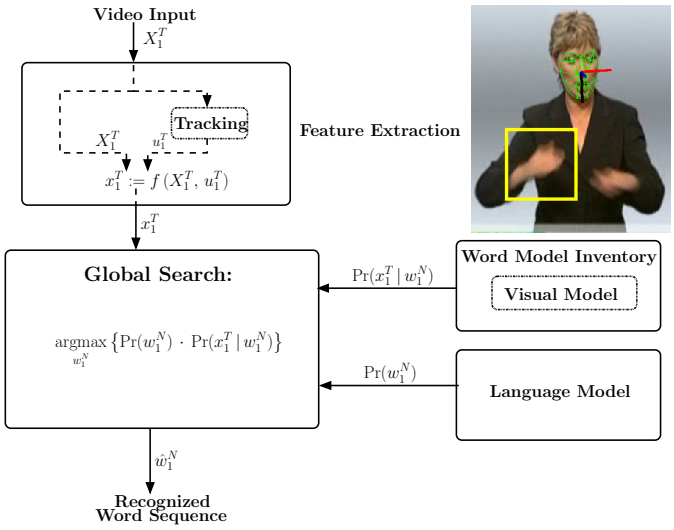

Figure 5: Continuous Recognition System Overview

	SIGNUM	SIGNUM MS	PHOENIX
Von Agris et al. [69]	12.7	-	-
Gweth et al. [29]	11.9	-	-
Forster et al. [22]	10.7	-	41.9
Forster et al. [21]	-	22.1	38.6

Table 3: Published WERs in [%] on SIGNUM single and multi-signer and RWTH-PHOENIX-Weather single signer subsets.

4. Statistical Modelling for Automatic Sign Language Recognition

Statistical approaches to ASR have matured to a point where they are used in daily life by millions of people around the globe. The great benefit of statistical approaches is their ability to learn from data foregoing the need for hand crafted recognition and grammar rules. Due to the ability to learn from data, statistical approaches lend themselves to ASLR because rules, lexica and even basic concepts such as sentence boundaries are not (yet) defined in the area of sign language linguistics.

The ASLR system used in this work is based on the freely available state-of-the-art open source speech recognition system RASR [57] and follows the system schematic in Figure 5. Given a sequence of features $x_1^T = x_1, \ldots, x_T$, the system searches for an unknown sequence of words $w_1^N = w_1, \ldots, w_N$ for which the sequence of features x_1^T best fits the learned models. To this end, the posterior probability $Pr(w_1^N|x_1^T)$ over all possible word sequences w_1^N with unknown number of words N is maximised. Using Bayes' decision rule, casting the *visual model* $Pr(x_1^T|w_1^N)$ as the marginal over all possible HMM temporal state sequence $s_1^T = s_1, \ldots, s_T$ for word sequence w_1^N, as well as assuming a first order Markov dependency and maximum approximation,

$$x_1^T \rightarrow [w_1^N]_{\text{opt}} = \underset{w_1^N}{\text{argmax}} \left\{ Pr(w_1^N) \max_{s_1^T} \left\{ Pr(x_t|s_t, w_1^N) \cdot Pr(s_t|s_{t-1}, w_1^N) \right\} \right\} \quad (1)$$

where $Pr(w_1^N)$ is the *language model*. $Pr(\cdot)$ denotes the true probablity density function (PDF) of the investigated entities. In reality, the true PDFs of the system are unknown and must be estimated from data. We denote, PDFs estimated from data by $p(\cdot)$. Considering Equation (1), the *visual*, the *language*, as well as the *state transition* model $Pr(s_t|s_{t-1}, w_1^N)$ need to be estimated. The state transition model can be estimated from a lattice of HMM alignments using the extended Baum-Welch algorithm but for speech related tasks it is typically sufficient to pool the model over all words w and assign fixed values (see Equation (2)) [57] without losing recognition performance.

$$p(s_t|s_{t-1}, w_1^N) = p(s_t - s_{t-1}) = \begin{cases} f_0 / \sum_{i=0}^{2} f_i & : \quad s_t - s_{t-1} = 0 \\ f_1 / \sum_{i=0}^{2} f_i & : \quad s_t - s_{t-1} = 1 \\ f_2 / \sum_{i=0}^{2} f_i & : \quad s_t - s_{t-1} = 2 \\ -\infty & : \quad \text{otherwise} \end{cases}, \quad (2)$$

where state sequence variables s_i denote the state with number y at time i. Thereby $s_t - s_{t-1} = 0$ defines a loop in the HMM (stay in the same state) while $s_t - s_{t-1} = 1$ describes a forward transition.

The *visual* model forms the core of the HMM for each word w, modelling for each state s of a word w how on average w is represented in the feature space. As the amount of available data for ASLR is not large enough to benefit from recent advancements in the field of artificial neural network (ANN) for ASR, Gaussian mixture models (GMMs) are used in this work.

In particular, for state s of word w

$$p(x|s,w) = \sum_{m=1}^{M} c_m \cdot \mathcal{N}(x, \mu_m, \Sigma) \tag{3}$$

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

where $\mathcal{N}(x, \mu, \Sigma)$ is a multi-variate Gaussian with mean μ, covariance matrix Σ and M is the number of mixture components (can differ between states of the same word). The proposed system uses a globally pooled covariance matrix Σ (see Equation (3)) to cope with the low amount of training samples per state and word. The expectation maximization (EM) algorithm is used to estimate the sufficient statistics of the GMMs. The number of EM iterations is optimised during the training phase of the system. The *language* model forms a distribution over the target sequence of words w_1^N and is learned form text sequences. In this work, a n-gram language model ($n = 3$ and $n = 4$) is constructed using modified Kneser-Ney discounting [5]. Discounting allows to shift probability mass from seen n-grams in training such as *HEUTE REGEN STARK* to lower-order n-grams such as *REGEN STARK* and especially unseen words. The standard toolbox *SRILM* [62] is used to estimate a 3-gram language model for all databases used in this work.

Moving from modelling towards decoding, time-synchronous word-conditioned tree search with dynamic programming is used expanding all *state hypotheses* $Q_v(t, s)$ in all trees for each time step t to decode an unknown SL sequence in a video.

$$Q_v(t,s) \quad = \quad \max_{\sigma} \; \{p(x_t, s|\sigma) \cdot Q_v(t-1, \sigma)\} \tag{5}$$

denoting the joint probability for the best partial path up to time t ending in state s with the best predecessor state σ and predecessor word v. In case of a word end state the state hypotheses of the previous time step are weighted by the language model to obtain new state hypotheses for the next word.

$$Q_v(t, s=0) \quad = \quad \max_{u} \{p(v|u) \cdot Q_u(t, S_v)\}, \tag{6}$$

where u is the predecessor word at the previous time step, v is the predecessor of the new state hypothesis, S_v is the ending state of word v, $s = 0$ is the virtual starting state, and $p(v|u)$ is a 2-gram language model for simplification. A new state tree with predecessor word v and virtual starting state $s = 0$ is then started and the whole process repeated until the end of the current sentence is reached.

The resulting system is implemented in C++ under Linux/Unix making use of multi-threaded linear algebra packages such as BLAS (Intel MKL and similar implementations), and ffmpeg for video processing. A typical recognition experiment in ASLR has a real time factor of 3 (1 minute of video footage takes 3 minutes to process) which can be brought down to close to real-time by pruning of the search space leading to reduced performance in recognition metrics. While model training is parallelised over a computing cluster of 200 computing nodes featuring up to 64GB of memory and I4 Intel CPUS, decoding is not parallelised as time-synchronous word-conditioned tree search is not parallelisable.

5. Tracking

When looking at a deaf person signing, it is immediately apparent that information is conveyed through several moving body parts. But how important is accurate tracking of body parts for ASLR? How much does it contribute to the recognition of SL and how does it compare to just using the whole image for feature extraction?

In the following subsections we describe the tracking methods used to create a state-of-the-art ASLR system capable of handling sign language data recorded outside the research lab. In particular, Sections 5.1 and 5.2 detail the algorithms for robustly tracking the signer's hands and face.

5.1. Tracking Hands

The hands of a signer convey the majority of information when signing in any given SL. Information is encoded in the appearance, shape and movement of the hands. Therefore, it is a necessity to track a signer's hands to extract movement information as well as to be able to extract features representing the appearance and shape of a particular hand.

State-of-the-art tracking systems mainly follow the tracking-by-detection paradigm in which a complex model of the object to be tracked is learned, the object is detected in every frame, and detections are linked between frames. Drawbacks of the tracking-by-detection paradigm are the taking of potentially wrong local decisions, limiting the context of a tracking decision to the detection result in the current video frame or preceding frames only, as well as additional image segmentation steps to aid the detection of the object of interest. These drawbacks often lead to a tracking loss of the object if the object is occluded, undergoes a change in appearance or shape (i.e. non-rigid object), or moves in a fast and unexpected way.

In this work, we employ a model-free tracking system that is based on dynamic programming allowing to adapt the system to arbitrary tracking tasks by choosing adequate local scoring functions. The dynamic programming tracking (DPT) system is part of the open-source, automatic large vocabulary speech recognition system RASR [15, 16, 14].

DPT avoids potentially wrong local decisions by optimising tracking decisions over time and tracing back the best (partial) sequence of tracking decisions at the end of a video sequence thus finding the optimal tracking path w.r.t. to a chosen optimisation criterion. The actual decision on the movement of the object over time is made by tracing back the best, in the sense of accumulated scores, sequence of decisions from the end of the video on. Using this two step procedure of first accumulating scores over time and second tracing back the best sequence of decisions, DPT avoids taking possibly wrong local decisions and yields the optimal solution for the tracking problem at hand [50, 49] guaranteeing a smooth tracking path.

Optimising over the whole video sequence mitigates the problem of self-occlusion unless the self-occlussion continues for a prolonged period of time (typically 1 second for a 25 fps video).

Taking inspiration from the time alignment problem that needs to be solved in ASR, DPT uses dynamic programming to break down the complex tracking problem in a set of smaller sub-problems. These sub-problems correspond to tracking the object of interest in a certain time window,e.g. from time t to $t+1$, and lead to a series of decision steps over time. As depicted in Figure 6, DPT is composed of a forward and backward step. In the forward step, every possible area of interest is associated with a score which is maximised over time in the dynamic programming framework. In the backward step, the best object path is created by tracing back the decisions that led to the best overall score after the forward step.

In the following, let X_t denote a video frame of size $I \times J$, $I, J \subset \mathbb{N}$, Pixel at time t and $l_t = (i, j) : i \in I, j \in J$ denote a *location* at time t in X_t. Finding the best tracking path $l_1^T = l_1, \ldots, l_t, \ldots, l_T$, $1 < t < T$, for an object in the image sequence $X_1^T = X_1, \ldots, X_t, \ldots, X_T$ corresponds to maximising the log-likelihood of l_1^T given X_1^T:

$$[l_1^T]_{\text{opt}} = \underset{l_1^T}{\operatorname{argmax}} \left\{ \sum_{t=1}^{T} \log p(l_t | l_1^{t-1}, X_1^T) \right\} \tag{7}$$

Assuming a first-order Markov process, i.e. the location of the object to be tracked at time t depends only on its location at time $t-1$, Equation (7) is simplified to

$$[l_1^T]_{\text{opt}} = \underset{l_1^T}{\operatorname{argmax}} \left\{ \sum_{t=1}^{T} \log p(l_t | l_{t-1}, X_{t-1}^t) \right\} \tag{8}$$

In the DPT framework, Equation (8) is reformulated by expressing $p(l_t | l_{t-1}, X_{t-1}^t)$ via a relevance scoring function $\tilde{q}\left(l_t, l_{t-1}, X_{t-1}^t\right)$ depending on the object's position in the current and previous video frame. Normalising to fulfil the requirements of a probability density function and dropping the logarithm because of its monotonicity, Equation (9) describes the final optimisation criterion of the used DPT framework

$$[l_1^T]_{\text{opt}} = \underset{l_1^T}{\operatorname{argmax}} \left\{ \sum_{t=1}^{T} \frac{\tilde{q}\left(l_t, l_{t-1}, X_{t-1}^t\right)}{\sum_l \tilde{q}\left(l, l_{t-1}, X_{t-1}^t\right)} \right\}, \tag{9}$$

where $\tilde{q}\left(l_t, l_{t-1}, X_{t-1}^t\right)$ is split into an image-independent smoothness function $T(l_t, l_{t-1})$ called *jump penalty* and an image-dependent scoring function $q\left(l_t, l_{t-1}, X_{t-1}^t\right)$.

$$\tilde{q}\left(l_t, l_{t-1}, X_{t-1}^t\right) = q\left(l_t, l_{t-1}, X_{t-1}^t\right) - T(l_t, l_{t-1}) \tag{10}$$

Dropping the normalisation term in Equation (9)[5] and using Equation (10) it is straight forward to define the necessary auxiliary quantities to define the dynamic programming recursive equations. Let $D(t, l_t)$ be the score of the best partial tracking path that starts at time $t = 1$ and ends at time t at location l_t and $B(t, l_t)$ the predecessor location corresponding to $D(t, l_t)$. Then the dynamic programming equations are defined by

$$D(t, l_t) = \max_{l' \in \mathcal{M}(l)} \left\{ D(t-1, l') - T(l_t, l') + q(l_t, l', X_{t-1}^t) \right\} \tag{11}$$

$$B(t, l_t) = \underset{l' \in \mathcal{M}(l)}{\operatorname{argmax}} \left\{ D(t-1, l') - T(l_t, l') + q(l_t, l', X_{t-1}^t) \right\} \tag{12}$$

where $\mathcal{M}(l)$ is the set of possible predecessor locations of l according to a chosen dynamic model.

While Equations (11) and (12) describe the forward step of DPT, the backward step is accomplished by tracing back the locations belonging to the best, in the sense of maximal score, tracking path found in the forward step. In particular starting with the location at time T

$$l_{T,\text{opt}} = \underset{l}{\operatorname{argmax}} \{D(T, l)\} \tag{13}$$

belonging to the best tracking path, the remaining locations are found by iteratively looking up

$$l_{t,\text{opt}} = B(t+1, l_{t+1}) \tag{14}$$

for $t = 1, \ldots, T-1$.

Please note that if $\mathcal{M}(l)$ is small and the image dependent scoring function considers only a small area around a location or only the locations themselves, DPT can be considered an instance of correlation-based optical flow [32] or zero-order warping between images [37, 26, 27].

Considering real-world sign language videos with complex object interactions, it is very difficult to consistently describe and track an object using only one relevance scoring function. Therefore, the DPT framework allows to use a weighted sum of local scoring functions to track an object of interest.

$$q(l_t, l_{t-1}, X_{t-1}^t) = \sum_{n=1}^{N} \alpha_n \cdot q_n(l_t, l_{t-1}, X_{t-1}^t) \tag{15}$$

The local scoring functions $q_n(l_t, l_{t-1}, X_{t-1}^t)$ cover different aspects of an object and can be calculated on dense probability images calculated from the original video frames. Examples of such dense probability images are skin-colour probability images or face probability images computed from responses of the Viola & Jones face detector [67].

In the following, Q_t denotes the *tracking window* and it contains all locations within a rectangle of size $w \times h$ centred on the image location l_t.

$$Q_t := \{l_t + l : l \in Q\} \tag{16}$$

$$Q := \{(i, j) : -w \leq i \leq w,\ -h \leq j \leq h\} \tag{17}$$

The size of the tracking window is fixed and equal for all local scoring functions used within the DPT framework and is a crucial parameter to be adjusted to the video sequence at hand. If w or h is chosen too small tracking performance degrades because the object of interest is not enclosed in the tracking window. Conversely, if w or h is chosen too large the score reflects the

[5] constant with regard to *argmax* function

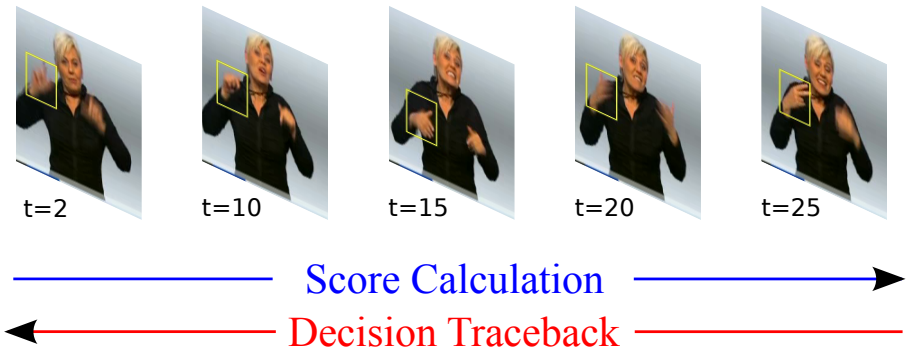

Figure 6: Tracking principle: Tracing back the best tracking path on the RWTH-PHOENIX-Weather database; yellow rectangles visualise tracked window.

background rather than the object of interest. $X[l]$ denotes the feature vector (can be as simple as a single gray-scale value) at Pixel $l = (i, j)$ in image X. If $X[l]$ describes a vector of size m in the following equations the mathematical operations are carried out channel-wise, i.e. $X[l] - X[u] = \sum_m (X[l]_m - X[u]_m)$ but the channel-wise notation is dropped for readability.

The majority of scoring functions used in this work are based on the assumption that the object of interest moves more than every other object within the sequence of images under consideration. This assumption holds even in the context of complex and cluttered backgrounds as long as the background remains static. Scoring function used in this work are:

5.1.1. Constant Object Appearance:

Assuming a high enough frame rate, i.e. temporal sampling, the object of interest is nearly constant in its appearance leading to a scoring function dubbed *constant-object-appearance*, implying a small distance between the appearance of the same object in consecutive video frames

$$q(l_t, l_{t-1}, X_{t-1}^t) = -\sum_{l \in Q} (X_t[l_t + l] - X_{t-1}[l_{t-1} + l])^2 \qquad (18)$$

Equation (18) describes a negative distance in order to fit into the maximisation framework of Equation (9). The taken assumption is similar to the base assumption of optical-flow [32] and a prerequisite for particle filtering techniques [33].

5.1.2. Constant Background:

Following the high frame rate assumption, we assume that the background of the object to be tracked is constant or nearly constant between consecutive video frames. This implies that only those regions change between consecutive frames where the object is in the current frame X_t and where it has been in the previous frame X_{t-1}. Accordingly, the difference between all other parts of the consecutive video frame pair should be minimal.

$$\begin{aligned} q(l_t, l_{t-1}, X_{t-1}^t) = &\sum_{l \in Q_t} (X_t'[l])^2 + \sum_{l \in Q_{t-1}} (X_t'[l])^2 \\ &- \sum_{l \in Q_t \cap Q_{t-1}} (X_t'[l])^2 \end{aligned} \qquad (19)$$

where $X_t' = X_t - X_{t-1}$.

5.1.3. Soft Spatial Pruning:

The human body has certain kinematic constraints with regard to where in a video frame e.g. the right hand of a person can be when facing the camera. We encode a soft form of this constraint into the tracking framework via

$$q(l_t = (i, j), l_{t-1}, X_{t-1}^t) = \begin{cases} \lfloor \lambda \cdot \tau_i \rfloor & : i > \lfloor \lambda \cdot \tau_i \rfloor \\ i - \tau_i & : \tau_i \leq i \leq \lfloor \lambda \cdot \tau_i \rfloor \\ -f(\tau_i - i) & : 0 \leq i < \tau_i \end{cases} \qquad (20)$$

where $0 < \tau_i < I$ and $\lambda \geq 1$ are constants, and $f : \mathbb{N} \mapsto [0, \infty) \subset \mathbb{R}$ is a continuous function on $[0, \tau_i] \subset \mathbb{N}$. τ_i denotes the horizontal axis which the object of interest should not cross due to kinematic constraints. λ governs the width of the linear part of the scoring function.

Considering the problem of tracking both hands of a signer, soft spatial pruning is used to partition the video frame in a region for the dominant hand and one for the non-dominant hand by choosing τ_i accordingly. Adjusting Equation (20) for the non-dominant hand is straight-forward.

5.1.4. Face Suppression:

In the context of tracking a person's hand while the person is signing the issue of the tracker getting stuck at the person's face arises. This happens primarily when the tracker utilises non-skincolour suppression to increase tracking performance. To reduce the probability of the tracker getting stuck at the signer's face we use a spring-like function centered on the face position to reduce the score of the hand tracker in the face region.

$$q(l_t, l_{t-1}, X_{t-1}^t) = \sum_{l \in Q_t} 1 - f_{fp}(X_t)[l] \qquad (21)$$

where $f_{fp}(X) : \mathbb{R}^{I \times J} \mapsto [0, 1]^{I \times J} \subset \mathbb{R}^{I \times J}$ denotes the face probability image of X which is obtained via probabilistic face detectors [67, 59] or the face tracker described in Section 5.2.

5.1.5. Other Hand Suppression:

An issue unique to hand tracking for humans shown in frontal pose in a video frame is that both hands are nearly indistinguishable from an algorithmic point of view. Using a spring-like function, we can utilise the tracked position of e.g. the right hand to reduce the search space when tracking the e.g. left hand

since it is unlikely that both hands overlap for a prolonged period of time when a person is signing.

$$\lambda(l_t, l_{t-1}|l_t^{\Xi} = (u,v)) = -\sum_{l=(i,j)\in Q_t\cap Q_t^{\Xi}} 1.0 - \frac{\sqrt{(u-i)^2+(v-j)^2}}{\min(w_{\Xi}, h_{\Xi})} \tag{22}$$

$$q(l_t, l_{t-1}, X_{t-1}^t|l_1^{T,\Xi})) = \begin{cases} \lambda(l_t, l_{t-1}|l_t^{\Xi}) : \lambda(l_t, l_{t-1}|l_t^{\Xi}) \geq 0.0 \\ 0.0 \quad : \quad \text{otherwise} \end{cases} \tag{23}$$

where l_t^{Ξ} is the hand location detected in the first tracking pass with tracking window Q_t^{Ξ} of size $w_{\Xi} \cdot h_{\Xi}$.

Using other hand suppression, DPT processes the video sequence twice by first tracking the dominant hand and then using the resulting tracking locations as an additional information source in tracking the non-dominant hand.

5.2. Tracking Facial Landmarks

Active Appearance Models (AAMs) were introduced by Edwards et al. [18] in 1998 and notably reformulated by Matthews et al. [46] in 2004. They attempt to recover an object's shape $\mathbf{s}$ by generatively fitting a deformable shape model to the image data. $\mathbf{s}$ is defined as a vector of v 2-dimensional landmark points representing a meaningful part of the object, such as an eye corner in the human face:

$$\mathbf{s} = (x_1, y_1, x_2, y_2, \ldots, x_v, y_v)^{\intercal} \tag{24}$$

AAMs model shape deformation using a point density model (PDM), which is a parametric linear subspace model learned statistically by principal components analysis (PCA) on a set of training shape examples, such as shown in Figure 7. Thereby, any shape $\mathbf{s}$ of the deformable object can be expressed as a linear combination of a base shape $\mathbf{s}_0$ and n shape vectors $\mathbf{s}_i$:

$$\mathbf{s} = \mathbf{s}_0 + \sum_{i=1}^{n} p_i \mathbf{s}_i \tag{25}$$

AAMs propose to model the coupling between the PDM and the image data, i.e. the predictions on the PDM's landmarks locations given a target image, using an appearance model representing the object. This is also a parametric linear subspace model, obtained by applying PCA to shape-normalised training example images, which involves the warping to a reference frame. This is typically done by piece-wise affine warping functions defined between each example shape and the base shape $\mathbf{s}_0$. The generative appearance model is then used to express any object's appearance $A(\mathbf{x})$ as a base appearance $A_0(\mathbf{x})$ plus a linear combination of m appearance images $A_i(\mathbf{x})$:

$$A(\mathbf{x}) = A_0(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i A_i(\mathbf{x}) \qquad \forall \mathbf{x} \in \mathcal{R}(\mathbf{s}_0) \tag{26}$$

where $\mathcal{R}(\mathbf{s}_0)$ denotes the set of pixel locations within the region defined by the base shape $\mathbf{s}_0$, i.e. the reference frame for the object's appearance.

Figure 7: Visualisation of facial annotations

Given these two generative models and following the independent AAM formulation proposed in [46], registration can be seen as an image matching problem between the synthetic model image and the shape-normalised target image; the fitting goal can therefore be expressed as finding the parameters $\mathbf{p} = (p_1, p_2, \ldots, p_n)^{\intercal}$ and $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_m)^{\intercal}$ that minimise

$$\sum_{\mathbf{x}\in\mathcal{R}(\mathbf{s}_0)} \left[A_0(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i A_i(\mathbf{x}) - I(\mathbf{W}(\mathbf{x};\mathbf{p})) \right]^2 \tag{27}$$

where I is the target image and $\mathbf{W}(\mathbf{x};\mathbf{p})$ is a piece-wise affine warping function which projects a pixel location $\mathbf{x}$ from the reference frame to the target image frame, depending on the PDM's parameters $\mathbf{p}$. The minimisation of this quantity is non-linear in the parameters $\mathbf{p}$ and must be solved iteratively by linear approximation, typically using the Gauss-Newton algorithm.

AAM variants mostly differ in the way they parametrise the linear approximation to derive the parameters' update equation. We chose to use the efficient version of the simultaneous inverse-compositional AAM (SICAAM) proposed in [28], which is more robust to large variations in shape and appearance. Moreover, we follow [72] in order to cope with large off-plane head rotations, which are also common in sign language and can lead a 2D AAM to failure. Thus, in the present work a 3D PDM is estimated using a non-rigid structure-from-motion algorithm on the training shapes, and is then involved in the optimisation process which incorporates a regularisation term encouraging the 2D shape controlled by the 2D PDM to be a valid projection of the 3D PDM. Similar to the 2D PDM, the 3D PDM expresses any 3D shape $\mathbf{S}$ as a 3D base shape $\mathbf{S}_0$ plus a linear combination of $\bar{n}$ 3D shape vectors $\mathbf{S}_i$:

$$\mathbf{S} = \mathbf{S}_0 + \sum_{i=1}^{\bar{n}} \bar{p}_i \mathbf{S}_i \tag{28}$$

Note that the 3D PDM is also involved in the calculation of the high-level facial features described in Section 6.

5.3. Tracking Experiments

In SLs, the dominant hand of a signer carries more information than the non-dominant hand. In our datasets, the majority of signers is right hand dominant. Hand-patch features based on tracked locations of the hands allow to put more emphasis on the signs themselves. Hand and face locations have

been tracked using the scoring functions detailed previously on lighting and contrast normalised video frames (unless a scoring function requires probability images). In the following experiments, for RWTH-PHOENIX-Weather, hand-patches of size 50×70 have been cropped around the tracked position where 50×70 pixel is the average size of an hand in the RWTH-PHOENIX-Weather database. Three consecutive hand-patches are concatenated and each colour channel (red, green and blue) is reduced to it's 70 most discriminative components. For the SIGNUM database, gray-scale hand-patches of size 30×30 have been cropped centred on the tracked location. A temporal context of ± 1 frame is applied before reducing the feature to its 200 most discriminative components. To gauge if tracking is necessary for ASLR in contrast to taking the whole video frame as feature, we down-scale the original full video frame from 210×260 to 53×65 for the RWTH-PHOENIX-Weather corpus (see Table 1) and apply the same feature extraction pipeline as for the tracking based features (channel-wise PCA to 70 dimensions). For the SIGNUM database, gray-scale full images down-scaled from 575×575 (cropped from the original resolution 578×776) to 32×32 are used. Baseline results for both databases are reported in Table 4.

In the face related experiments the face patches represent down-sampled face crops (RWTH-PHOENIX-Weather 22x35 pixels and SIGNUM 32x32 pixels), 3 frames temporally concatenated and PCA reduced to 200 dimensions. The AAM-face features represent 15 shape and 15 texture AAM coefficients originating from the AAM described in Section 5.2. For RWTH-PHOENIX-Weather 5 temporal frames have been concatenated and PCA reduced to 210 dimensions, for SIGNUM 9 frames have been reduced to 200 dimensions.

Table 4 clearly shows that full video image features are outperformed by hand-patch features. In case of RWTH-PHOENIX-Weather the overall result is improved by 22% WER absolute and on the SIGNUM database the result is improved by 12% WER where WER measures the the required numbers of deletions, insertions and substitution operations to transform the recognised word sequence into the correct word sequences. Results in Table 4 show that full image features contain too much variation to be effectively handled by the current prototype. Furthermore, the necessary down-scaling of the full images reduces the information contained in the images and renders the identification of individual signs difficult.

	PHOENIX		SIGNUM	
	del/ins	WER	del/ins	WER
Full Frame	25.3/5.5	77.0	7.2 / 2.3	28.2
Tracked Hand-patch	20.3/5.7	55.0	2.2 / 3.2	16.0
Face-patch	53.6/1.2	95.1	37.2 / 2.6	83.1
Tracked Face-points	23.4/3.5	62.6	19.5/2.7	56.2

Table 4: Impact of using tracking. First two lines present hand tracking instead of full frames, second two lines present AAM landmark tracking instead of face patches for RWTH-PHOENIX-Weather and SIGNUM single signer subsets. Error rates in [%].

Tracking based features like hand-patches suffer from error propagation in the overall pipeline. If the tracked position of the object deviates too far from the real position, the corresponding hand-patch either does not match the trained model of the recogniser or introduces strong variation in the training process, severely limiting the model's ability to generalise over unseen data. Table 5 shows the influence of the tracking performance measured in TrackEr($\tau = 20$) (Equation 29) against the recognition performance measured in WER.

TrackEr is the tracking error, measured as the average discrepancy between an annotated ground truth object location and the location found by automatic tracking:

$$\text{TrackEr} = \frac{1}{T}\sum_{t=1}^{T}\delta_\tau(l_t, \hat{l}_t), \quad \delta_\tau(l,m) := \begin{cases} 0 & \|l-m\| < \tau \\ 1 & \text{otherwise} \end{cases} \tag{29}$$

where l and m are two 2D coordinates in a video frame and τ is the deviation threshold.

To measure the impact of the tracking performance, the parameters of the DPT framework have been adjusted to reach a specific TrackEr. For all experiments the parameters of the proposed ASLR system have been kept fixed for training and testing being optimised of the best result obtained at TrackEr of 11.6. For each TrackEr level the whole system has been retrained using features extracted based on a tracking achieving the performance level in question. The first row of the table showing a TrackEr of 0 is based on ground-truth annotation for the whole training and testing set for the RWTH-PHOENIX-Weather signer specific subset.

TrackEr	del/ins	WER
0	13.1/7.6	48.3
11.6	20.3/5.7	55.0
20.0	13.0/6.5	56.2
25.0	22.6/6.2	63.0
30.0	24.8/5.5	68.4
40.0	24.0/9.4	76.6

Table 5: Influence of tracking performance measured in TrackEr at $\tau = 20$ on recognition for RWTH-PHOENIX-Weather single signer subset. Error rates in [%].

Results depicted in Table 5 show a clear connection between TrackEr and WER for continuous sign language recognition. The higher the TrackEr gets the worse the recognition result is. An improvement of the used tracking method in order to achieve better and more consistent tracking results will have a positive impact on sign language recognition performance. The result of 48.3% of the perfect tracking result indicates the limit of performance reachable with hand-patch features alone and should not be interpreted as the overall gain possible with improved tracking on the used database. It stands to reason that features capturing the overall shape of hand while being invariant to the majority of noise present in the video frame (e.g. motion blur) will benefit from better tracking in the same manner as hand-patch features.

6. Features

SLs convey meaning by several information streams in parallel: besides the hand shape, orientation and position, the upper body pose, and also facial expression, such as mouthing, eye brows and eye gaze are important. The goal of this section is to assess the impact of each feature adding up to a state-of-the-art recognition system.

Instead of PCA-reduced hand patches as in Section 5.3, we employ **HOG-3D Features** [45], which explicitly capture the edges of the hands spatially and also temporally and are therefore much more robust against illumination differences. The HOG-3D histograms are computed using a temporal context of 7 frames. A large part of the information in SL is contained in the temporal sequence. We thus add more temporal context by stacking together ±4 video frames for SIGNUM and ±2 frames for PHOENIX. Subsequently, we perform a PCA reduction to 200 and 210 dimensions for SIGNUM and PHOENIX respectively. HOG-3D are used for ASLR in this work and not for tracking.

Trajectories with Position constitute a second important manual feature. The trajectory motion is understood as a main direction and a shape. Given the hand position $u_t = (x, y)$ at a time t, the velocity vector $m_t = u_t - u_{t-\delta}$ points in the direction of the movement. However, a more robust method is used in this work. It is based on the estimation of the covariance matrix within a time window $2\delta + 1$ around time t, as shown in Equation (30),

$$\Sigma_t = \frac{1}{2\delta + 1} \sum_{t'=t-\delta}^{t+\delta} (u_{t'} - \mu_t)(u_{t'} - \mu_t)^T \tag{30}$$

with $\mu_t = \frac{1}{2\delta+1} \sum_{t'=t-\delta}^{t+\delta} u_{t'}$.

$$\Sigma_t \cdot v_{t,i} = \lambda_{t,i} \cdot v_{t,i}, i \in \{1, 2\} \tag{31}$$

The eigenvector $v_{t,i}$ with the larger corresponding eigenvalue points towards the direction of highest variance. The eigenvalues $\lambda_{t,i}$ characterise the motion. If both values are similar, it is a curved motion, otherwise a line. In order to capture temporal variation on different levels, the feature vectors are composed of the eigenvalues and main eigenvectors, calculated over the tracked trajectory points of three different temporal windows with $\delta \in \{4, 5, 6\}$ for RWTH-PHOENIX-Weather and $\delta \in \{8, 9, 10\}$ for SIGNUM. Additionally, the position of the dominant hand w.r.t. the signer's nose is added to the feature vector.

In DGS there are one- and two-handed signs. Two-handed signs either have a symmetric or anti-symmetric movement or the non-dominant hand serves as location for the dominant hand. In either case, the relative movement of the hands towards each other is a good indicator for this behaviour. We define **Handedness** features as the eigenvectors and eigenvalues of the movement of both hands relative to each other over multiple time windows of δ video frames. This corresponds to Equation (30), with the relative distance between both hands rather than the hand position. δ has been optimised to be $4, 5, 6$ for RWTH-PHOENIX-Weather and $8, 9, 10$ for SIGNUM.

Semantic description	Related point distances	
mouth vertical openness	{18, 24, 25, 21}	{18, 26, 27, 21}
mouth horiz. openness	{18}	{21}
lower lip to chin distance	{26, 27}	{32, 33}
upper lip to nose distance	{16, 15, 17}	{18, 24, 25, 21}
left eyebrow state	{0, 1, 2}	{6, 8}
right eyebrow state	{3, 4, 5}	{10, 12}
gap between eyebrows	{2}	{3}

Table 6: High-level facial features and the related lower-level points (refer to Figure 7 for the landmark indexes)

High-Level Face Features consist of seven continuous distance measurements across landmarks around the signer's face, as described by Table 6. They correspond to key locations on the cheeks and chin outlines, the nose ridge and nose base, the eyelids and eye corners, the eyebrow outlines and the lip and mouth corners. These measurements are based on the tracked landmarks as described in Section 5.2 and are expected to capture the information encoded in the non-manual parameters used in SL. To estimate the high-level mouth distances we project the registered shape and remove its global translation and rotation by means of the 3D PDM (refer to Section 5.2). Then, for each point subset given in Fig. 6, we estimate the corresponding local area-based measurement and normalise it between 0 and 1 according to the minimum and maximum values obtained during training. See Figure 8 for a visualisation of these features.

Temporal Derivatives constitute a well known feature in ASR. They capture additional context being the temporal change of the features they are applied to. The derivatives denoted in the following as Δ for the first derivatives and $\Delta\Delta$ for the second derivatives are calculated around the current frame X_t as:

$$\Delta(X_t) = X_{t+1} - X_{t-1} \tag{32}$$

$$\Delta\Delta(X_t) = X_{t+2} - 2X_t + X_{t-2} \tag{33}$$

As discussed previously in this section, the system uses HOG-3D as well as other trajectory-based features to describe the different information channels of SL. Handedness, hand trajectory as well as high-level face features implicitly contain temporal derivatives and are therefore not used to extract additional delta features. HOG-3D features on the other hand, contain only quantised temporal information but do not encode the speed of the temporal change. Thus, Δ and $\Delta\Delta$ features are only extracted from the HOG-3D features.

6.1. Discussion of Experimental Feature Impact

In Table 7 we present results using the advanced features introduced in the previous section. The features' impact is evaluated on the signer dependent subsets of both corpora presented in Section 3. We can clearly see the positive impact of each added feature, which all help to reduce the WER from 43.5% with only HOG-3D to 38.6% with added dominant (right) hand trajectory and position, handedness, high-level face features and first and second order derivatives on the PHOENIX corpus.

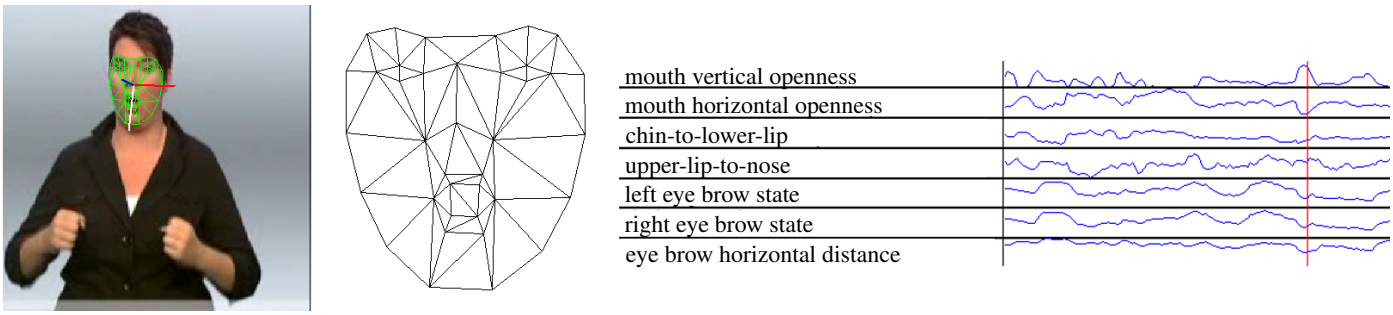

Figure 8: High-level feature extraction, left: the grid of fitted AAM points, center: rotated and normalised AAM points, right: high-level feature values over time

On SIGNUM the features reduce the error rate from 12.5% to 10.0% outperforming previous approaches (see Table 3). However, the derivatives do not improve the recognition. This may be explained by the controlled signing speed in SIGNUM as described in Section 3. Due to the staged nature of the corpus the artificial signing can be observed as rhythmic with consistent speed. Thus, features capturing this speed do not add additional discriminative information.

Comparing HOG-3D only results to the results obtained by PCA reduced hand-patch features (see Tables 4 and 5), HOG-3D features extracted from tracked locations outperform the PCA hand-patch features extracted from ground-truth annotation. This clearly shows the importance of noise robust features. HOG-3D features consist of spatial and temporal edges focusing on the shape of the hands.

The fact that each of the other individual features improves the recognition is important, as it underlines the multimodality of SL. This corresponds to sign linguists' findings: The information in SL is perceived through manual and non manual channels simultaneously. The question how to best combine the features is still unsolved, a thorough analysis can be found in [22]. Only minor gain from asynchronous modelling is reported there. This is why we preferred a feature combination by stacking in the scope of this work. Stacking features together in a HMM framework bears a difficulty with respect to the feature weight. Each feature dimension adds to the final decision and its scaling w.r.t. to the other dimensions is a detailed, but crucial factor determining possible reduction of WER. This constitutes a big difference to ASR systems, where the set of features is quite standard and lower dimensional. **Feature Preprocessing** plays, thus, a key role in our HMM approach. Therefore, we apply a global variance normalisation per dimension. Adaptive feature transformations (such as constrained maximum likelihood linear regression (CMLLR) presented in Section 8), linear discriminant analysis (LDA) or discriminative feature learning approaches, such as neural networks also help controlling this problem. The employed statistical modelling approach uses the Mahalanobis distance and diagonal covariance matrices (refer to Section 4 for details). This makes it crucial that the features are decorrelated, for which we use PCA.

	PHOENIX		SIGNUM	
Feature	Del / Ins	WER	Del / Ins	WER
HoG-3D	19.5 / 4.9	43.5	2.8 / 2.4	12.5
+ RH Traj / Place	14.6 / 8.2	42.5	3.0 / 1.5	11.9
+Handedness	22.0 / 3.3	39.4	2.1 / 2.1	11.7
+ Highlvl-Face	22.4 / 3.6	39.2	1.7 / 1.7	10.0
+ 50 Δ + 1$\Delta\Delta$	17.7 / 4.9	38.6	1.1 / 2.8	10.2

Table 7: Recognition results on RWTH-PHOENIX-Weather and SIGNUM single signer subsets showing the impact of the features. Error rates in [%].

7. Visual Modelling

The way we model the features crucially influences the system's performance. An often neglected aspect of building statistical models for sign language recognition tasks is how to model non-speech. Modelling non-speech in the context of ASLR, i.e. non-gesture, is a challenging task because it is not clear which parts of the temporal signal are not part of a sign. While in ASR the transition between words may be marked by a decline in speech volume, the transition between signs depends on both enclosing signs. For example consider the case of a sign 'A' ending its manual part with the dominant hand raised to the eye level of the signer and the following sign 'B' starting its manual component at waist height. In this case neither the movement from the ending location of sign 'A' to the starting location of sign 'B' nor the hand's change in shape and orientation are part of either signs. Thus, the transition part called movement epenthesis should not be part of the learned statistical models for signs 'A' and 'B'. If the movement epenthesis spans only a couple of time frames it can be compensated for by HMMs but in 'real-life' video footage this is often not the case.

Besides movement epenthesis effects, signers tend to switch hands while signing causing non-gesture effects for the models trained to recognise the manual aspects of the respective other hands.

In this section we provide insights about the question of how to model sign language on three levels:

- Improving the HMM state-alignment by non-gesture garbage models,
- Length model HMMs to account for length in whole sign models,

- Account for co-articulation effects by adding threshold or transition model.

7.1. Non-Gesture Garbage Modelling

Figure 9 shows the video frame to HMM state alignments for one sentence of the RWTH-PHOENIX-Weather corpus. On the left side without non-gesture models and on the right side using non-gesture models. The time in video frames is depicted on the x-axes with the ground truth glosses including the gloss/sign boundaries overlayed on the axis. The y-axis depicts the states of the individual HMM models. The blue circles represent that the frame in question has been aligned to this HMM state where the colour change in the background illustrates the time period that is aligned to the same state. The red straight line linking the origin of the plot to the top right corner is an optical aid not related to the alignment but illustrates the theoretical ideal alignment.

Comparing the left part to the right part in Figure 9 one can see that the depicted alignment on the right side is closer to the optical aid representing an ideal alignment when the length modelling is perfect. Furthermore, the large 'plateau' areas in the alignment on the left side have vanished from the alignment indicating a better distribution of the data to the learned models. Finally, the non-gesture block on the left side of Figure 9 indicated by the white blank in the ground truth gloss annotation has correctly been assigned to a non-gesture block. Non-gesture blocks for RWTH-PHOENIX-Weather have been inferred from ground truth annotation accounting for hand changes and non-gesture facial expressions, such as mimics.

Applying non-gesture garbage models to a recognition system using HOG-3D and movement trajectory features extracted from ground truth annotation of the hand positions, the recognition system result is improved from 42.1% to 39.8% WER for the signer dependent sub set of RWTH-PHOENIX-Weather. This underlines the importance of modelling non-gesture for ASLR.

The used non-gesture garbage modelling makes use of the fact that a subset of the RWTH-PHOENIX-Weather database is annotated on the sentence and the sign level. This allows to identify temporal gaps between signs and assign a garbage tag to these gaps. Furthermore, annotations include tags for the left vs. right hand, also allowing to add offhand specific tags for modelling. These kind of annotations are not available for the SIGNUM database, preventing the use of he proposed garbage models to the SIGNUM database. Nevertheless, results on SIGNUM are expected to improve, if a similar kind of garbage modelling is applied to it.

7.2. Length Modelling

Length modelling is analysed to account for the variability of sign duration. Models incorporating the average length of signs corresponding to a certain gloss are compared to models having a constant length for all glosses.

In Table 8 the impact of length modelling can be verified. To evaluate the impact of length modelling, it is compared to the baseline system using a standard 3-2 Bakis HMM. The number of states S is determined by half of the median of the total number N of running lengths l belonging to a certain gloss i. Only if this value is bigger than the shortest length $l_{i,min}$ reduced by 20%, then the number of states has to be adapted to ensure all training samples can reach the end state of the model. Refer to Equation 34 for a mathematical description.

$$S_i = \min\left(0.8 \cdot l_{i,min}, \frac{1}{2} \cdot \tilde{l}_i\right) \tag{34}$$

$$\tilde{l}_i := \begin{cases} l_{i,\lceil \frac{N}{2} \rceil} & n \text{ odd} \\ \frac{1}{2}\left(l_{i,\frac{N}{2}} + l_{i,\lceil \frac{N}{2} \rceil}\right) & n \text{ even} \end{cases}, \qquad l_{i,n} \leq l_{i,n+1} \; \forall \; n \tag{35}$$

Referring to Table 8, on the single signer subset of RWTH-PHOENIX-Weather an absolute gain of more than 5% can be observed by using length modelling and basic PCA-reduced hand-patch features for the tracked dominant hand. SIGNUM does not show this behaviour, here the WER increases from 16.0% to 17.5% with length modelling. Similarly, for RWTH-PHOENIX-Weather multi-signer, the error rate increases with length modelling. Within the single signer setup of RWTH-PHOENIX-Weather gloss lengths have been manually annotated (all gloss boundaries throughout the corpus are annotated), whereas within SIGNUM and RWTH-PHOENIX-Weather 2014 multi-signer they have been estimated from the recogniser state alignment. This fact is likely to account for the difference. It can be concluded that an accurate length modelling helps to improve recognition performance. Similar observations have been made in the early times of ASR before the transition from whole-word models to phoneme-based models.

7.3. Universal Transition Model

SLs, particularly when looking at natural signing, contain a lot of variability. The preceding and following signs influence the starting position and the execution of the current sign. Also, noises in the recording material and errors in the manual annotation render the learning of clean and accurate models difficult.

In this subsection, we evaluate a method to better cope with noise in the data, particularly originating from movement epenthesis in SL. Similar to Yang [74], we implemented a threshold model to account for outliers in the data, whenever they do not match the gloss model well enough. During Viterbi training, when searching for the optimal alignment between features x_1^T and the HMM state sequence s_1^N, we allow a threshold garbage model to account for the input features by upper-bounding the log-likelihood score to the threshold value λ. However, the model showed no significant improvement in our large scale recognition pipeline.

Improvement could be achieved by learning a universal transition model, being a background HMM with a single state and separately optimised transition probabilities. During Viterbi training, this model has an empty pronunciation and can thus be inserted in between two signs to account for the movement epenthesis. As can be seen in Table 9, an improvement by 1.9% WER (58.3% $\rightarrow$ 56.4%) on the development set and 1.4% on the PHOENIX MS test set could be achieved.

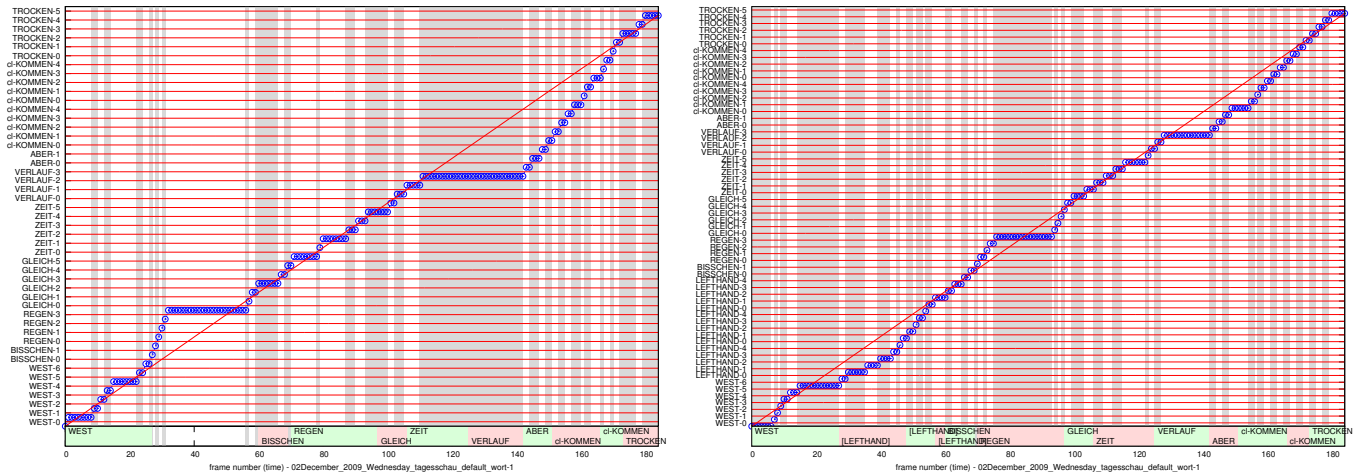

Figure 9: HMM state alignment visualisation for RWTH-PHOENIX-Weather single signer sub set using non-gesture garbage modelling for the same video sequence. Left: without non-gesture modelling, Right: with non-gesture modelling. The alignment on the right is superior to the left as it is close to the theoretical optimal alignment (red line from origin to upper right corner) and the number of state 'plateau' areas is reduced. See Section 7.1 for detailed discussion.

	PHOENIX manual		PHOENIX MS auto				SIGNUM auto	
	Test		Dev		Test		Test	
Length modelling	Del / Ins	WER	Del / Ins	WER	Del / Ins	WER	Del / Ins	WER
Yes	20.3 / 5.7	55.0	25.7 / 4.4	59.2	23.4 / 4.6	56.7	3.1/2.8	17.5
No	22.6 / 6.0	60.8	23.4 / 4.4	58.3	22.0 / 4.3	55.5	2.2 / 3.2	16.0

Table 8: Influence of Length Modelling on Recognition Performance for RWTH-PHOENIX-Weather both single and multi signer and SIGNUM single signer subset. Both single signer setups employ hand-patch features, while the multi signer setup employs HOG-3D + trajectory + handedness + highlevel face. 'manual' refers to manually annotated sign boundaries, whereas 'auto' denotes automatic length estimation. Error rates in [%].

8. Generalising to Multiple Signers

ASLR is currently transitioning from signer specific systems trained on a single signer to systems capable of dealing with multiple signers. A requirement for building systems for multiple signers or even signer independent systems is a suitable amount of annotated training data for each of the signers in question.

Considering the multi-signer setups of SIGNUM and RWTH-PHOENIX-Weather (see Section 3), they benefit from the increased amount of training data. However, on the other hand, the varying signing styles of the different signers contribute to a higher inter-signer variability in addition to the strong intra-signer variability present also in the single signer corpora.

In the area of large vocabulary speech recognition, a common technique to address intra-signer variability is speaker adaptive training and speaker adaptation. Speaker adaptive training is a two pass training procedure in which the model trained in the first pass is adapted to the data. Constrained maximum likelihood linear regression (CMLLR) [25] allows to adapt the features to the learned model from the first pass and to re-train the whole system using the adapted features. To learn the necessary linear transformation one needs to know which speaker/signer performed which sentence and a time alignment of the individual feature frames to the states of the learned HMM. Both information is available for RWTH-PHOENIX-Weather and SIGNUM.

For this experiment, the feature evaluation from Table 7 has been re-evaluated using the multi-signer corpora. Table 10 subsumes the results for RWTH-PHOENIX-Weather and SIGNUM multi-signer for adding each additional feature, in a standard approach (column 'no CMLLR') and with speaker adaptive training and CMLLR (column 'CMLLR'). All parameters have been optimised on the development set only. HOG-3D, movement trajectory and handedness features have been extracted using tracking results obtained via the DPT system (Section 5). All features have been variance normalised and individually PCA reduced to maintain 99.5% of their variance. This decorrelation has been seen to be crucial in order to successfully apply CMLLR.

Standard no-CMMLR results obtained on RWTH-PHOENIX-Weather multi-signer improve from 58.1% WER with the HOG-3D features to 55.6% with added trajectory, handedness and highlevel face features on the test set. Reported single signer results (refer to Table 7) have been more than 20% absolute better. However, recognition results on both sets are not comparable as a) annotated ground truth object locations are used for signer-dependent experiments in Table 7 instead of automatic DPT tracking, and b) length modelling is done using annotated gloss lengths for signer dependent experiments while they are estimated from training state alignments in this section. While these error sources are difficult to avoid, it gives a realistic estimate of the system performance when deployed in a TV studio (e.g. providing automatic subtitling). Furthermore,

	PHOENIX MS				SIGNUM MS	
	Development		Test		Test	
Setup	Del / Ins	WER	Del / Ins	WER	Del / Ins	WER
no transition model	23.4 / 4.4	58.3	22.0 / 4.3	55.5	12.7 / 16.2	62.0
universal transition model	22.4 / 4.6	56.4	22.3 / 4.3	54.1	5.5 / 2.2	16.5

Table 9: Recognition results RWTH-PHOENIX-Weather multi signer: HOG-3D + trajectory + handedness + high-level face - 4-gram for PHOENIX and 3-gram LM for SIGNUM. Error rates in [%].

	no CMLLR				CMLLR			
	Development		Test		Development		Test	
Setup	Del / Ins	WER	Del / Ins	WER	Del / Ins	WER	Del / Ins	WER
Righthand HoG3D	25.8 / 4.2	60.9	23.2 / 4.1	58.1	24.1 / 4.1	58.9	22.0 / 4.7	57.3
+ RH Traj / Place	30.2 / 2.6	60.9	27.6 / 3.0	58.8	24.4 / 3.8	60.6	22.9 / 4.6	58.6
+ handedness	25.7 / 3.2	58.6	24.1 / 4.0	56.9	24.9 / 3.4	58.3	22.3 / 5.1	57.4
+ highLvl Face	23.6 / 4.0	57.3	23.1 / 4.4	55.6	21.8 / 3.9	55.0	20.3 / 4.5	53.0
+ 50Δ +1ΔΔ	24.1 / 4.0	57.5	23.5 / 4.9	56.1	21.5 / 4.6	57.4	20.3 / 5.5	55.6

Table 10: Recognition results RWTH-PHOENIX-Weather multi signer with 4-gram lm. Error rates in [%].

	no CMLLR	
	Test	
Setup	Del / Ins	WER
Righthand HoG3D	6.0 / 3.1	19.1
+ RH Traj / Place	5.5 / 2.8	18.8
+ handedness	6.4 / 2.2	18.3
+ highLvl Face	5.0 / 2.7	16.4
+ 50Δ +1ΔΔ	5.8 / 2.4	17.3

Table 11: Recognition results on SIGNUM multi-signer with 3-gram lm. Error rates in [%].

the signers present in RWTH-PHOENIX-Weather are trained hearing interpreters hailing from different areas of Germany. They differ in pronunciation and show dialectic differences.

The overall trend found in the previous section with regard to the impact of individual features is preserved with the exception of adding the derivatives. Both the development and the test set show similar performance while the results on the test set are slightly better than the results obtained on the development set showing that the learned models are able to generalise to unseen data. This also applies to comparable 'no CMLLR' results on the SIGNUM multi-signer set as shown in Table 11. Here, the WER decreases from 19.1% to 16.4% on the 25 signer corpus. This underlines the fact that the RWTH-PHOENIX-Weather database represents a bigger challenge than the SIGNUM database. This is because the data in the RWTH-PHOENIX-Weather database constitutes natural signing, which has not been controlled particularly for research purposes.

The CMLLR results shown in Table 10 denote the improvement achieved by this linear speaker-based transformation. The lowest error rate with HOG-3D, trajectory, handedness and high-level face features reaches 53.0% on the PHOENIX MS test set, while being 2.6% better than without CMLLR. Note that the speakers have been manually identified during recognition. Apparently, CMLLR doesn't work as well with the trajectory and handedness features. The added trajectory features show only an improvement of 0.2%, reaching a higher WER than the HOG-3D alone and adding handedness increases the error by 0.5% on the test set after the transformation. However, 55.0% on the dev set and 53.0% on the test set represent the best results published so far.

Figure 10 shows the WERs per signer of the best RWTH-PHOENIX-Weather multi-signer setup (being 55.6% vs. 53.0%). Note that the error rate decreases in all cases but for Signers #2 and #6. The reason is that those two signers occur seldomly in the training set (< 1%) (see Figure 4) and thus CMLLR transformation matrices cannot be reliably estimated.

On the SIGNUM data set CMLLR improved the HOG-3D result from 19.1% (see Table 11) to 18.6%.

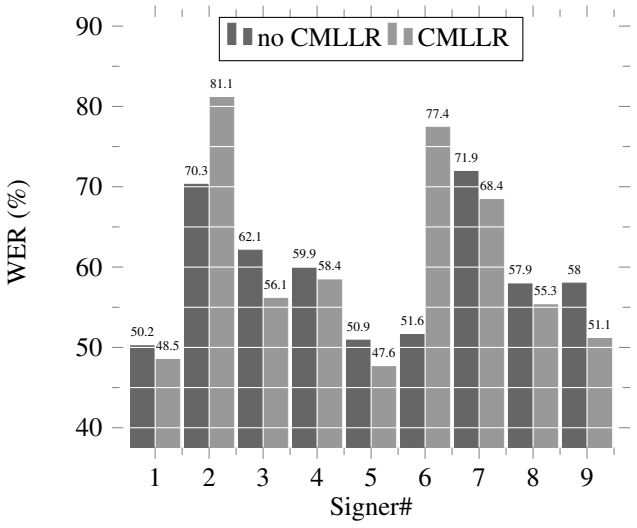

Figure 10: WER per signer, comparing PHOENIX MS test set 'no CMLLR' (55.6%) and 'CMLLR' (53.0%) results. Employed features are HOG-3D + traj/place + handedness + highlevel face with a 4-gram lm.

9. Language Modelling

ASLR systems employ a statistical language model (LM) in addition to the HMM-based visual models as a knowledge source. The LM models the probability of a sign or word occurring in the context of other signs and is learned from text data. In ASR the LM is learned from millions of running sentences in the target language and domain of the recognition system. Since SLs are purely visual languages without a normed or even agreed writing system, the LM can only be learned from the annotated trained data of the corpus under consideration. The low amount of training data for the LM reduces the power of the language model during ASLR because the majority of glosses will only be seen once or twice in any given context.

9.1. Class Language Models

Statistical language modelling of SLs suffers from the low amount of available textual training data, and high sparsity in the sense of a low frequency of the vocabulary entries being seen in any given SL sentence. One way to improve the consistency and frequency in which a specific gloss is seen in context of other glosses is to introduce gloss classes. During training of the LM, the gloss in question is replaced by its class effectively pooling all occurrences of class members into one context n-gram. For example, the gloss sequences `MORGEN 22 GRAD`, `MORGEN 4 GRAD`, and `MORGEN 12 GRAD` result in different 3-grams with a support of one each. Conversely, using a class for numbers the three gloss sequences transform into `MORGEN <number> GRAD` leading to one 3-gram with a support of three.

These classes can be inferred automatically from training data by clustering glosses according to context. Creating classes automatically suffers from the underlying data sparseness problem preventing it's application to current SL corpora.

Another way to define classes is by performing a manual analysis of the recognition errors of any given recognition system. In case of the system setup described in Section 7.1 achieving 39.4% (HOG-3D + RH Traj/Place + Handedness), 3.8% absolute WER of all recognition errors are due to wrongly recognised numbers. Numbers occur in the description of temperatures, dates and altitudes. Furthermore, 2.2% absolute of all errors can be attributed to falsely recognised orientation descriptions such as `NORD` or `NORDWEST` (north and north-west).

Table 12 shows the impact of augmenting a 3-gram LM using Kneser-Ney discounting by classes for numbers and orientations for the single signer subset of RWTH-PHOENIX-Weather. The domain of weather forecasts featured in the RWTH-PHOENIX-Weather corpus is structured and lends itself well to the class LM approach.

Adding classes to the LM improves in both cases the perplexity of the LM. The perplexity, as defined in Equation (36), is an inverse probability between how many classes the model chooses on average to hypothesise every word of a text of length N.

$$\text{Perplexity} = p(w_1^N)^{-\frac{1}{N}} = \left[\prod_{n=1}^{N}(p(w_n|h_n)\right]^{-\frac{1}{N}} \tag{36}$$

Class	Perplexity	Del / Ins	WER
None	34.9	22.0/3.3	39.4
Orientations	31.2	18.1/5.3	39.2
Numbers	29.3	19.3/4.1	38.8
+ Orientations	25.7	16.2/6.2	**38.6**

Table 12: Effect adding classes to a 3-gram LM on ASLR performance on the single signer subset of the RWTH-PHOENIX-Weather. Perplexity measured on the test set for a 3-gram. Error rates in [%].

Adding only numbers or orientations to the LM improves the LM perplexity by more than 10% relative leading to improvements in the ASLR performance. Adding both classes to the LM improves the perplexity by more than 25% relative leading to an overall recognition improvement of 1.2%. Applying the class LM (numbers and orientations) on top of the proposed system using high-level face features, 34.3% WER (20.5% deletions / 1.8% insertions) is achieved. In case of added temporal derivatives no further gain is observed.

Repeating the same process for the multi-signer setup of RWTH-PHOENIX-Weather the LM perplexity is again reduced from 47.9 without gloss classes to 38.2 using a class for numbers and finally to 33.8 by using both numbers and orientations as classes. In contrast to the signer dependent subset, no improvement in recognition performance is observed for the multi-signer setup. Inspecting the errors made by the system with and without class LM it is observable that while the right class (e.g. orientations) from LM perspective is predicted by the ASLR system, the visual model proposes the wrong sign (e.g. NORTH instead of SOUTH). Thus, potential improvements by the class LM are obscured by the lower discriminative power of the visual model.

Class LMs are not considered for the SIGNUM database because the of the artificial nature of the sentence construction.

10. Summary and Conclusion

In this paper we have shown our recent advances in system design for ASLR. We evaluated our approach on two large publicly available continuous sign language data sets representing lab-data (SIGNUM database: 25 signer, 455 sign vocabulary, 19k sentence) and unconstrained 'real-life' sign language (RWTH-PHOENIX-Weather database 9 signer, 1081 sign vocabulary, 7k sentences) reflecting the community's moving from artificial lab-generated data to 'real-life' data. Compared to the current best published results, we are able to improve recognition on lab-data from 10.7% WER to 10.0% for a single signer and from 21.4% down to 16.4% for a multi-signer setup. On the challenging 'real-life' data set we improve the previously best known result of 38.6% WER to 33.4% for a single signer and set a new best result for the PHOENIX multi-signer setup at 53.0% WER.

In numerous detailed ASLR experiments targetting features, visual modelling, signer-dependency and language modelling, we show the impact and benefit of

- tracking of the hands and facial landmarks,

- multimodal sign language features describing hand shape, hand position and movement, inter-hand-relation and detailed facial parameters, as well as temporal derivatives,
- tackling ASLR jointly with hand and face features
- non-gesture-models, length modelling and universal transition models,
- CMLLR as strategy to cope with inter-signer variation in multi-signer corpora and
- class language modelling.

To sum up, we showed that the statistical approach works for sign language recognition and that the results remain consistent with what is expected, even on larger corpora of continuous 'real-life' signing. Applying techniques from speech recognition is useful, as long as the particularities of sign language are being taken care of. We present guidelines to open up the field for newcomers who can benefit from the insights presented here jointly with public access to our large-vocabulary continuous sign language corpus RWTH-PHOENIX-Weather.

11. Future Work

ASLR is still a brittle technology in the sense of the used amounts of data and modelling techniques. Developments in the area of consumer cameras and the advent consumer priced 2.5D camera systems such as Microsoft's Kinect 2.0 system promise to benefit ASLR systems greatly by reducing the complexity of object tracking and feature extraction. Accessing depth information as an additional knowledge source, ASLR can be used to analyse the usage of the signing space in front of a signer allowing to improve recognition performance.

The increase in sign language corpora size over the last years makes findings for ASLR more reliable fostering research into multi-signer and even signer independent ASLR. Furthermore, advances in sign language linguistics coupled with bigger corpora will allow to investigate sub-sign units comparable to the idea of phonemes for spoken languages. Such sub-sign units will allow the application of modelling techniques from conventional ASR that are currently not transferable to ASLR.

In terms of features for sign languages, neural networks and especially deep neural networks show promising results in the area of conventional ASR allowing to automatically learn optimal features for a given language.

[1] Aran, O., Burger, T., Caplier, A., Akarun, L., 2009. A belief-based sequential fusion approach for fusing manual signs and non-manual signals. Pattern Recognition 42 (5), 812–822.

[2] Bauer, B., Nie\ssen, S., Hienz, H., Oct. 1999. Towards an automatic sign language translation system. In: Proc. of the International Workshop on Physicality and Tangibility in Interaction: Towards New Paradigms for Interaction Beyond the Desktop. Siena, Italy.

[3] Bowden, R., Windridge, D., Kadir, T., Zisserman, A., Brady, M., 2004. A linguistic feature vector for the visual interpretation of sign language. Computer Vision-ECCV 2004, 390–401.

[4] Braffort, A., Bolot, L., Chételat-Pelé, E., Choisier, A., Delorme, M., Filhol, M., Segouat, J., Verrecchia, C., Badin, F., Devos, N., 2010. Sign language corpora for analysis, processing and evaluation. In: LREC.

[5] Chen, S., Goodman, J., 1998. An empirical study of smoothing techniques for language modeling. Technical Report TR-10-98, Computer Science Group, Harvard University.

[6] Cooper, H., Holt, B., Bowden, R., Jan. 2011. Sign language recognition. In: Moeslund, T. B., Hilton, A., Krüger, V., Sigal, L. (Eds.), Visual Analysis of Humans. Springer London, pp. 539–562.

[7] Cooper, H., Ong, E.-J., Pugeault, N., Bowden, R., 2012. Sign language recognition using sub-units. The Journal of Machine Learning Research 13 (1), 2205–2231.

[8] Cooper, H., Pugeault, N., Bowden, R., Nov. 2011. Reading the signs: A video based sign dictionary. In: Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on. pp. 914–919.

[9] Crasborn, O., Zwitserlood, I., 2008. The corpus NGT: an online corpus for professionals and laymen. In: Construction and exploitation of sign language corpora. 3rd Workshop on the Representation and Processing of Sign Languages. ELDA, Parijs, pp. 44–49.

[10] Cui, Y., Weng, J., 2000. Appearance-based hand sign recognition from intensity image sequences. Computer Vision and Image Understanding 78 (2), 157–176.

[11] Dalal, N., Triggs, B., 2005. Histograms of oriented gradients for human detection. In: Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on. Vol. 1. IEEE, pp. 886–893.

[12] Deng, J., Tsui, H. T., 2002. A two-step approach based on PaHMM for the recognition of ASL. ACCV, Jan.

[13] Dick, T., Zieren, J., Kraiss, K.-F., 2006. Visual hand posture recognition in monocular image sequences. In: Pattern Recognition. Springer, pp. 566–575.

[14] Dreuw, P., Apr. 2012. Probabilistic sequence models for image sequence processing and recognition. Ph.D. thesis, RWTH Aachen University, Aachen, Germany.

[15] Dreuw, P., Deselaers, T., Rybach, D., Keysers, D., , Ney, H., April 2006. Tracking Using Dynamic Programming for Appearance-Based Sign Language Recognition. In: 7th IEEE International Conference on Automatic Face and Gesture Recognition (FG). pp. 293–29.

[16] Dreuw, P., Forster, J., Deselaers, T., Ney, H., Sep. 2008. Efficient approximations to model-based joint tracking and recognition of continuous sign language. In: IEEE International Conference on Automatic Face and Gesture Recognition. Amsterdam, The Netherlands, pp. 1–6.

[17] Dreuw, P., Forster, J., Ney, H., Sep. 2010. Tracking benchmark databases for video-based sign language recognition. In: ECCV International Workshop on Sign, Gesture, and Activity. Crete, Greece.

[18] Edwards, G. J., Taylor, C. J., Cootes, T. F., Jun. 1998. Interpreting face images using active appearance models. In: Proc. International Conference on Automatic Face and Gesture Recognition. IEEE, pp. 300–305.

[19] Efthimiou, E., Fotinea, S.-E., Hanke, T., Glauert, J., Bowden, R., Braffort, A., Collet, C., Maragos, P., Lefebvre-Albaret, F., 2012. Sign language technologies and resources of the dicta-sign project. In: Proc. of the 5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon. LREC. pp. 23–27.

[20] Evangelidis, G., Singh, G., Radu Horaud, P., Sep. 2014. Continuous gesture recognition from articulated poses. In: ChaLearn Looking at People Workshop in conjunction with ECCV 2014 – European Conference on Computer Vision. Zurich, Switzerland.

[21] Forster, J., Koller, O., Oberdörfer, C., Gweth, Y., Ney, H., Aug. 2013. Improving continuous sign language recognition: Speech recognition techniques and system design. In: Workshop on Speech and Language Processing for Assistive Technologies. Grenoble, France, pp. 41–46, satellite Workshop of INTERSPEECH 2013.

[22] Forster, J., Oberdörfer, C., Koller, O., Ney, H., Jun. 2013. Modality combination techniques for continuous sign language recognition. In: Iberian Conference on Pattern Recognition and Image Analysis. Lecture Notes in Computer Science 7887. Springer, Madeira, Portugal, pp. 89–99.

[23] Forster, J., Schmidt, C., Hoyoux, T., Koller, O., Zelle, U., Piater, J., Ney, H., 2012. RWTH-PHOENIX-weather: A large vocabulary sign language recognition and translation corpus. In: International Conference on Language Resources and Evaluation. Istanbul, Turkey, pp. 3785–3789.

[24] Forster, J., Schmidt, C., Koller, O., Bellgardt, M., Ney, H., May 2014. Extensions of the sign language recognition and translation corpus rwth-phoenix-weather. In: International Conference on Language Resources and Evaluation. Reykjavik, Island, pp. 1–6.

[25] Gales, M. J., 1998. Maximum likelihood linear transformations for

HMM-based speech recognition. Computer speech & language 12 (2), 75–98.

[26] Gass, T., Dreuw, P., Ney, H., Aug. 2010. Constrained energy minimisation for matching-based image recognition. In: International Conference on Pattern Recognition. Istanbul, Turkey, pp. 3304–3307.

[27] Gass, T., Pishchulin, L., Dreuw, P., Ney, H., Mar. 2011. Warp that smile on your face: Optimal and smooth deformations for face recognition. In: IEEE International Conference Automatic Face and Gesture Recognition. Santa Barbara, CA, USA, pp. 456–463.

[28] Gross, R., Matthews, I., Baker, S., 2005. Generic vs. person specific active appearance models. Image and Vision Computing 23 (12), 1080–1093.

[29] Gweth, Y., Plahl, C., Ney, H., Jun. 2012. Enhanced continuous sign language recognition using PCA and neural network features. In: CVPR 2012 Workshop on Gesture Recognition. Providence, Rhode Island, USA, pp. 55–60.

[30] Han, J., Awad, G., Sutherland, A., Apr. 2009. Modelling and segmenting subunits for sign language recognition based on hand motion analysis. Pattern Recognition Letters 30 (6), 623–633.

[31] Hanke, T., 2004. HamNoSys - representing sign language data in language resources and language processing contexts. In: LREC Workshop on the Representation and Processing of Sign Languages. Lisbon, Portugal, pp. 1–6.

[32] Horn, B., Schunk, B., 1981. Determing optical flow. Artificial Intelligence 17, 185–203.

[33] Isard, M., Blake, A., August 1998. CONDENSATION – conditional density propagation for visual tracking. International Journal of Computer Vision 29 (1), 5–28.

[34] Kadous, M. W., 1996. Machine recognition of auslan signs using PowerGloves: Towards large-lexicon recognition of sign language. In: Proceedings of the Workshop on the Integration of Gesture in Language and Speech. pp. 165–174.

[35] Kelly, D., McDonald, J., Markham, C., Sep. 2009. Recognizing spatiotemporal gestures and movement epenthesis in sign language. In: Machine Vision and Image Processing Conference, 2009. IMVIP '09. 13th International. IEEE, pp. 145–150.

[36] Kelly, D., McDonald, J., Markham, C., Apr. 2011. 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.

[37] Keysers, D., Deselaers, T., Gollan, C., Ney, H., Aug. 2007. Deformation models for image recognition. PAMI 29 (8), 1422–1435.

[38] Klaser, A., Marszalek, M., 2008. A spatio-temporal descriptor based on 3d-gradients.

[39] Koller, O., Ney, H., Bowden, R., Apr. 2013. 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, pp. 1–6.

[40] Koller, O., Ney, H., Bowden, R., Sep. 2014. Read my lips: Continuous signer independent weakly supervised viseme recognition. In: Proceedings of the 13th European Conference on Computer Vision. Zurich, Switzerland.

[41] Lee, H.-K., Kim, J. H., Oct. 1999. An HMM-based threshold model approach for gesture recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 21 (10), 961–973.

[42] Liang, R.-H., Ouhyoung, M., 1998. A real-time continuous gesture recognition system for sign language. In: Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on. IEEE, pp. 558–567.

[43] Liu, J., Liu, B., Zhang, S., Yang, F., Yang, P., Metaxas, D. N., Neidle, C., 2014. Non-manual grammatical marker recognition based on multi-scale, spatio-temporal analysis of head pose and facial expressions. Image and Vision Computing.

[44] Lowe, D. G., 1999. Object recognition from local scale-invariant features. In: Computer vision, 1999. The proceedings of the seventh IEEE international conference on. Vol. 2. Ieee, pp. 1150–1157.

[45] Mark Dilsizian, Polina Yanovich, Shu Wang, Carol Neidle, Dimitris Metaxas, May 2014. A new framework for sign language recognition based on 3D handshape identification and linguistic modeling. In: Proceedings of 9th Language Resources and Evaluation Conference (LREC 2014). Reykjavík, Iceland.

[46] Matthews, I., Baker, S., 2004. Active appearance models revisited. International Journal of Computer Vision 60 (2), 135–164.

[47] Nayak, S., Duncan, K., Sarkar, S., Loeding, B., Sep. 2012. Finding recurrent patterns from continuous sign language sentences for automated extraction of signs. Journal of Machine Learning Research 13, 2589–2615.

[48] Neidle, C., Vogler, C., 2012. A new web interface to facilitate access to corpora: development of the ASLLRP data access interface. In: Proceedings of the International Conference on Language Resources and Evaluation.

[49] Ney, H., Apr. 1984. The use of a one-stage dynamic programming algorithm for connected word recognition. IEEE Transactions on Speech and Audio Processing 32 (2), 263–271.

[50] Ney, H., Dreuw, P., Gass, T., Pishchulin, L., 2010. Image recognition and 2d warping. Lecture Notes, RWTH Aachen.

[51] Ong, E.-J., Koller, O., Pugeault, N., Bowden, R., Jun. 2014. Sign spotting using hierarchical sequential patterns with temporal intervals. In: IEEE Conference on Computer Vision and Pattern Recognition. Columbus, OH, USA, pp. 1931–1938.

[52] Ong, S. C., Ranganath, S., 2005. Automatic sign language analysis: A survey and the future beyond lexical meaning. IEEE Pattern Analysis and Machine Intelligence 27 (6), 873–891.

[53] Pfister, T., Charles, J., Zisserman, A., 2013. Large-scale learning of sign language by watching TV (using co-occurrences). In: Proceedings of the British machine vision conference. U. K. Leeds.

[54] Pigou, L., Dieleman, S., Kindermans, P.-J., Schrauwen, B., 2014. Sign language recognition using convolutional neural networks. In: European Conference on Computer Vision, Workshop.

[55] Pitsikalis, V., Theodorakis, S., Vogler, C., Maragos, P., Jun. 2011. Advances in phonetics-based sub-unit modeling for transcription alignment and sign language recognition. In: 2011 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 1–6.

[56] Rutkowski, P., \Lacheta, J., Mostowski, P., Filipczak, J., \Lozińska, S., 2013. The corpus of polish sign language (PJM): Methodology, procedures and impact.

[57] Rybach, D., Hahn, S., Lehnen, P., Nolden, D., Sundermeyer, M., Tüske, Z., Wiesler, S., Schlüter, R., Ney, H., Dec. 2011. RASR - the RWTH aachen university open source speech recognition toolkit. In: IEEE Automatic Speech Recognition and Understanding Workshop.

[58] Schmidt, C., Koller, O., Ney, H., Hoyoux, T., Piater, J., 2013. 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.

[59] Schmidt, C., Koller, O., Ney, H., Hoyoux, T., Piater, J., 2013. Using viseme recognition to improve a sign language translation system. In: International Workshop on Spoken Language Translation. Heidelberg, Germany, pp. 197–203.

[60] Starner, T., Weaver, J., Pentland, A., 1998. Real-time american sign language recognition using desk and wearable computer based video. IEEE Pattern Analysis and Machine Intelligence 20 (12), 1371–1375.

[61] Stein, D., Dreuw, P., Ney, H., Morrissey, S., Way, A., Sep. 2007. Hand in hand: Automatic sign language to speech translation. In: Conference on Theoretical and Methodological Issues in Machine Translation. Sk"ovde, Sweden, pp. 214–220.

[62] Stolcke, A., 2002. {SRILM}- an extensible language modeling toolkit. In: Proc. International Conference on Spoken Language Processing (ICSLP). Denver, Colorado.

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

[64] Tamura, S., Kawasaki, S., 1988. Recognition of sign language motion images. Pattern Recognition 21 (4), 343–353.

[65] Theodorakis, S., Katsamanis, A., Maragos, P., 2009. Product-HMMs for automatic sign language recognition. In: Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on. IEEE, pp. 1601–1604.

[66] Tokuda, M., Okumura, M., Jan. 1998. Towards automatic translation from japanese into japanese sign language. In: Mittal, V. O., Yanco, H. A., Aronis, J., Simpson, R. (Eds.), Assistive Technology and Artificial Intelligence. No. 1458 in Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 97–108.

[67] Viola, P., Jones, M., 2004. Robust real-time face detection. International Journal on Computer Vision (IJCV) 57 (2), 137–154.

[68] Vogler, C., Metaxas, D., 1999. 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, pp. 116–122.
[69] von Agris, U., Knorr, M., Kraiss, K.-F., 2008. The significance of facial features for automatic sign language recognition. In: Automatic Face & Gesture Recognition, 2008. FG'08. 8th IEEE International Conference on. IEEE, pp. 1–6.
[70] Waldron, M. B., Kim, S., 1995. Isolated ASL sign recognition system for deaf persons. IEEE Transactions on Rehabilitation Engineering, 261–271.
[71] Wang, C., Chen, X., Gao, W., 2006. Expanding training set for chinese sign language recognition. In: Automatic Face and Gesture Recognition, 2006. FGR 2006. 7th International Conference on. IEEE, pp. 323–328.
[72] Xiao, J., Baker, S., Matthews, I., Kanade, T., 2004. Real-time combined 2D+ 3D active appearance models. In: CVPR (2). pp. 535–542.
[73] Xiujuan Chai, Hanjie Wang, Xilin Chen, 2014. The DEVISIGN large vocabulary of chinese sign language database and baseline evaluations. Tech. rep., Key Lab of Intelligent Information Processing of Chinese Academy of Sciences, 00000.
[74] Yang, R., Sarkar, S., Loeding, B., Mar. 2010. Handling movement epenthesis and hand segmentation ambiguities in continuous sign language recognition using nested dynamic programming. IEEE Transactions on Pattern Analysis and Machine Intelligence 32 (3), 462 –477.
[75] Zafrulla, Z., Brashear, H., Hamilton, H., Starner, T., 2010. A novel approach to american sign language (asl) phrase verification using reversed signing. In: Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on. IEEE, pp. 48–55.
[76] Zahedi, M., Dreuw, P., Rybach, D., Deselaers, T., Ney, H., Sep. 2006. Using geometric features to improve continuous appearance-based sign language recognition. In: British Machine Vision Conference. Vol. 3. Edinburgh, UK, pp. 1019–1028.

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

Jens Forster studied computer science at RWTH Aachen University, Germany, specalising in computer vision and speech recognition. In October 2008, he joined the Human Language and Pattern Recognition Group at RWTH Aachen University as doctoral student researcher where he worked on statistical approaches to automatic sign language recognition. In 2015, he joined Amazon to work on automatic speech recognition. His main research interests include sign language and gesture recognition, multimodal fusion of asynchronous information streams, and automatic speech recognition.

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.