Improving Continuous Sign Language Recognition: Speech Recognition Techniques and System Design

Jens Forster, Oscar Koller, Christian Oberdörfer, Yannick Gweth, Hermann Ney

Human Language Technology and Pattern Recognition Group
Computer Science Department
RWTH Aachen University
Aachen, Germany
{surname}@cs.rwth-aachen.de

Abstract

Automatic sign language recognition (ASLR) is a special case of automatic speech recognition (ASR) and computer vision (CV) and is currently evolving from using artificial lab-generated data to using 'real-life' data. Although ASLR still struggles with feature extraction, it can benefit from techniques developed for ASR. We present a large-vocabulary ASLR system that is able to recognize sentences in continuous sign language and uses features extracted from standard single-view video cameras without using additional equipment. ASR techniques such as the multi-layer-perceptron (MLP) tandem approach, speaker adaptation, pronunciation modelling, and parallel hidden Markov models are investigated. We evaluate the influence of each system component on the recognition performance. On two publicly available large vocabulary databases representing lab-data (25 signer, 455 sign vocabulary, 19k sentence) and unconstrained 'real-life' sign language (1 signer, 266 sign vocabulary, 351 sentences) we can achieve 22.1% respectively 38.6% WER.

Index Terms: Continuous Sign Language Recognition, Large Vocabulary, ASR, Computer Vision, Recognition System

1. Introduction

Sign languages are natural languages that develop in communities of deaf people around the world and vary from region to region. A sign consists of manual and non-manual components that partly occur in parallel but are not perfectly synchronous [1]. Manual components comprise hand configuration, place of articulation, hand movement and hand orientation while non-manual components include body pose and facial expression. ASLR is a subfield of CV and ASR allowing methods of both worlds to be deployed but it also inherits their respective challenges. Large inter-/intra-personal signing variability, strong coarticulation effects, context dependent classifier gestures, no agreed written form or phoneme-like definition in conjunction with partly parallel information streams, high signing speed inducing motion blur, missing features and the need for automatic hand and face tracking make video-based ASLR a notoriously challenging research field.

Although ASLR is starting to tackle 'real-life' data, the majority of work in the community still focusses on the recognition of isolated signs, particularly in the context of gesture recognition. Deng and Tsui [2] and Wang et al. [3] use parallel HMMs to recognize isolated signs in American Sign Language or Chinese Sign Language, respectively, achieving recognition accuracies over 90%. Ong et al. [4] use boosted sequential pattern trees to recognize isolated signs in British sign language (BSL) allowing to combine partly parallel, not perfectly synchronous, automatically mined phoneme-like units in the recognition process. Pitsikalis et al. [5] extract subunit definitions from linguistic annotation in HamNoSys [6], whereas Koller et al. [7] employ an open SignWriting [8] dictionary to produce and align linguistically meaningful subunits to signs in German sign language (GSL).

However, in real tasks ASLR is more likely to face continuous signing, that is what this work focusses on. In this context, Cooper et al. [9] compare boosted sequential pattern trees to HMMs using linguistically inspired subunits and 3D tracking information finding that the trees outperform HMMs for BSL. Forster et al. [10] investigate techniques to combine not perfectly synchronous information streams within an HMM-based ASLR system finding that synchronization just at word boundaries improves the recognition performance. Recognizing a sign language sentence by spotting individual signs has been investigated by several authors [11, 12, 13, 14] reporting promising results. Finally Yang et al. [15] use a nested dynamic programming approach to handle coarticulation movements between signs.

Given the cited work and the works described in the survey on sign language recognition by Ong and Ranganath[16], two approaches to ASLR are observable. On the one hand, ASLR is viewed as a pure CV problem neglecting the natural language processing nature of the task and focussing on developing tailor-made solutions for gestures. However, we believe to be soon able to tackle real-world problems, ASLR should much more be seen as application of ASR, exploiting previous knowledge gained in that area. Following that track, we provide systematically gathered knowledge on how to create a large vocabulary ASLR system for continuous SL evaluating which techniques from ASR are applicable. Specifically, we investigate the impact of CV and ASR techniques on the recognition performance. Among others, the impact of the performance of automatic hand tracking on the recognition performance is investigated. Tackling the question of suitable features for non-rigid objects such as the hands, HoG3D [17] features proposed in the area of action recognition, appearance based features and learned MLP features used in ASR are investigated. Addressing inter signer variability, the technique of automatic signer adaptation is adopted from ASR (speaker adaptation) and tested within our proposed large-vocabulary, HMM-based sign language recognition system. Additionally, techniques to combine

SLPAT 2013, 4th Workshop on Speech and Language Processing for Assistive Technologies, pages 41–46,
Grenoble, France, 21–22 August, 2013.

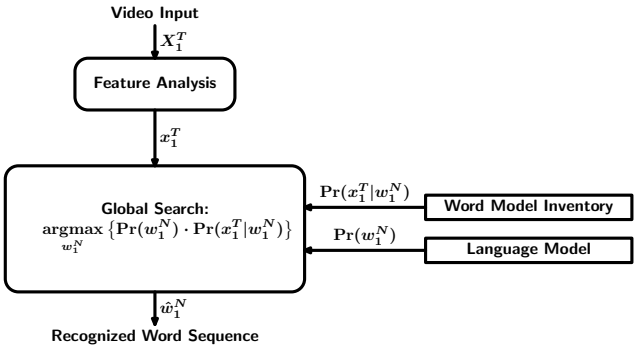

Figure 1: Bayes' decision rule used in ASLR.

partly parallel information streams/modalities are presented and evaluated. The system and its components are tested in the context of continuous ASLR for two publicly available, large-vocabulary databases. One database represents lab-data created for pattern recognition purposes and one database represents 'real-life' data recorded from German public TV. Comparing findings on lab-data and 'real-life' data we investigate which findings on lab-data generalize to 'real-life' scenarios.

2. System overview and features

The ASLR system described here follows the system design proposed in [18] and is based on Bayes's decision rule but differs in several aspects. Specifically, features adapted from action recognition, learned features, a number of techniques to combine different modalities within the system, class-based language models, gap/noise models and signer adaptation techniques for multi-signer data are employed.

The recognition result of the system is the sequence of words that best fits the trained word models and the language model (see Figure 1). One has to note that linguistically this represents a major simplification but the use of gloss annotations (see Section 2.1 for a short definition) is a common practice within the recognition community to deal with the non-availability of a common writing system for sign languages. While linguistically motivated writing notations such as HamNoSys[6] or SignWriting [8] cover information about different modalities used within sign languages, they are still a weak labeling scheme for signs because they do not give an annotation of the movement, facial expression, etc. per time frame. Furthermore, using glosses as target classes and annotation scheme allows for faster annotation of large amounts sign language data which is needed for a automatic statistical recognition approach.

Finally, the proposed recognition system has been tested on the two publicly available databases SIGNUM [19] and RWTH-PHOENIX-Weather (PHOENIX) [20] for GSL which are among the biggest datasets available for continuous ASLR.

2.1. Visual modeling

Albeit the cited work on automatic subunit extraction from sign language videos, it is still unclear how signs can be split into subunits. Furthermore, the majority of sign language corpora including those used in this work (see Section 3) is annotated using *glosses* effectively labeling the meaning of a sign rather than its appearance. Therefore, the proposed system is based on whole-word models. The visual model (VM) of a sign consists of a left-to-right HMM in Bakis topology [21] where each segment of the model (each pair of consecutive states) is modelled by a separate Gaussian mixture model (GMM) with globally pooled covariance matrix. The number of segments per model is estimated from manually annotated sign boundaries on the training data. Due to strong visual pronunciation variances (3 different signs for Sunday exist in GSL), the effect of explicit visual pronunciation modelling is investigated in Section 3.

2.2. Language models

Language models (LMs) play a crucial role in state-of-the-art ASR and ASLR systems. Dreuw et al. [18] showed that the impact of the well-known LM scale on the recognition performance of an ASLR system is in the same order of magnitude as in an ASR system. Therefore, the LM scale is optimized for all experiments presented in this work.

In contrast to ASR where it is possible to obtain language-specific almost arbitrarily large text collections for every language and domain, here the LM can only be trained on the transcribed training data of any given database for ASLR inheriting the problem of singletons and infrequent signs which often make up more than 40% of the available vocabulary of typically 200 to 500 signs. Inspired by the idea of class and topic LMs in ASR [22, 23, 24] and statistical sign language translation [25], we propose to use classes of visually and contextual similar signs within the LM. Class selection is based on the analysis of errors of a baseline system without LM classes. In this work, all LMs are trained using the SRILM toolkit [26] with modified Kneser-Ney discounting with interpolation [27].

2.3. Manual and non-manual features

GSL conveys information through manual and non-manual parameters. Manual parameters comprise both hands' shape, their orientation and position. There are two-handed, as well as single-handed signs. Single-handed signs are usually signed using the dominant hand which in the databases used in this work corresponds to the right hand for all subjects in PHOENIX and all but two in the SIGNUM database.

Manual features: For full coverage of a sign, manual features of both hands are used as well as non-manual features of the face and upper-body. To extract hand features, tracking is performed for both hands separately using a robust tracking algorithm with decision back-tracing originally proposed in [28]. Four different kinds of manual features are extracted. The first one are colored image patches cut out around the tracked positions of the dominant hand with a size of 32×32 Pixel for SIGNUM and 53×65 Pixel for PHOENIX. As second feature, histograms of oriented image gradients in 3D space (HoG3D) [17] are extracted using a non-dense spatio-temporal grid from video volumes of ± 4 cropped patches. Third, the movement trajectory of the right hand is extracted, represented by the position relative to the nose and the eigenvectors and eigenvalues of the movement within a time window of $2\delta + 1$ frames. Fourth, MLP features have been successfully used in ASR [29] and optical character recognition [30]. Here a feed-forward network with one hidden layer of 2000 nodes is trained using frame alignments from a previously trained HMM system as labels and PCA reduced hand patches in case of SIGNUM and HoG3D and trajectory features in case of PHOENIX. The training of the MLP has been performed on the training set of the HMM system. Cross validation is used to adjust the learning rate and to avoid over-fitting.

Non-manual features: Face patches are extracted using the same tracking approach as described above. Furthermore, a position and orientation invariant active appearance model (POIAAM) [31] is fitted to each frame obtaining a 109 dimensional shape descriptor, including shape model parameters, head rotation in space, mouth and eye openings and degrees of eyebrow raise. Finally, every frame of a video sequence is scaled down to 32×32 and 53×65 respectively to get a simple upper body feature as originally proposed in [18].

For all features, temporal context is included by stacking ± 4 video frames for SIGNUM and ± 2 frames for PHOENIX. Since the resulting feature dimension is too high to robustly estimate HMM parameters, PCA is applied. All features but the movement trajectory are reduced to 200 dimensions. In case of the colored hand and face patches PCA is applied to each color channel (red, green, blue) separately, yielding a final feature dimension of 210. The movement trajectory feature itself has only limited discriminative power and is therefore combined with the HoG3D features of the right hand.

2.4. Signer adaptation and modality combination

Sign languages use partly parallel, but not perfectly synchronous information streams/modalities to convey meaning. These modalities must be handled in the recognition process but it is an open question how to incorporate different modalities within such a system. A similar situation exists in audio-visual speech recognition (AVSR) where acoustic features and visual features of the mouth are combined. Following the work in AVSR, we investigated feature combination (concatenation), system combination using (i)ROVER [32] as well as combination between HMMs on state level (synchronous combination) and at word boundaries (asynchronous combination). Experimental results show that the first two types of combination are not effective for current ASLR because either the resulting feature space dimension is too high or the systems make too similar recognition errors [10]. Here, only results for synchronous and asynchronous combination are presented.

Signer adaptation: ASR systems trained on different speakers have to address the speakers' voice and speech patterns to achieve good recognition performance. A common approach is to use speaker adaptive training (SAT) and learn speaker dependent feature transformation matrices using constraint maximum likelihood linear regression (CMLLR). Analogous to ASR, ASLR has to tackle signing styles. Therefore, SAT/CMLLR is evaluated in the context of ASLR for 25 signers.

3. Experimental results

The SIGNUM database [19] contains lab recordings of 25 signers wearing black long-sleeve clothes in front of a dark blue background signing predefined sentences. Videos are recorded at 780×580 Pixel and 30 frames per second (fps). Each signer signs the 603 unique training and 177 testing sentences once, whereas they are signed thrice in the single signer setup. 3.6% of the glosses are out of vocabulary (OOV). Table 1 shows statistics of the single signer setup only. The multi signer setup has the same vocabulary and OOV rate but 15k sentences (92k running glosses) for training and 4.4k sentences (23k running glosses) for testing. If not stated explicitly otherwise, all presented SIGNUM results refer to the single signer setup.

The PHOENIX [20] database contains 'real-life' sign language footage recorded from weather-forecasts aired by the public German TV-station PHOENIX. 'Real-life' is meant from a computer vision point of view, where the signers were not artificially restricted in any sense in their signing (sentence structure, choice of vocabulary, size and intensity of signs, ...) and where the recording conditions have a much larger variance than on other signing corpora (lighting, camera-signer position, ...). The video footage has not been created for pattern recognition purposes or linguistic research. From a linguistic point of view the employed language has to be classified as non-native, as the signer is a hearing interpreter, whose parents are deaf. The videos (210×260 Pixel, 25 fps interlaced) show the interpreter wearing dark clothes in front of an artificial gray gradient background and pose a strong challenge to CV and ASLR due to high signing speed (majority of signs spans less than 10 frames), strong coarticulation effects and more than 30% of the vocabulary being singletons. Statistics of both databases are shown in Table 1.

Table 1: Statistics for SIGNUM single signer and PHOENIX

	SIGNUM		PHOENIX	
	Train	Test	Train	Test
# sentences	1809	531	304	47
# running glosses	11,109	2805	3309	487
vocabulary size	455	-	266	-
# singletons	0	-	90	-
# OOV [%]	-	3.6	-	1.6
perplexity (3-gram)	17.8	72.2	15.9	34.9

The system is trained using maximum likelihood and the EM-algorithm. The number of Gaussian densities and the LM-scale are optimized. For PHOENIX, the system uses 1433 emission distributions with a total of 4k Gaussians and a globally pooled covariance matrix. The same applies to SIGNUM, but the numbers are 1366 emission distributions with 24k Gaussians for single signer and 198k for multi-signer. Recognition uses word-conditioned tree search and Viterbi approximation.

Basic Features: In order to build a well performing ASLR system, the feature selection plays a crucial role. The full video images can be seen as a global descriptor of manual and non-manual parameters and are, thus, a good starting point. As the hands are known to carry the most information in signing, tracked and cut out hand patches have often been preferred [18] over full frames. Comparing both features, hand patches outperform full images on both databases (see Table 2, Row 1).

Model length estimation: In ASR, the HMM model of a word is formed by the linked models of the word's subunit HMMs. Thereby, the typical temporal length of a word is modelled. This approach is not yet possible in ASLR because the definition and extraction of subunits is still an open research question. PHOENIX includes word boundary annotations from which the number of segments for each gloss HMM can been estimated by choosing the median of the lengths minus 20% and adjusting the length in case the adapted median is shorter than the shortest utterance of the gloss. The hand patch baseline presented above uses this approach. Using uniform length for all glosses, the recognition result is 60.8% instead of 55.0% WER. 'Bootstrapping' the initial system alignment using the word boundary ground truth, we achieve 57.5% WER.

No word boundary ground truth is available for SIGNUM. Model length estimation is performed using statistics on the

Table 2: WERs for competing features (Rows 1.-6.), WERs without and with specific techniques (Rows 7.-11.). '+' denotes a synchronous, asynchronous or feature combination. Please see corresponding text parts for explanations. HoG3D uses tracked hand locations. For PHOENIX, in rows 3.-5., manual ground truth annotation has been used instead.

Competing Features		PHOENIX		SIGNUM	
1. Full image	Hand patch	80.1	55.0	31.6	16.0
2. Hand patch	HoG3D	55.0	49.7	16.0	12.5
3. HoG3D	+Traj	45.2	42.1	12.5	14.2
4. HoG3D+Traj	+Face	42.1	41.9	14.2	14.2
5. HoG3D	+Full	45.2	45.2	12.5	10.7
Impact of Techniques		WER [%]		WER [%]	
6. Model Length Estimation		60.8	55.0	16.0	17.5
7. Temporal Context		51.3	49.7	12.7	12.4
8. MLP		39.8	43.3	16.0	13.0
9. Manual Tracking Annotation		55.0	48.3	–	–
10. Gap Models		42.1	39.8	–	–
11. Class LM		39.8	38.6	–	–

frame alignment of an HMM system with uniform length. No improvement over uniform length is observed due to the estimation on the frame alignment having limited accuracy and the signs in the video already sharing a similar length.

Visual pronunciation variants: Sign languages exhibit strong pronunciation variation which manifest in visual sign variants. Visual variants are not explicitly labeled in PHOENIX or SIGNUM. While in SIGNUM no variants exist because of the artificial nature of the database, PHOENIX shows high variability within signs annotated by the same gloss. This arises mainly from the interpreter mixing different dialects.

We have manually annotated the variants with regard to the visual appearance and the motion of the hand yielding on average 2.7 variants per gloss and a total of 711 different variants. Using these annotations, each variant is modelled by a distinct HMM with model length estimation achieving 56.5% WER in contrast to the baseline of 55.0%. Further, both systems outperform the 62.2% WER of a 'nearest-neighbor' style system where each gloss occurrence is modelled independently. Apparently, increasing the number of dedicated HMMs per gloss worsens recognition. Coherent manual definition of variants is likely to be a problematic factor, as well as the HMMs not generalizing well over unseen data because of the reduction in training data per HMM and strong coarticulation effects.

Tracking Influence: The presented hand patch baselines rely on tracking to localize the hands of the signer. Tracking is not perfect and errors propagate through the recognition system. Figure 2 shows the impact of tracking quality measured in tracking error rate (TrackEr) [28] counting a tracked position as wrong if it differs by more than 20 Pixel from ground truth on ASLR for PHOENIX. The TrackEr of 0 at 48.3% WER refers to using ground truth tracking annotation (see Table 2, Row 9).

HoG3D: HoG3D features encode the shape and its change over time of a tracked hand. The latter aspect is not covered by hand patch features. Further, HoG3D features are more compact than hand patches, and robust against local illumination changes. Comparing to the hand patch baselines, recognition results are improved from 55.0% to 49.7% WER for PHOENIX and from 16.0% to 12.5% WER for SIGNUM. The result on PHOENIX

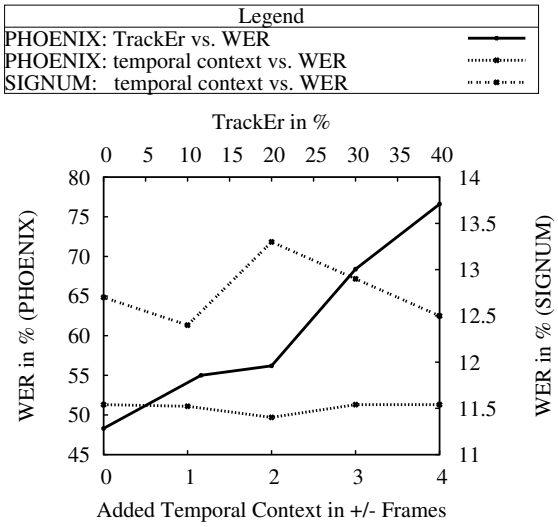

Figure 2: Solid Black: Influence of tracking performance in TrackEr on WER for PHOENIX using right hand patches features (read top x-axis vs. left y-axis). Dotted: Impact of temporal context using HoG3D (right hand) on WER for PHOENIX (read bottom x-axis vs. left y-axis). Dashed: Impact of temporal context using HoG3D (right hand) on WER for SIGNUM (read bottom x-axis vs. right y-axis)

is almost as good as using ground truth tracking information for the hand patches.

Temporal Context: The temporal context of a feature includes information that cannot easily be learned by an HMM system but has been shown to improve results in ASR [33].

Although HoG3D features already incorporate temporal context, we find that including additional context benefits the recognition, as can be seen in Figure 2. More context than ± 2 frames degrades recognition accuracy on PHOENIX, capturing too much information of the following glosses. On SIGNUM, we observe only marginal recognition improvement indicating that the context included in HoG3D is sufficient. The chosen system defaults are at ± 2 frames for PHOENIX and ± 4 frames for SIGNUM and are, thus, well chosen for both cases.

Modalities: In addition to the body pose (full image) and the right hand (HoG3D), we evaluate the performance using facial expressions (POIAAM), the left hand (HoG3D) and the movement of the right hand (Traj). For both databases, the left hand tracking quality is worse than the right hand. Henceforth ground truth tracking annotations are used for PHOENIX to avoid tracking bias. Thus, the HoG3D baseline improves to 45.2% WER. Using left hand features 63.9% respectively 51.0% WER are achieved for PHOENIX and SIGNUM. The stronger recognition degradation for PHOENIX reflects the difficulty of the database. With facial features, the recognition result is 62.6% respectively 89.3% WER for PHOENIX and SIGNUM. The high WER for SIGNUM is due to the fact that hardly any facial expressions are present here. Concatenating movement trajectory and right hand HoG3D, results are improved for PHOENIX but not for SIGNUM (Table 2, Row 3).

Using synchronous (Table 2, Row 4) and asynchronous (Table 2, Row 5) modality combination techniques, recognition

results for both databases are improved if the respectively best single modalities are combined. For a full overview of modality combination techniques and results refer to [10].

Gap Models: The SIGNUM database is designed to contain only one-handed signs and no switching of the hand. Contrarily, in PHOENIX signers partly switch hands and use the left hand for signing while holding the right. This effect introduces missing features in the information stream of the right and left hand. One way to remedy this problem is to borrow the idea of noise models from ASR and to augment the system's vocabulary by two such models. One model subsuming signs performed by the left hand only and one for long gaps between signs of more than five frames that are part of the sentence but do not belong to either neighboring sign. The training data annotation is automatically augmented by labels for both aspects using ground truth annotation. Using these gap models, the WER is improved from 42.1% to 39.8% on PHOENIX, due to the models only being populated with clean and complete data. Further, we observe an improved feature to HMM state alignment (measured as distance to the ground truth annotation).

MLP-tandem: The MLP-tandem approach was evaluated for SIGNUM and PHOENIX. For SIGNUM the MLP is trained on hand patch features resulting in 13.0% WER that outperforms the baseline by 3%. This result is comparable to the 12.5% obtained using HoG3D features. For PHOENIX, the MLP is trained on concatenated HoG3D with Trajectory features. The recognition result is with 43.3% WER (at $\pm$ 1 frame temporal context) 3.5% worse than the baseline of 39.8% obtained by the HoG3D+Traj features alone. Including more temporal context does not help because it is already included in the MLP posterior estimates. Two aspects feature into the performance of the MLP features on PHOENIX. On the one hand, it is not clear if the MLP can reliably extract the relevant information from the HoG3D+Traj features although following the ASR praxis of using the best feature available. On the other hand, the MLPs for PHOENIX and SIGNUM have about the same number of parameters but the MLP on SIGNUM is trained using ten times the data of PHOENIX. Anyhow, the results show that MLP features as used in ASR achieve comparable results to specialized features from CV although requiring training themselves.

Class LM: With regard to PHOENIX the analysis of the recognition errors showed that 3.8% absolute of all errors are due to misclassified numbers and 2.2% absolute are due to orientations such as *north*. Further, both classes appear in a specific context such as a number before the gloss TEMPERATURE which is not adequately captured by sign-level LMs. Additionally, numbers have a low frequency in the LM training data appearing on average less than ten times. Augmenting the LM for PHOENIX with a class for numbers, the perplexity (PPL) on the test data is reduced from 34.9 to 29.3. Orientations reduce PPL to 31.2 and using both classes PPL is reduced to 25.7.

Table 3 shows that using the orientation category the recognition performance is only marginally improved but using the number category alone improves the overall recognition result by 1.2% WER. Other categories as used in sign language translation [25] did not improve results. For SIGNUM, class LMs have not been used because of the special and artificial structure of the sentences.

Table 3: Class LM results for PHOENIX. Error rates in %.

Class	del/ins	WER
None	20.7/4.5	39.8
Orientation	18.1/5.3	39.2
Numbers	19.3/4.1	38.8
+ Orientation	16.2/6.2	**38.6**

Signer adaptation: Applying the findings on SIGNUM single signer to the case of 25 signers and using tracked hand patches of the right hand as features, the system achieves 23.6% WER.

In ASR SAT is used to adapt the features to better fit the learned models. In the same fashion, we use SAT to adapt the baseline system to the signers sign patterns. In a second training pass, signer specific feature transformation matrices are estimated using CMLLR. In SIGNUM the signer ids are annotated and hence no signer clustering is performed.

Using the signer ids of the test data, it is possible to evaluate what is the maximal achievable improvement in terms of WER using SAT/CMLLR on the given test data. In the typical recognition setup the ids of the signers in the test data are not known and the resulting improvement is lower due to errors in the clustering process. Adapting the proposed recognition system build for the SIGNUM multi-signer database using SAT/CMLLR, the WER of 23.6% is improved to 22.1% showing that the standard approach from ASR is applicable to ASLR without any modifications.

4. Summary and conclusion

In this work, a large-vocabulary ASLR system for continuous sign language using single-view videos as well as the process of feature selection, technology transfer from ASR and CV and system design have been presented. Techniques from ASR and CV have been evaluated in the context for ASLR for challenging 'real-life' data and data designed for pattern recognition.

Some aspects were found to generalize over both data sets: HoG3D alone outperforms all other tested features with MLPs being a close second. The combination of the two best single performing modalities achieves the best combination result and the system benefits from including temporal context in features.

Other findings are related to particularities of the given corpora: On PHOENIX, gap models improve results but use specific annotations not necessarily available in other corpora. The improvement by class LMs exploits domain-specific knowledge and model length estimation relies on accurate sign boundaries.

To sum up, the WER on 'real life' data has been reduced from over 80% to 38.6% and on lab data from over 30% to 10.7% for single signer and to 22.1% for multi signer. Although this might sound very high compared to the state-of-the-art in ASR, this is one of the first times that recognition results have been published on 'real-life' data. We believe that our work helps pushing ASLR towards more realistic application scenarios, which come along with challenges most of the current sign language data sets ignore. This goes especially for the use of single-view video material in contrast to using special hardware such as bulky cyber gloves, or stereo cameras.

Future work will investigate sub units and coarticulation effects.

5. Acknowledgements

This work has been partly funded by the European Community's Seventh Framework Programme FP7-ICT-2007-3, grant agreement 231424 - SignSpeak project and the Janggen-Pöhn-Stiftung. Special thanks to Thomas Hoyoux (Technical University Innsbruck) and Yvan Richard (Centre de Recerca i Innovacio de Catalunya (CRIC)) for providing AAM and HoG3D features.

6. References

[1] C. Vogler and D. Metaxas, "A framework for recognizing the simultaneous aspects of american sign language," *Computer Vision and Image Understanding*, vol. 81, no. 3, pp. 358–384, 2001.

[2] J. Deng and H. T. Tsui, "A Two-step Approach based on PaHMM for the Recognition of ASL," in *ACCV*, Jan 2002.

[3] C. Wang, X. Chen, and W. Gao, "Expanding Training Set for Chinese Sign Language Recognition," in *FG*, 2006.

[4] E.-J. Ong, H. Cooper, N. Pugeault, and R. Bowden, "Sign language recognition using sequential pattern trees," in *CVPR*, Jun. 16 – 21 2012, pp. 2200 – 2207.

[5] V. Pitsikalis, S. Theodorakis, C. Vogler, and P. Maragos, "Advances in phonetics-based sub-unit modeling for transcription alignment and sign language recognition," in *CVPR*, 2011, pp. 1 – 6.

[6] T. Hanke, "HamNoSys - representing sign language data in language resources and language processing contexts," in *LREC 2004, Workshop proceedings : Representation and processing of sign languages*, 2004, pp. 1 – 6.

[7] O. Koller, H. Ney, and R. Bowden, "May the force be with you: Force-aligned signwriting for automatic subunit annotation of corpora," in *FG*, Apr. 2013.

[8] V. Sutton and Writing, Deaf Action Committee for Sign, *Sign writing*. Deaf Action Committee (DAC), 2000.

[9] H. Cooper, E.-J. Ong, N. Pugeault, and R. Bowden, "Sign language recognition using sub-units," *JMLR*, vol. 13, pp. 2205–2231, Jul 2012.

[10] J. Forster, C. Oberdörfer, O. Koller, and H. Ney, "Modality combination techniques for continuous sign language recognition," in *IbPRIA*, Jun. 2013.

[11] P. Buehler, M. Everingham, and A. Zisserman, "Learning sign language by watching TV (using weakly aligned subtitles)," in *CVPR*, 2009.

[12] H. Cooper and R. Bowden, "Learning signs from subtitles: A weakly supervised approach to signlanguage recognition," in *CVPR*, Jun 2009, pp. 2568 – 2574.

[13] H.-D. Yang, S. Sclaroff, and S.-W. Lee, "Sign language spotting with a threshold model based on conditional random fields," *PAMI*, vol. 31, no. 7, pp. 1264–1277, July 2009.

[14] S. Najak, K. Duncan, S. Sarkar, and B. Loeding, "Finding recurrent patterns from continuous sign language sentences for automated extraction of signs," *JMLR*, vol. 13, pp. 2589 – 2615, Dec 2012.

[15] R. Yang, S. Sarkar, and B. Loeding, "Handling movement epenthesis and hand segmentation ambiguities in continuous sign language recognition using nested dynamic programming," *PAMI*, vol. 32, no. 3, pp. 462–477, Mar 2010.

[16] S. Ong and S. Ranganath, "Automatic sign language analysis: a survey and the future beyond lexical meaning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 6, pp. 873–891, 2005.

[17] A. Kläser, M. Marsza ek, and C. Schmid, "A spatio-temporal descriptor based on 3d-gradients," in *BMVC*, Sep 2008, pp. 995–1004.

[18] P. Dreuw, D. Rybach, T. Deselaers, M. Zahedi, and H. Ney, "Speech recognition techniques for a sign language recognition system," in *Interspeech*, Aug. 2007, pp. 2513–2516, iSCA best student paper award.

[19] U. von Agris, M. Knorr, and K.-F. Kraiss, "The significance of facial features for automatic sign language recognition," in *FG*, Sep 2008, pp. 1–6.

[20] J. Forster, C. Schmidt, T. Hoyoux, O. Koller, U. Zelle, J. Piater, and H. Ney, "Rwth-phoenix-weather: A large vocabulary sign language recognition and translation corpus," in *LREC*, May 2012.

[21] R. Bakis, "Continuous speech word recognition via centisecond acoustic states," in *91st Meeting of the Acoustical Society of America (ASA)*, Washington, DC, USA, April 1976.

[22] A. Emami and S. Chen, "Multi-class model m," in *ICASSP*, May 2011, pp. 5516–5519.

[23] S. F. Chen and S. M. Chu, "Enhanced word classing for model m," in *Interspeech*, 2010, pp. 1037–1040.

[24] W. Naptali, M. Tsuchiya, and S. Nakagawa, "Topic-dependent-class-based n-gram language model," *ASL*, vol. 20, no. 5, pp. 1513–1525, 2012.

[25] D. Stein, C. Schmidt, and H. Ney, "Analysis, preparation, and optimization of statistical sign language machine translation," *Machine Translation*, vol. 26, no. 4, pp. 325–357, Dec. 2012.

[26] A. Stolcke, J. Zheng, W. Wang, and V. Abrash, "SRILM at Sixteen: Update and outlook," in *ASRU*, Waikoloa, Hawaii, December 2011.

[27] S. F. Chen and J. Goodman, "An empirical study of smoothing techniques for language modeling," *Computer Speech & Language*, vol. 13, no. 4, pp. 359–393, 1999.

[28] P. Dreuw, T. Deselaers, D. Rybach, D. Keysers, and H. Ney, "Tracking using dynamic programming for appearance-based sign language recognition," in *FG*, 2006, pp. 293–298.

[29] Z. Tüske, M. Sundermeyer, R. Schlüter, and H. Ney, "Context-dependent mlps for lvcsr: Tandem, hybrid or both?" in *Interspeech*, Portland, OR, USA, Sep. 2012.

[30] G. R. J. Schenk, "Novel hybrid nn/hmm modelling techniques for on-line handwriting recognition," in *IWFHR*, Oct 2006, pp. 619 – 623.

[31] J. Piater, T. Hoyoux, and W. Du, "Video Analysis for Continuous Sign Language Recognition," in *4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies*, May 2010, lREC.

[32] B. Hoffmeister, R. Schlüter, and H. Ney, "icnc and irover: The limits of improving system combination with classification?" in *Interspeech*, Sep. 2008, pp. 232–235.

[33] A. Zolnay, R. Schlüter, and H. Ney, "Acoustic feature combination for robust speech recognition," in *ICASSP*, 2005.