SignSpeak Project

Scientific understanding and vision-based technological development for continuous sign language recognition and translation

Month 15 Evaluation Report

Major Deliverable D.7.1.M15

Release version: V1.0

Grant Agreement Number 231424

Small or medium-scale focused research project (STREP)
FP7-ICT-2007-3. Cognitive Systems, Interaction, Robotics

Project start date: 1 April 2009
Project duration: 36 months

Dissemination Level		
PU	Public (can be made available outside of SignSpeak Consortium without restrictions)	X
RE	Restricted to SignSpeak Programme participants and a specified group outside of SignSpeak consortium	
IN	SignSpeak Internal (only available to (all) SignSpeak programme participants)	
LI	SignSpeak Limited (only available to a specified subset of SignSpeak programme participants)	
Distribution list (only for RE or LI documents)		

0 General Information

0.1 Document

Title	Month 15 Evaluation Report
Type	Major Deliverable
Ref	D.7.1.M15
Target version	V1.0
Current issue	V1.0
Status	Draft
File	`D.7.1.M15.EvaluationReport.tex`
Author(s)	Philippe Dreuw, Jens Forster, and Hermann Ney / RWTH Justus Piater, Du Wei, and Thomas Hoyoux / ULg Gregorio Martínez, Jaume Vergés-Llahí, and Juan D. García-Arteaga / CRIC
Reviewer(s)	Gregorio Martinez / CRIC
Approver(s)	Gregorio Martinez / CRIC
Approval date	
Release date	15/09/2010

0.2 History

Date	Version	Comment
15/09/2010	V0.2	updated description of D.7 report including preliminary CNGT results
28/07/2010	V0.1	released first description of D.7 report

0.3 Document scope and structure

The tasks in WP7 are intended to evaluate the deliverables generated within the technical work packages (WP3, WP4 and WP5), with the aim of providing a constant monitoring of progress and obtaining feedback for the next developments. The document consists of a first part describing the objectives of the project in the present project, followed by the evaluation of the different elements composing each WP, namely, the multi-modal visual analysis and the sign language recognition and translation tasks.

Authors	Group
Dreuw, Philippe	RWTH
Forster, Jens	RWTH
Gweth, Yannick	RWTH
Ney, Hermann	RWTH
Stein, Daniel	RWTH
Zelle, Uwe	RWTH
Piater, Justus	ULg
Wei, Du	ULg
Hoyoux, Thomas	ULg
Martinez, Gregorio	CRIC
Vergés-Llahí, Jaume	CRIC
García-Arteaga, Juan D.	CRIC

0.4 Content

Table 1: Expected corpus annotation progress of the RWTH-PHOENIX and Corpus-NGT corpora in comparison to the limited domain speech (Verbmobil II) and translation (IWSLT) corpora.

	BOSTON-104	Phoenix		Corpus-NGT		Vermobil II	IWSLT
year	2007	2009	2011	2009	2011	2000	2006
recordings	201	78	400	116	300	-	-
running words	0.8k	10k	50k	30k	80k	700k	200k
vocabulary size	0.1k	0.6k	$<$ **2.5k**	3k	$<$ **5k**	10k	10k
T/T ratio	8	15	$>$ **20**	10	$>$ **20**	70	20
Performance	11% WER [18]	- -		- -		15% WER [27]	40% TER [28]

1 Project Objectives for the Period

The tasks in WP7 are intended to evaluate the deliverables generated within the technical work packages (WP3, WP4 and WP5), with the aim of providing a constant monitoring of progress and obtaining feedback for the next developments.

For the first year, prototypes have been created for the multimodal visual analysis (D.3.2), the sign language recognition (D.4.1) and sign language translation (D.5.1). This deliverable D7.1 gathers the evaluation of these three prototypes.

Major achievements in reaching scientific and technological project objectives were:

- A survey of video databases that can be used within a continuous sign language recognition scenario to measure the performance of head and hand tracking algorithms either w.r.t. a tracking error rate or w.r.t. a word error rate criterion is presented in this report.
- Robust tracking algorithms are required as the signing hand frequently moves in front of the face, may temporarily disappear, or cross the other hand.
- Only few studies consider the recognition of continuous sign language, and usually special devices such as colored gloves or blue-boxing environments are used to accurately track the regions-of-interest in sign language processing.
- Ground-truth labels for hand and head positions have been annotated for more than 30k frames in several publicly available video databases of different degrees of difficulty, and preliminary tracking results are presented for different tracking approaches, such as model-free or person-dependent approaches. Furthermore, evaluation of WP4 and WP5 has also been accomplished by measuring error rates in the processes of sign recognition and translation.

2 Technical Accomplishment

The present section will establish the kind of data employed throughout the document in order to evaluate and validate the performance of the different WP.

In order to build a Sign-Language-to-Spoken-Language translator, reasonably sized corpora have to be created for statistically-based data-driven approaches. For a limited domain speech recognition task (Verbmobil II) as e.g. presented in [27], systems with a vocabulary size of up to 10k words should be trained with at least 700k words to obtain a reasonable performance, i.e. about 70 observations per vocabulary entry. Similar values should be obtained for a limited domain translation task (IWSLT) as e.g. presented in [28].

Similar corpora statistics can be observed for other ASR or MT tasks. The requirements for a sign language corpus suitable for recognition and translation can therefore be summarized as follows:

- annotations for a limited domain (i.e. broadcast news, etc.)
- for a vocabulary size smaller than 4k words, each word should be observed at least 20 times
- the singleton ratio should ideally stay below 40%

Existing corpora should be extended to achieve a good performance w.r.t. recognition and translation [23]. During the SignSpeak project, the existing RWTH-PHOENIX corpus [38] and Corpus-NGT [11] will be extended to meet these demands (c.f. Table 1). Novel facial features [33] developed within the SignSpeak project are

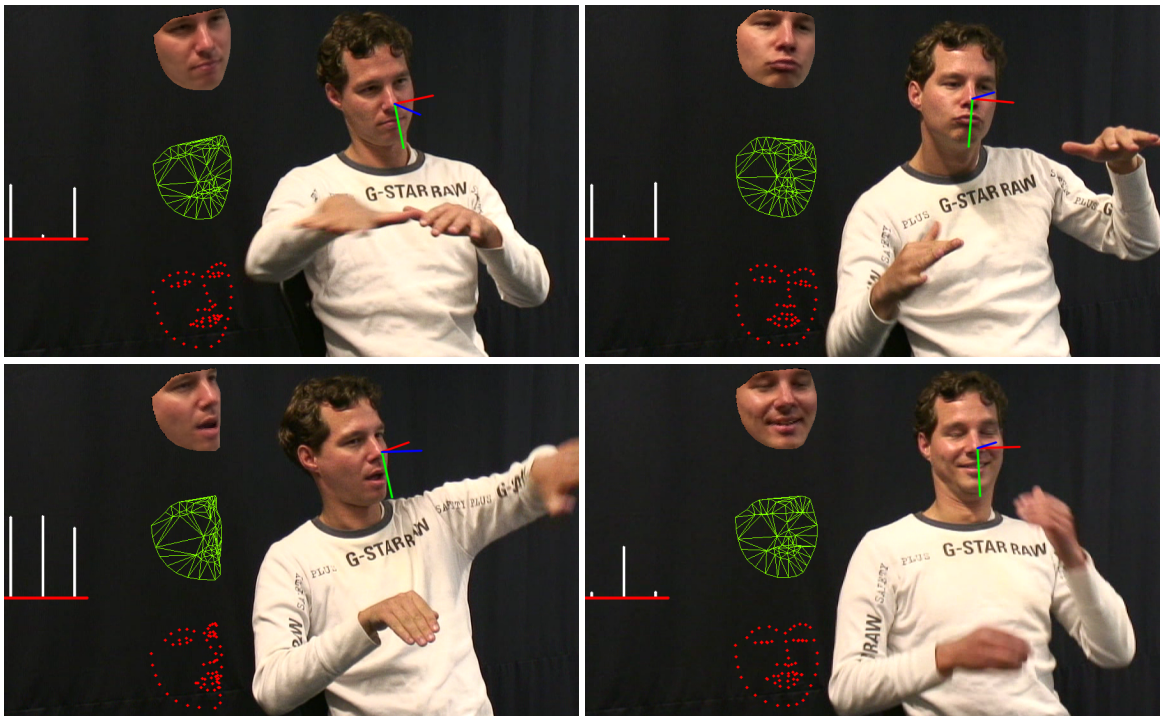

Figure 1: Facial feature extraction on the Corpus-NGT database (f.l.t.r.): three vertical lines quantify features like left eye aperture, mouth aperture, and right eye aperture; the extraction of these features is based on a fitted face model, where the orientation of this model is shown by three axis on the face: red is X, green is Y, blue is Z, origin is the nose tip.

shown in Figure 1 and will be analyzed for continuous sign language recognition w.r.t. WER and TER criterions using the annotated benchmark corpora described in paragraph 2.1.2.2.

2.1 Evaluation of the Multimodal Visual Analysis (Task 7.1)

The objective of this Task is to provide a first quantitative evaluation of the individual components for visual analysis developed under WP3 by this time.

2.1.1 Workpackage objectives and starting point at the beginning of the period

Quantitative evaluation requires the availability of

- annotated training data for the construction of appearance models (for model-based hand and face tracking methods),
- annotated ground truth as a reference for evaluation.

At the beginning of the project, no annotated data for face model construction were available. Annotations useful for hand model construction and ground truth for evaluation were available in limited quantities for some datasets only.

2.1.2 Progress towards objectives

ULg expended considerable effort on labeling face images for the construction and evaluation of baseline face models, permitting an initial, indicative performance evaluation. An exhaustive evaluation would require substantially more investment, consuming resources better spent on developing new methods intended to provide better performance using less manual intervention.

A major activity of RWTH in WP7 has been the annotation ground-truth in order to evaluate the performance of hand and head tracking algorithms. RWTH labeled hand and head positions in a subset of 7.891 images of Corpus-NGT and distributed the ground-truth to ULg and CRIC. Furthermore, RWTH distributed complete ground-truth for tracking of head and hand configurations of the RWTH-BOSTON-104 database consisting of 15.746 images to the project partners.

The activity of CRIC in WP7 has consisted in evaluating the performance of the head/hands detection and tracking algorithm using the aforementioned corpus NGT annotated by RWTH.

2.1.2.1 Introduction

Tracking is especially important if motion trajectories have to be recognized, e.g. for collision detection, gait analysis [34], marker-less motion capturing [8], or vision-based gesture or sign language recognition [3, 16]. Numerous tracking models of different complexity have been discussed in the literature [24, 1, 35, 12, 6], but they are typically task and environment dependent, or require special hardware . Under realistic circumstances, the performance of most current approaches decreases dramatically as it heavily depends upon possibly wrong local decisions [25].

A common assumption is that the target object is moving most over time. Opposed to a relatively rough bounding-box based tracking of e.g. persons or cars for tracking-only tasks, usually special devices such as colored gloves or blue-boxing environments are used to accurately track the regions-of-interest (such as the head, the hands, etc.) for tracking *and* recognition tasks in sign language processing.

Only few studies consider the recognition of continuous sign language. Most of the current sign language recognition systems use specialized hardware [22, 46] and are person dependent [43, 3, 6], i.e. can only recognize the signers they were designed for.

Furthermore, most approaches focus on the recognition of isolated signs or on the even simpler case of recognizing isolated gestures [45], which can often be characterized just by their movement direction. The recognition of continuous sign language is usually performed by hidden Markov model (HMM) based systems. An HMM-based approach for French Sign Language recognition has been proposed in [5], where a data glove was used to obtain hand appearance and position. Starner et al. presented an American Sign Language (ASL) recognition system [37], Holden et al. proposed an Australian Sign Language recognition system based on HMMs [26], and e.g. Bauer and Kraiss proposed a German Sign Language recognition system based on HMMs [2] in which the signer wore simple colored gloves to obtain data. Ong et al. [31] give a review on recent research in sign language and gesture recognition.

The main objectives of this report are:

- To provide a brief survey of video databases that can be used within a continuous sign language recognition scenario to measure the performance of head and hand tracking algorithms either w.r.t. a tracking error rate or w.r.t. a word error rate criterion
- To show that a conceptually simple model-free tracking model can be used in several sign language tracking and recognition tasks
- To briefly compare model-free and model-based tracking algorithms
- To show the impact of person-dependency in model-based tracking algorithms

2.1.2.2 Benchmark Databases

All databases presented in this section are used within the SignSpeak project and are either freely available or available on request. The SignSpeak[1] project tackles the problem of automatic recognition and translation of continuous sign language [20]. The overall goal of the SignSpeak project is to develop a new vision-based technology for recognizing and translating continuous sign language (i.e. provide Video-to-Text technologies).

Example images showing the different recording conditions are shown for each database in Figure 3, where Table 2 gives an overview how the different corpora can be used for evaluation experiments.

For an image sequence $X_1^T = X_1, \ldots, X_T$ and corresponding annotated hand positions $u_1^T = u_1, \ldots, u_T$, we define the tracking error rate (TER) of tracked positions $\hat{u}_1^T$ as the relative number of frames where the Euclidean distance between the tracked and the annotated position is larger than or equal to a tolerance τ:

$$\text{TER} = \frac{1}{T}\sum_{t=1}^{T}\delta_\tau(u_t,\hat{u}_t) \qquad \text{with} \qquad \delta_\tau(u,v) := \begin{cases} 0 & \|u-v\| < \tau \\ 1 & \text{otherwise} \end{cases} \tag{1}$$

Depending on the database format, a viewport for TER calculation can be specified in addition. Frames, in which the hands are not visible, are disregarded, resulting in a different number of frames to be evaluated (e.g. in Table 4, the dominant-hand is only visible in 12909 frames of the 15746 annotated frames, the head is always visible). Examples of annotated frames and evaluation viewport borders are shown in Figure 2: in the left image,

[1] http://www.signspeak.eu

Table 2: Freely available sign language corpora and their evaluation areas ($\times$: unsuitable or unannotated, $\surd$: already annotated, $\star$: annotations underway)

Corpus	Evaluation Areas			
	Isolated Recog.	Continuous Recog.	Tracking	Translation
Corpus-NGT	$\surd$	$\surd$	$\surd$	$\surd$
RWTH-BOSTON-50	$\surd$	$\times$	$\surd$	$\times$
RWTH-BOSTON-104	$\times$	$\surd$	$\surd$	$\times$
RWTH-BOSTON-400	$\times$	$\surd$	$\times$	$\times$
RWTH-PHOENIX-v1.0	$\surd$	$\surd$	$\star$	$\surd$
RWTH-PHOENIX-v2.0	$\times$	$\surd$	$\star$	$\surd$
ATIS-ISL	$\times$	$\surd$	$\surd$	$\surd$
SIGNUM	$\surd$	$\surd$	$\star$	$\times$

Figure 2: Example of ground-truth annotations and evaluation viewport borders: ground-truth annotations within the red-shaded area are disregarded in the corresponding TER calculation

all annotated ground-truth points are within a specified evaluation viewport border and will be considered for TER calculation, whereas in the right image both the dominant hand and non-dominant hand (i.e. right and left hand, annotated by the green and red circle, correspondingly) are out of the viewport border and will be ignored for TER calculation.

Corpus-NGT Database The Corpus-NGT[2] database is a 72 hour corpus of Sign Language of the Netherlands. It is the first large open access corpus for sign linguistics in the world. It presently contains recordings from 92 different signers, mirroring both the age variation and the dialect variation present in the Dutch Deaf community [11].

For the SignSpeak project, the limited gloss annotations that were present in the first release of 2008 have been considerably expanded, and sentence-level translations have been added. Currently, 280 video segments with about **8k frames** have been annotated to evaluate hand and head tracking algorithms (c.f. Table 3).

[2] `http://www.corpusngt.nl`

Table 3: Freely available tracking ground-truth annotations in sign language corpora for e.g. hand and face positions

Corpus	Annotated Frames
Corpus-NGT	7891
RWTH-BOSTON-50	1450
RWTH-BOSTON-104	15746
ATIS-ISL	5757

Boston Corpora All corpora presented in this section are freely available for further research in linguistics, tracking, recognition, and translation[3].

The data was recorded within the ASLLRP[4] project by Boston University, the database subsets were defined at the RWTH Aachen University in order to build up benchmark databases [19] that can be used for the automatic recognition of isolated and continuous sign language.

The RWTH-BOSTON-50 corpus was created for the task of isolated sign language recognition [47]. It has been used for nearest-neighbor leaving-one-out evaluation of isolated sign language words. About **1.5k frames** in total are annotated and are freely available (c.f. Table 3).

The RWTH-BOSTON-104 corpus has been used successfully for continuous sign language recognition experiments [16, 21]. For the evaluation of hand tracking methods in sign language recognition systems, the database has been annotated with the signers' hand and head positions. More than **15k frames** in total are annotated and are freely available (c.f. Table 3).

For the task of sign language recognition and translation, promising results on the publicly available RWTH-BOSTON-104 corpus have been achieved for automatic sign language recognition [16] and translation [21, 17] that can be used as baseline reference for other researchers. However, the preliminary results on the larger RWTH-BOSTON-400 corpus show the limitations of the proposed framework and the need for better visual features, models, and corpora [19].

Phoenix Weather Forecast Corpora The RWTH-PHOENIX corpus with German sign language annotations of weather-forecasts has been first presented in [38] for the purpose of sign language translation (referred to as RWTH-PHOENIX-v1.0 corpus in this work). It consists of about 2k sentences, 9k running words, with a vocabulary size of about 1.7k signs. Although the database is suitable for recognition experiments, the environment conditions in the first version are more challenging for robust feature extraction such as hand tracking (c.f. Figure 3). During the SignSpeak project, a new version RWTH-PHOENIX-v2.0 is recorded and annotated to meet the demands described in paragraph 2.1.2.4. Due to simpler environment conditions in the RWTH-PHOENIX-v2.0 version (see also Figure 3), promising feature extraction and recognition results are expected. Ground-truth annotations are currently added for about **8k frames** and will be freely available in the near future (c.f. Table 3).

The ATIS Irish Sign Language Corpus The ATIS Irish sign language corpus (ATIS-ISL) has been presented in [7], and is suitable for recognition and translation experiments. The Irish sign language corpus formed the first translation into sign language of the original ATIS data, a limited domain corpus for speech recognition and translation tasks. The sentences from the original ATIS corpus are given in written English as a transcription of the spoken sentences. The ATIS-ISL database as used in [39] contains 680 sentences with continuous sign language, has a vocabulary size of about 400 signs, and contains several speakers. For the SignSpeak project, about **6k frames** have been annotated with hand and head positions to be used in tracking evaluations (c.f. Table 3).

SIGNUM Corpus The SIGNUM[5] corpus has been first presented in [44] and contains both isolated and continuous utterances of various signers. This German sign language corpus is suitable for signer independent continuous sign language recognition tasks. It consists of about 33k sentences, 700 signs, and 25 speakers, which results in approximately 55 hours of video material. Ground-truth annotations will be added in the near future (c.f. Table 3).

2.1.2.3 Hand and Head Tracking for Sign Language Recognition

For feature extraction, relevant body parts such as the head and the hands have to be found. To extract features which describe manual components of a sign, at least the dominant hand has to be tracked in each image sequence. A robust tracking algorithm is required as the signing hand frequently moves in front of the face, may temporarily disappear, or cross the other hand.

Hand Tracking The head and hand tracking tracking algorithm described in [15] (DPT) is based on dynamic programming and is inspired by the time alignment algorithm in speech recognition which guarantees to find the optimal path w.r.t. a given criterion and prevents taking possibly wrong local decisions.

[3] http://www-i6.informatik.rwth-aachen.de/aslr/
[4] http://www.bu.edu/asllrp/
[5] http://www.phonetik.uni-muenchen.de/forschung/Bas/SIGNUM/

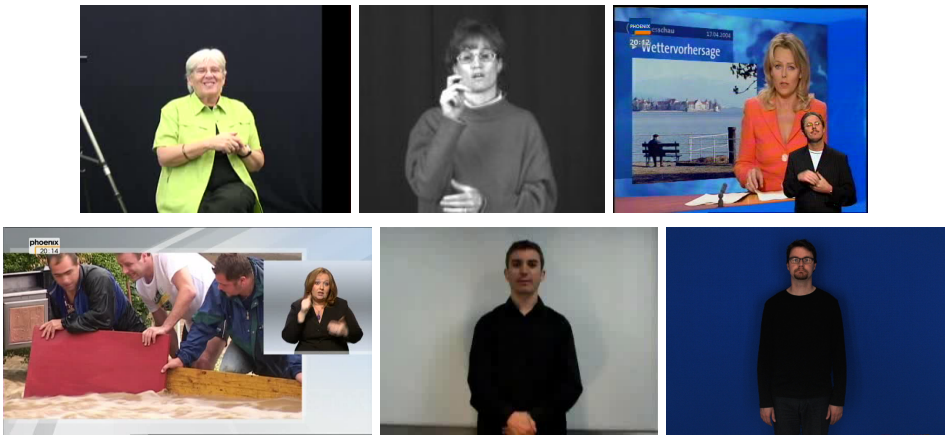

Figure 3: Example images from different video-based sign language corpora (f.l.t.r.): Corpus-NGT, RWTH-BOSTON, RWTH-PHOENIX v1.0, RWTH-PHOENIX v2.0, ATIS-ISL, and SIGNUM

Instead of requiring a near perfect segmentation for these body parts, the decision process for candidate regions is postponed to the end of the entire sequences by tracing back the best decisions. No training is required, as it is a model-free and person independent tracking approach.

Another approach used in this work is based on the work of [33]. The hand tracking system contains two steps, skin-color region segmentation followed by PCA-based template matching. Individual PCA models are trained for the left and right hands of each signer.

Due to the nature of the sequences, which have a controlled illumination, static background, and distinguished singer's body parts, a model-based and personal-independent approach can be achieved by taking into account cues such as color and movement. This way, an adaptive skin-color based segmentation (ASCS) algorithm [13] is able to extract image regions where candidates to be hands and face are to be found. Tracking of candidate regions is performed with CamShift algorithm [10, 4]. A posterior heuristic is employed to only keep track of the most likely candidates. This is the least discriminative of all previous approaches. Though it has advantages in process speed and being personal-independent, meaning that no training is required and thus it can be readily used, on the one hand, it is also less precise in locating the body parts due to the kind of areas obtained and unable to distinguish by itself dichotomies such as head/hands and left/right hands, on the other hand.

Head Tracking In an Eigenface approach[40], the distance to the face-space can be seen as a measure of faceness and can thus be used as a score. To train the eigenfaces in [15], the BioID[6] database has been used, i.e. the head tracking approach is model-based but person-independent (c.f. Table 4). As faces generally are skin colored, a skin color model can be used as an additional score within the DPT approach.

The active appearance model (AAM) based face tracker proposed by [33] is composed of an offline part, where a statistical face model containing the facial appearance variation information is trained, and an online part, where the facial features are tracked in real time using that model. Because the fitting method is a local search, they initialize the AAM using the face detector by Viola and Jones [42].

In contrast to the tracking approaches, a model-based face detection approach is used for comparison where the faces have been automatically detected using the OpenCV implementation of the Viola & Jones [42] face detector. As the cascades have been trained on different data, the detection approach is model-based but person-independent (c.f. Table 4).

As explained in previous section, tracking regions based on the adaptive skin-color segmentation approach (ASCS) does not *a priori* distinguish between head and hands. Therefore, tracking using CamShift algorithm is performed equally for head and hand candidates. Heads are selected only afterwards by means of an heuristic which takes size, shape and appearance into account. Therefore, in the case of head tracking, this scheme has the same advantages and disadvantages previously stated.

2.1.2.4 Experimental Results and Requirements

Tracking Results For $\tau = 20$, the model-free and person independent DPT [15] tracking approach can achieve already 8.37% TER on the 12909 frames of full RWTH-BOSTON-104 dataset, and 8.83% TER on the

[6] http://www.bioid.com

Table 4: Hand and head tracking on RWTH-BOSTON-104 dataset

Tracking	Model	Pers. dep.	# Frames	Setup	TER τ=5	τ=10	τ=15	τ=20
Dominant Hand	no	no	12909	DPT [15]	73.59	42.29	18.79	8.37
	no	no	2603	DPT [15]	74.79	44.33	20.43	8.83
	yes	yes	2603	Robust PCA [33]	89.86	77.41	64.50	47.48
Non-Dominant Hand	yes	yes	842	Robust PCA [33]	80.19	57.78	39.39	24.06
Head	yes	no	15732	DPT + PCA [15]	26.77	17.32	12.70	10.86
	yes	no	15732	Viola & Jones [42]	9.75	1.23	1.09	1.07
	yes	no	15732	Viola & Jones + Kalman	10.04	0.81	0.73	0.68
	yes	yes	15732	AAM [33]	10.17	6.85	6.82	6.81
	yes	no	15732	AAM [33]	10.92	7.92	7.88	7.76

Table 5: Hand and head tracking on Corpus-NGT dataset

Tracking	Model	Pers. dep.	# Frames	Setup	TER τ=5	τ=10	τ=15	τ=20
Dominant Hand	no	no	7891	DPT [15]	97.26	85.62	67.88	52.15
	yes	no	7891	ASCS[13, 10, 4]	92.11	75.08	56.20	41.16
Non-Dominant Hand	yes	no	7891	ASCS[13, 10, 4]	89.82	69.72	53.28	42.62
Head	yes	no	7891	DPT [15]	98.18	92.13	75.82	59.43
	yes	no	7891	Viola & Jones [42]	78.13	62.07	59.59	58.52
	yes	no	7891	Viola & Jones + Kalman	56.92	26.04	17.55	15.81
	yes	no	7891	ASCS [13, 10, 4]	93.05	82.00	65.40	45.36

2603 test frames, where the dominant-hand is visible (c.f. Table 4).

To train the person-dependent hand models by the Robust PCA [33] approach, a set of training data is collected for each signer from the training set of 161 sequences of the RWTH-BOSTON-104 database. On the full RWTH-BOSTON-104 database, the Robust PCA [33] approach achieved an averge error of 9.57 pixels for the left hand and 16.22 pixels for the right hand. On the test set of the RWTH-BOSTON-104 database, the average errors for the left and the right hand are 10.19 pixels and 17.94 pixels. Note that, although the Robust PCA [33] approach achieved large error rates than the DPT [15] approach, the hand regions in [33] are larger, 51 by 51, and considerablely large overlappings between the tracked regions and the ground-truth hand regions are obtained.

The AAM model used to conduct the person-independent evaluation of the the AAM-based face tracker on RWTH-BOSTON-104 dataset was built from 52 images from each of the three signers from this dataset. These images correspond to 156 frames from the test set, picked from 22 of the 40 test sequences.

The model-based and person-dependent AAM approach [33] does not outperform the Viola & Jones [42] approach due to model-fitting problems and thus missing face detections in about 700 frames. On the other hand, in contrast to the latter, the AAM approach produces much more detailed tracking results than are evaluated here, including detailed shape and appearance parameters.

To conduct the person-independent evaluation of the the AAM-based face tracker on RWTH-BOSTON-104 dataset, the same 156 face images used for the person-dependent evaluation were used to build 3 AAM models, using 104 frames from 2 signers and leaving out 52 frames from the third signer for each model. The averaged results are presented in Table 4. As expected the person-dependent AAM outperforms person-independent ones. However a person-independent model built from only two persons can only poorly generalize and in the future, experiments should be conducted using models built from more different persons.

The performance of both DPT tracking and Viola & Jones detection based approaches is relatively poor

Table 6: Hand tracking on ATIS-ISL dataset

Tracking	Model	Pers. dep.	# Frames	Setup	TER τ=5	τ=10	τ=15	τ=20
Dominant Hand	no	no	5660	DPT [15]	95.46	76.17	53.18	35.9

on the Corpus-NGT database (c.f. Table 5). This can be explained by the high number of near-profile head images in the database, as both person-independent models have been trained on near frontal images only. The proposed Kalman Filter-like tracking approach in combination with Viola & Jones detections can reduce this effect.

Although ULg conducted an evaluation of the AAM-based face tracker on RWTH-BOSTON-104 dataset, a rigourous evaluation of this face tracker on Corpus-NGT dataset is not currently available. The reason is that, in order for such evaluation to be relevant, a substantial effort is required in face image labeling as the currently available labeled examples are not in sufficient number, and ULg considers as more efficient to invest its resources in developing new methods able to cope with the difficult conditions present in Corpus-NGT dataset.

The performance of ASCS [13] combined with CamShift [10, 4] on the NGT Corpus sequences is, for TER $\tau = 20$, 41.16% for the right hand and 45.36% for the head, which is comparable to the results obtained with the DPT tracking [15] and Viola & Jones detection [42] alone, while it is significatively inferior when a Kalman Filter-like tracking approach is combined with Viola & Jones detection. In both cases, these results are due to the same cause: ASCS algorithm obtains regions which correspond, in the case of hands, most of the times to arms and, in the case of head, to faces plus other close body parts with similar color. Therefore, the points used to compute hands and head positions differe by construction from those employed in the corpus annotations. This is a skew in the computation of the correct hands and head positions, which can be only estimated indirectly from the position of these skin-colored regions. As consequence, both head and hands positions will always have bigger errors than other more discriminant approaches.

2.1.2.5 Conclusions

Ground-truth labels for hand and head positions have been annotated for more than 30k frames in several publicly available video databases of different degrees of difficulty, and preliminary tracking results have been presented, which can be used as baseline reference for further experiments.

The proposed benchmark corpora can be used for tracking as well as for word error rate evaluations in isolated and continuous sign language recognition, and furthermore allow for a comparison of model-free and person-independent / person-dependent tracking approaches.

2.2 Evaluation of the Sign Language Recognition (Task 7.2)

The goal of Task 7.2 is to evaluate the developed techniques for isolated and continuous sign language recognition on several languages:

- Sign Language of the Netherlands (NGT)
- American Sign Language (ASL)
- German Sign Language (DGS)
- Irish Sign Language (ISL)

The quality of the recognition results will be measured in WER. Novel features and tracking methods will be evaluated combined and independently to see which benefit recognition results.

2.2.1 Workpackage objectives and starting point at the beginning of the period

At the beginning of the period, RWTH invested considerable effort in analyzing the status of annotations in Corpus-NGT. During the period the number of gloss annotations in Corpus-NGT has been extended and subsets of data suitable for sign language recognition have been defined (c.f. paragraph 2.2.2.3). Furthermore, RWTH started to record weather forecast data in German Sign Language (DGS) from the German television station Phoenix with the goal to create a limited domain sign language recognition and translation data base. Existing databases such as the RWTH-BOSTON-104 have been distributed to ULg and CRIC for evaluation and extraction of novel features for tracking and recognition.

Main objectives for the first period were

- to establish database snapshots for
  - Corpus-NGT
  - RWTH-PHOENIX-v2.0
  - SIGNUM

- RWTH-BOSTON-104

– to establish baseline results for

- isolated sign language recognition
- continuous sign language recognition

2.2.2 Progress towards objectives

A major activity of RWTH in WP7 has been the evaluation of hand and head tracking features developed within the first year of the SignSpeak project in order to measure the performance of the proposed algorithms w.r.t. word error rate (WER). Furthermore, the databases Corpus-NGT and RWTH-PHOENIX-v2.0 have been significantly extended and prepared for recognition and translation experiments. All gloss annotations used in this section are freely available on request.

2.2.2.1 Introduction

We call the conversion of a video signal (images) into a sequence of written words (text) automatic sign language recognition (ASLR). Our ASLR system is based on Bayes' decision rule: the word sequence w_1^N which best explains the current observation x_1^T given the learned model is the recognition result.

For data capturing we use standard video cameras rather than special data acquisition devices. To model the video signal we use appearance-based features in our baseline prototype which are reduced by principal components analysis (PCA) reduction matrix.

As it is still unclear how sign language words can be split up into sub-word units, e.g. phonemes, suitable for sign language recognition, our corpus (c.f. paragraph 2.1.2.4) is annotated in *glosses*, i.e. whole-word transcriptions, and the system is based on whole-word models. This means for Equation 2 that the phoneme inventory in combination with a pronunciation lexicon is replaced by a word model inventory without a lexicon. Each word model consists of several *pseudo-phonemes* modeling the average word length seen in training. Each such phoneme is modeled by a 3-state left-to-right hidden Markov model (HMM) with three separate Gaussian mixture models (GMMs) and a globally pooled diagonal covariance matrix [16].

Small differences between the appearance and the length of the utterances are compensated by the HMMs, but different pronunciations of a sign must be modelled by separate models, i.e. a different number of states and different GMMs. Therefore, we added pronunciation information to the corpus annotations and adjusted our language models.

The language models aim at representing syntax and semantics of natural language (spoken or written). They are needed in automatic language processing systems that process speech (i.e. spoken language) or language (i.e. written language).

2.2.2.2 System Overview

For purposes of linguistic analysis, signs are generally decomposed into hand shape, orientation, place of articulation, and movement [3] (with important linguistic information also conveyed through non-manual means, i.e., facial expressions and head movements).

In a vision-based ASLR system for continuous sign language, at every time-step $t := 1,\ldots,T$, tracking-based features are extracted at positions $u_1^T := u_1,\ldots,u_T$ in a sequence of images $X_1^T := X_1,\ldots,X_T$. We are searching for an unknown word sequence w_1^N, for which the sequence of features $x_1^T = f(X_1^T, u_1^T)$ best fits to the trained models. Opposed to a recognition of isolated gestures, in continuous sign language recognition we want to maximize the posteriori probability $\Pr(w_1^N|x_1^T)$ over all possible word sequences w_1^N with unknown number of words N. This can be modeled by Bayes' decision rule [3, 16]:

$$x_1^T \longrightarrow \hat{w}_1^N = \arg\max_{w_1^N} \left\{\Pr(w_1^N|x_1^T)\right\} = \arg\max_{w_1^N} \left\{\Pr(w_1^N) \cdot \Pr(x_1^T|w_1^N)\right\} \tag{2}$$

where $\Pr(w_1^N)$ is the a-priori probability for the word sequence w_1^N given by the language model (LM), and $\Pr(x_1^T|w_1^N)$ is the probability of observing features x_1^T given the word sequence w_1^N, referred to as visual model (VM).

Hand and head tracking algorithms for sign language recognition can be evaluated on the one hand w.r.t. a tracking error rate (TER) criterion, but on the other hand w.r.t. the well known word error rate (WER) criterion which consists of errors that are due to deletions, substitutions, and insertions of words. In this part we focus on the evaluation of tracking approaches by a word error rate criterion.

Table 7: RWTH-BOSTON-104 Corpus Statistics

	Training	Test
# sentences	161	40
# running words	710	178
# frames	12422	3324
vocabulary size	103	65
# singletons	27	9
# OOV	-	1

Table 8: RWTH-BOSTON-104 language model perplexities

LM type	PP
zerogram	106.0
unigram	36.8
bigram	6.7
trigram	4.7

2.2.2.3 Corpora

The RWTH-BOSTON-104 corpus[7] is a subset of a much larger database of sign language sentences that were recorded at Boston University for linguistic research [29]. The RWTH-BOSTON-104 corpus consists of 201 sequences, and the vocabulary contains 104 words. The sentences were signed by 3 speakers (2 female, 1 male) and the corpus is split into 161 training and 40 test sequences. An overview on the corpus is given in Table 7: 26% of the training data are singletons, i.e. a “one-shot training” occurs. The sentences have a rather simple structure and therefore the language model perplexity (PP) is low (c.f. Table 8). The test corpus has one out of vocabulary (OOV) word. Obviously, this word cannot be recognized correctly using whole-word models.

Corpus-NGT is a 72h hour corpus of Sign Language of the Netherlands and consists of several different domains. For the SignSpeak project all domains have been analyzed from a speech recognition and translation point of view. It was found that only the two discussion domains were promising to tackle during the SignSpeak project. The main reasons being

- hardly any use of classifier signs
- predefined discussion topics
- largest quantity of annotated data within Corpus-NGT

Unfortunately, Corpus-NGT in its original version provided annotations on gloss level without annotations for sentences boundaries and provided no translations from Sign Language of the Netherlands into Dutch. During the first period of the SignSpeak project glossing of the session of the discussion domains was continued and sentence boundary annotations were added to the existing gloss annotations.

RWTH received gloss and sentence annotations for 129 recorded sessions of two signers discussing on the 8th of July 2010. The derived statistics of this data snapshot are shown in Table 9 in column Current. In total the snapshot contains about 36k glosses out of which are about 2.8k unique and about 2.3k sentence boundary annotations. It turned out that the sentence boundary annotations covered only half the annotated glosses. Due to fact that RWTH received the data on 8th of July and sentence boundaries only covering half of the data, it was not possible for RWTH to finish evaluating the first prototype for continuous sign language recognition for this report. For the second project period RWTH plans to evaluate all developed methods using continuous sign language. To achieve this goal, the partners agreed to use 10% of the data amount of Verbmobil II [27] database as reference for the next evaluation. The goal for first of March is detailed in Table 9.

The refined suitable Corpus-NGT sub-corpus for tracking is also suitable for recognition experiments. It consists of a limited vocabulary (0.1k most frequent glosses, no indexing signs) and is suitable for signer-dependent or multi-signer recognition of isolated signs (c.f. Table 10).

[7] http://www-i6.informatik.rwth-aachen.de/~dreuw/database.html

Table 9: Corpus-NGT status as of 8 July 2010 for right hand glosses

	Current	Goal - 01 March 2011
# sessions	129	≈300
# sentences	2,361	10,000
# glosses in sentences	17,050	80,000
# glosses total	36,598	– ” –
vocabulary size	2,846	$<$ 4,000
# singeltons (rate)	1,308 (46%)	$<$ 1,600 (40%)
avg. sentence length	7.2	$\approx$ 8
avg. TTR	12.8	$\geq$20

Table 10: Corpus-NGT 0.1k snapshot statistics suitable for tracking and recognition of isolated signs

	Training	Devel	Test
# signers	24	22	22
# running glosses	4,235	525	567
vocab. size	101	91	70
# signer	24	24	24
# OOV	–	0	0
# singeltons	0	–	–
avg. TTR	41.9	–	–

Table 11: Detailed word count statistics for Corpus-NGT 0.1k snapshot

Word Counts		
Training	Devel	Test
583 DOOF	13 ZIEN	9 ZWANGER
344 GEBAREN	13 ZELFDE	9 ZO
253 KUNNEN	13 ZELF	9 ZIEN
236 NU	13 ZEGGEN	9 ZELFDE
207 HOREND	13 VROEGER	9 ZELF
171 ZEGGEN	13 VOELEN	9 ZEGGEN
142 ZELF	13 VINDEN	9 WILLEN
138 ZELFDE	13 TOCH	9 WIE
116 ZIEN	13 TAAL	9 WETEN
103 PRATEN	13 SCHOOL	9 WERKEN
100 SCHOOL	13 SAMEN	9 WERELD
74 TAAL	13 S	9 WEG
73 VROEGER	13 PRATEN	9 WAT
58 TOCH	13 PLUS	9 VROEGER
54 S	13 NU	9 VRAGEN
49 SAMEN	13 KUNNEN	9 VOOR
47 VOELEN	13 HOREND	9 VOLGEN
44 PLUS	13 GEBAREN	9 VOELEN
43 VINDEN	13 DOOF	9 VINGERSPELLEN
42 ZO	12 STEL	9 VINDEN
42 WETEN	11 ZO	9 VERTELLEN
42 VEEL	8 PERSOON	9 VERSCHILLEND
42 STEL	7 WETEN	9 VERANDEREN
42 SENSEO	5 VEEL	9 VEEL
42 PERSOON	4 ZWANGER	9 VAN
41 VAN	4 ZWAAR	9 V
38 WILLEN	4 ZORGEN	9 TOLK
33 VERSCHILLEND	4 ZOON	9 TOCH
Continues . . .		

Table 11: (cont.)

Word Counts (cont.)		
Training	Devel	Test
33 T	4 ZELFSTANDIG	9 TIJD
32 ROEPEN	4 WOORDEN	9 TEN-TWEEDE
31 VRAGEN	4 WILLEN	9 TEN-EERSTE
30 WERELD	4 WIE	9 TAAL
27 WAT	4 WERKEN	9 T
27 ROLSTOEL	4 WERELD	9 STEM
26 WEG	4 WEG	9 STEL
26 TEN-EERSTE	4 WAT	9 SENSEO
26 R	4 WAAROM	9 SCHRIJVEN
24 WERKEN	4 VROUW	9 SCHOOL
24 STEM	4 VREEMD	9 SAMEN
22 TEN-TWEEDE	4 VRAGEN	9 S
20 SCHRIJVEN	4 VOORBEELD	9 ROLSTOEL
19 PROBLEEM	4 VOOR	9 ROEPEN
18 TIJD	4 VOLGEN	9 R
18 PRIMA	4 VINGERSPELLEN	9 PROBLEEM
16 TOLK	4 VERTELLEN	9 PRIMA
14 ZWANGER	4 VERSTAND	9 PRATEN
14 WIE	4 VERSCHILLEND	9 PLUS
14 VOOR	4 VERGETEN	9 PERSOON
13 VERTELLEN	4 VERANDEREN	9 NU
12 VERANDEREN	4 VAN	9 KUNNEN
11 ZWEMMEN	4 V	9 HOREND
11 ZWAAR	4 U	9 GEBAREN
11 ZUS	4 TOLK	9 DOOF
11 ZORGEN	4 TOEKOMST	8 WOORDEN
11 ZOON	4 TIJD	8 VREEMD
11 ZITTEN	4 THUIS	8 VERGETEN
11 ZIN	4 TEST	7 WAAROM
11 ZELFSTANDIG	4 TEN-TWEEDE	7 TEGEN
11 WOORDEN	4 TEN-EERSTE	7 SOMS
11 WOORD	4 TEN-DERDE	6 ST-MICHIELSGESTEL
11 WIJ	4 TEGEN	5 VERSTAND
11 WEL	4 T	5 U
11 WEINIG	4 ST-MICHIELSGESTEL	4 ZORGEN
11 WACHTEN	4 STEM	4 ZOON
11 WAAROM	4 SOMS	4 VOORBEELD
11 VROUW	4 SNEL	4 THUIS
11 VRIENDEN	4 SENSEO	4 TEST
11 VREEMD	4 SCHRIJVEN	4 TEN-DERDE
11 VRAAG	4 ROLSTOEL	3 SNEL
11 VOORBEELD	4 ROEPEN	2 VROUW
11 VOLGEN	4 R	
11 VINGERSPELLEN	4 PROBLEEM	
11 VERSTAND	4 PROBEREN	
11 VERSCHILLENDE	4 PRIMA	
11 VERSCHIL	3 ZUS	
11 VERGETEN	3 SLECHT	
11 VADER	3 PRECIES	
11 V	3 PAS	
11 U	2 ZWEMMEN	
11 TWEEDE	2 WOORD	
11 TOLKEN	2 WEINIG	

Continues . . .

Table 11: (cont.)

Word Counts (cont.)		
Training	Devel	Test
11 TOEVALLIG	2 VRIENDEN	
11 TOEKOMST	2 VADER	
11 THUIS	2 TWEEDE	
11 TEST	2 TERUG	
11 TERUG	2 TEGENHOUDEN	
11 TEN-DERDE	2 SPORTEN	
11 TEGENHOUDEN	2 SLURF	
11 TEGEN	1 ZIN	
11 ST-MICHIELSGESTEL	1 WEL	
11 SPORTEN	1 VRAAG	
11 SOMS		
11 SNEL		
11 SLURF		
11 SLECHT		
11 REGELEN		
11 PROBEREN		
11 PRECIES		
11 PRACHTIG		
11 PAS		
$\sum$ 4,235	$\sum$ 525	$\sum$ 567
End		

Data analysis showed that the Corpus-NGT data is dominated by indexing signs. The four most frequent indexing signs make up for 21% of the overall annotated glosses. It showed In preliminary experiments that the indexing signs dominate the overall recognition performance and led to word error rates (c.f. (Equation 3)) above 80% for isolated sign language recognition. Therefore, indexing signs have been excluded from the vocabulary of the snapshot detailed in Table 10.

2.2.2.4 Experimental Results

All developed methods necessary for isolated and continuous sign language recognition are measured, similar to that done for speech recognition, in terms of WER. For isolated sign language recognition the WER is simply the error rate, but for continuous sign language recognition the WER is composed of errors that are due to deletion, insertion, or substitution of words:

$$\text{WER} = \frac{\#\text{deletions} + \#\text{insertions} + \#\text{substitutions}}{\#\text{observations}} \tag{3}$$

RWTH-BOSTON-104 As our corpora are annotated in glosses, i.e. whole-word transcriptions, the system is based on whole-word models. Each word model for the RWTH-BOSTON-104 database consists of one to three pseudo- phonemes modeling the average word length seen in training. Our RWTH-BOSTON-104 lexicon defines 247 pseudo-phonemes for 104 words. Each pseudo-phoneme is modeled by a 3-state left-to-right hidden Markov model (HMM) with three separate Gaussian mixtures (GMM) and a globally pooled covariance matrix.

As a baseline experiment we used simple appearance-based features by downscaling the video frames to 32x32 pixels (i.e. the Frame features), which can be reduced by a PCA transformation to a lower dimensional feature vector (i.e. the Frame+PCA features). The results in Table 12 show that 43 errors are made when using PCA reduced frame features, resulting in a 24.16% WER.

To show the impact of tracking based features, the DPT and AAM based frameworks (c.f. paragraph 2.1.2.3) have been used to extract dominant-hand and head based tracking features. The results in Table 12 suggest that the tracking accuracy itself has only a small impact for the chosen appearance-based hand-patch features, as the 33.71% WER achieved by the DPT approach, which itself achieves an 8.37% tracking error rate (TER) in Table 4, is only slightly worse than the 30.34% WER achieved in a cheating experiment, where we extracted the hand features based on the ground-truth annotation (i.e. a 0% TER). We did not yet evaluate hand features based on the Robust PCA tracking method described in paragraph 2.1.2.3 due to the relatively poor tracking

Table 12: Hand and head features on the RWTH-BOSTON-104 dataset

Tracker	Features	del	ins	sub	errors	WER [%]
-	Frame	39	10	20	69	38.76
-	+ PCA (200)	14	7	22	43	24.16
DPT [15]	Dom. Hand-Patch	27	8	31	66	37.08
DPT [15]	+ PCA (30)	17	13	30	60	33.71
Ground-truth	+ PCA (30)	9	12	33	54	30.34
AAM [33]	AAM Face Coordinates + PCA (30)	26	12	36	74	41.57

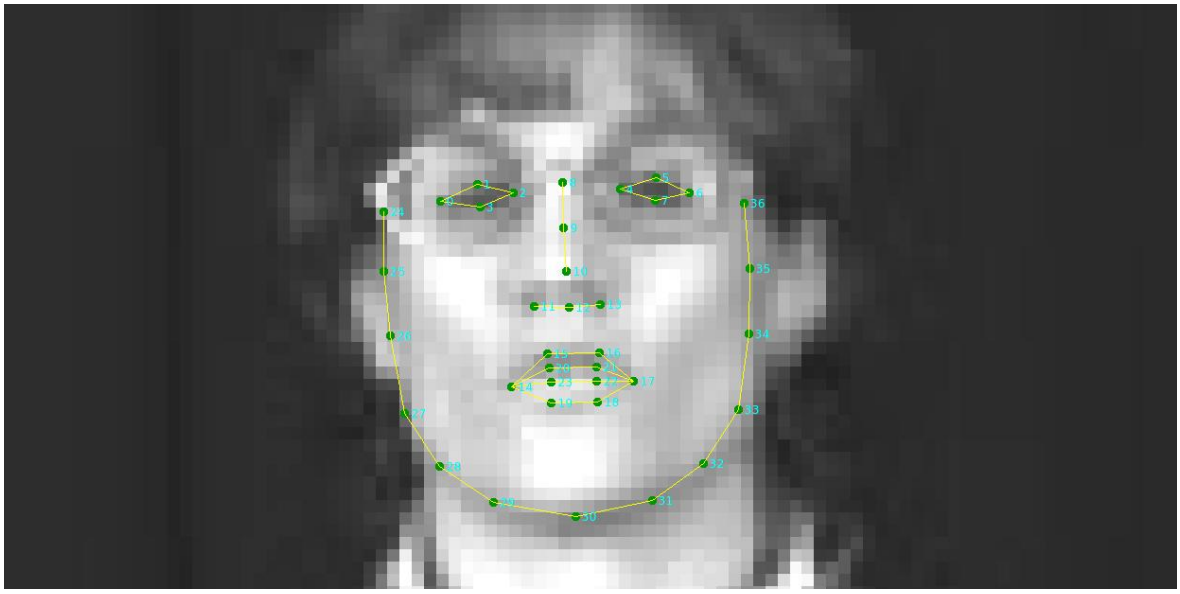

Figure 4: Example of AAM based face coordinate features

performance in Table 4. In the future it will be interesting to analyze more sophisticated hand features and feature combinations.

I another experiment we analyzed the facial features tracked and extracted by the AAM method described in paragraph 2.1.2.3. In informal experiments we simply used the coordinates of the fitted AAM model as shown in Figure 4 as features, and reduced them by a PCA transformation to 30 components. Again, the preliminary results in Table 12 suggest that it will be interesting to analyze more sophisticated head features and feature combinations.

RWTH-PHOENIX-v2.0 The annotation of the RWTH-PHOENIX-v2.0 database is still ongoing. The numbers presented in Table 13 cover the status as of the 30th of June 2010. A preliminary analysis of the data showed that the database did not reach yet a size and consistency necessary to effectively carry out recognition experiments. Therefore, RWTH refrained from conducting experiments on the RWTH-PHOENIX-v2.0 database in the first project period.

ATIS-ISL In the case of the ATIS-ISL database RWTH, preliminary recognition experiments showed poor results. It was found that the ATIS-ISL database contains very difficult lighting conditions, background clutter, and compression artefacts leading to poor results. Therefore, RWTH decided to focus on tracking and preprocessing techniques for the ATIS-ISL database. Tracking results are depicted in Table 6.

SIGNUM The SIGNUM database features conditions similar to the RWTH-BOSTON-104 database. RWTH carried out preliminary speaker dependent recognition experiments to calibrate system parameters as well as to evaluate frame and tracking features.

The results shown in Table 14 use a trigram language model using modified Kneser-Ney smoothing and image intensities as features. As such the results are to be consider as base line results. Because of a language

Table 13: Statistics of the RWTH-PHOENIX-v2.0 database from recognition point of view

	30 June 2010
# sentences	1187
# running words	10626
vocab. size	632
# singletons	242
# signer	6
avg. type-token-ratio	16.81
avg. sentence length	8.95

Table 14: Hand features on the SIGNUM dataset

Tracker	Features	del	ins	sub	errors	WER [%]
-	Frame + PCA (200)	243	161	752	1156	41.26
DPT [15]	Dom. Hand-Patch	209	166	787	1162	41.47

model scale of 250, the recognition system derives its performance mainly from the language model and not from the learned visual models. Further investigation into features and visual modelling will be necessary. RWTH is currently analyzing the trained visual models.

Corpus-NGT 0.1k Snapshot Due to the difficult conditions in the Corpus-NGT database and relatively high tracking error rates reported in Table 5, RWTH can not yet provide meaningful evaluation results for Corpus-NGT in this report. Even for words with a relatively high amount of training data such as `DOOF` (c.f. Table 11) the performance of the baseline system without using high-level features is low. Unfortunately, more sophisticated and high-level features such as robust two-hand tracking and body pose estimation are not yet available within SignSpeak and can not be evaluated.

Experimental results for the Corpus-NGT snapshot described in Table 10 will be presented in the next evaluation report. However the word-count statistics in Table 11 suggest that another subset partitioning of the full Corpus-NGT database might be necessary, as word frequencies in train and test sets are currently unbalanced.

2.2.2.5 Conclusions

The current baseline prototype for sign language recognition allows for the recognition of isolated and continuous sign language data. It offers many configuration possibilities and will allow for the recognition of other languages in the future, such as the Corpus-NGT database (Sign Language of the Netherlands or Nederlandse Gebaren Taal (NGT)) or the RWTH-PHOENIX-v2.0 database (German Sign Language (GSL)).

The prototype can be used now for a variety of video types and languages, and novel features can be easily loaded into the framework. The current prototype has been trained on appearance-based image features, hand- and head-tracking based features. The next steps will include the integration of more sophisticated features and feature combinations.

2.3 Evaluation of the Sign Language Translation (Task 7.3)

The goal of the Sign Language Translation system is to provide automatic translations from glosses to a spoken language. For this, we have worked with the RWTH-PHOENIX-v2.0 (c.f. Table 15) and the Corpus-NGT (c.f. Table 16).

We use an in-house statistical translation system similar to [9]. It is able to process hierarchical phrases in a context-free grammar with a variation of the CYK algorithm. For a given sentence f, the best translation $\hat{e}$ is chosen as the target sentence e that maximizes the sum over m different models h_m, scaled by the factors λ_m:

$$\hat{e} := \underset{e}{\operatorname{argmax}} \left(\sum_m \lambda_m h_m(e, f) \right) . \tag{4}$$

The alignment is created for both translation directions with GIZA++[8] and merged with a variation of the grow-diag-final algorithm. We employ a trigram language model using modified Kneser-Ney discounting which

[8] http://www.htlpr.rwth-aachen.de/~och/software/GIZA++.html

		glosses	German
Train:	Sentences	2711	
	Running Words	15 499	21 679
	Vocabulary	916	1476
	Singletons	337	633
Dev:	Sentences	338	
	Running Words	1924	2689
	Vocabulary	366	547
	OOVs	33	65
	Trigram ppl	19.7	52.4
Test:	Sentences	338	
	Running Words	1750	2629
	Vocabulary	362	517
	OOVs	48	49
	Trigram ppl	20.7	50.8

Table 15: Corpus Statistics for the RWTH-PHOENIX-v2.0 translation corpus

		glosses	Dutch
Train:	Sentences	1061	
	Running Words	7062	11992
	Vocabulary	1130	1574
	Singletons	547	823
Dev:	Sentences	132	
	Running Words	871	1515
	Vocabulary	332	480
	OOVs	77	123
	Trigram ppl	119.9	84.7
Test:	Sentences	132	
	Running Words	969	1612
	Vocabulary	362	489
	OOVs	72	127
	Trigram ppl	143.6	79.3

Table 16: Corpus Statistics for the Corpus-NGT translation corpus

is trained with the SRI toolkit[9]. The scaling factors of the log-linear model are optimized on the development set with Och's Minimum Error Rate Training [30], which is a variation of Powell's method working on n-best translations. The resulting factors are then used to translate the test set.

2.3.1 Results and Outlook

The translation system is up and running. It has been released into open source in [41]. Due to the size of the data, the translation is very fast and real-time compatible. TCP/IP support for remote access is provided.

We evaluate our translation system with the commonly used Translation Edit Rate (TransER)[10] [36], and for the RWTH-PHOENIX-v2.0, we also provide the performance measure BLEU [32]. For Corpus-NGT, the development and testing set are currently still quite small, and we therefore opted to use the word error rate (WER) and the position-independent word error rate (PER) instead, since BLEU is known for misleading results on small test sets.

RWTH-PHOENIX-v2.0 already provides nice results (c.f. Table 17), while the translation set of the Corpus-NGT is still very small with a larger domain. Due to these challenges, the results are encouraging but leave

[9] http://www-speech.sri.com/projects/srilm/

[10] We use an uncommon abbreviation here to distinguish the Translation Edit Rate from the Tracking Error Rate

	BLEU	TransER [%]
German–Glosses	21.4	63.9
Glosses–German	24.3	64.6

Table 17: Translation Results for RWTH-PHOENIX-v2.0

	WER [%]	PER [%]	TransER [%]
Glosses–Dutch	82.7	76.1	82.2

Table 18: Translation Results for Corpus-NGT

much room for improvements (c.f. Table 18). Now that we have a better understanding of the nature of this corpus, this situation is expected to improve soon, especially with new data.

Applying additional linguistic data for Dutch seems to be crucial. We therefore acquired the Corpus Eindhoven[11] with $\approx 600\,000$ running words, which also features spoken dialogue transcription. For syntactic models, the Dutch parser "tadpole" [14] has been selected. We will work them in to our translation system for the upcoming prototype.

3 Objectives for the next Evaluation

Next evaluation will be delivered on month 27 (end of June 2011). The evaluation will be carried out over the extended prototypes generated during the second year of the project: extended prototypes for multimodal visual analysis (D.3.4), extended prototypes for sign language recognition (D.4.2) and extended prototypes for sign language translation (D5.2).

4 References

[1] S. Baker and I. Matthews. Lukas-kanade 20 years on: A unifiying framework. *International Journal of Computer Vision*, 69(3):221255, 2004.

[2] B. Bauer and K.F. Kraiss. Video-based sign recognition using self-organizing subunits. In *International Conference on Pattern Recognition*, pages 434–437, August 2002.

[3] R. Bowden, D. Windridge, T. Kadir, A. Zisserman, and M. Brady. A linguistic feature vector for the visual interpretation of sign language. In *ECCV*, volume 1, pages 390–401, May 2004.

[4] Dr. Gary Rost Bradski and Adrian Kaehler. *Learning opencv, 1st edition*. O'Reilly Media, Inc., 2008.

[11] http://www.inl.nl

number of sentences	topic	number of tokens
6 600	daily newspaper	120 000
6 600	opiniebladen	120 000
7 800	gezinsbladen	120 000
10 000	books and novels	120 000
6 000	popular science	120 000
14 000	free conversations and interviews	120 000

Table 19: Corpus Eindhoven: Monolingual data for Dutch

[5] A. Braffort. Argo: An architecture for sign language recognition and interpretation. In *International Gesture Workshop: Progress in Gestural Interaction*, pages 17–30, April 1996.

[6] P. Buehler, M. Everingham, D. P. Huttenlocher, and A. Zisserman. Long term arm and hand tracking for continuous sign language TV broadcasts. In *Proceedings of the British Machine Vision Conference*, 2008.

[7] Jan Bungeroth, Daniel Stein, Philippe Dreuw, Hermann Ney, Sara Morrissey, Andy Way, and Lynette van Zijl. The ATIS Sign Language Corpus. In *LREC*, Marrakech, Morocco, May 2008.

[8] K. Cheung, S. Baker, and T. Kanade. Shape-from-silhouette across time part i: Theory and algorithms. *International Journal on Computer Vision*, 62(3):221–247, 2005.

[9] David Chiang. A Hierarchical Phrase-Based Model for Statistical Machine Translation. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 263–270, Ann Arbor, Michigan, USA, June 2005.

[10] Dorin Comaniciu, Visvanathan Ramesh, and Peter Meer. Kernel-based object tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25:564–577, 2003.

[11] Onno Crasborn, Inge Zwitserlood, and Johan Ros. Corpus-ngt. an open access digital corpus of movies with annotations of sign language of the netherlands. Technical report, Centre for Language Studies, Radboud University Nijmegen, 2008. `http://www.corpusngt.nl`.

[12] D. Cremers, M. Rousson, and R. Deriche. A review of statistical approaches to level set segmentation: Integrating color, texture, motion and shape. *International Journal of Computer Vision*, 72(2):195215, April 2007.

[13] Farhad Dadgostar and Abdolhossein Sarrafzadeh. An adaptive real-time skin detector based on hue thresholding: A comparison on two motion tracking methods. *Pattern Recogn. Lett.*, 27(12):1342–1352, 2006.

[14] A. Van den Bosch, G.J. Busser, W. Daelemans, and S. Canisius. An efficient memory-based morphosyntactic tagger and parser for dutch. In F. van Eynde, P. Dirix, I. Schuurman, and V. Vandeghinste, editors, *Selected Papers of the 17th Computational Linguistics in the Netherlands Meeting*, pages 99–114, Leuven, Belgium, 2007.

[15] P. Dreuw, T. Deselaers, D. Rybach, D. Keysers, and H. Ney. Tracking using dynamic programming for appearance-based sign language recognition. In *IEEE Automatic Face and Gesture Recognition*, pages 293–298, Southampton, April 2006.

[16] P. Dreuw, D. Rybach, T. Deselaers, M. Zahedi, and H. Ney. Speech recognition techniques for a sign language recognition system. In *Interspeech*, Antwerp, Belgium, August 2007. Best paper award.

[17] P. Dreuw, D. Stein, and H. Ney. Enhancing a sign language translation system with vision-based features. In *Intl. Workshop on Gesture in HCI and Simulation 2007*, pages 18–19, Lisbon, Portugal, May 2007.

[18] Philippe Dreuw, Jens Forster, Thomas Deselaers, and Hermann Ney. Efficient approximations to model-based joint tracking and recognition of continuous sign language. In *IEEE International Conference Automatic Face and Gesture Recognition*, Amsterdam, The Netherlands, September 2008.

[19] Philippe Dreuw, Carol Neidle, Vassilis Athitsos, Stan Sclaroff, and Hermann Ney. Benchmark databases for video-based automatic sign language recognition. In *LREC*, Marrakech, Morocco, May 2008.

[20] Philippe Dreuw, Hermann Ney, Gregorio Martinez, Onno Crasborn, Justus Piater, Jose Miguel Moya, and Mark Wheatley. The signspeak project - bridging the gap between signers and speakers. In *International Conference on Language Resources and Evaluation*, Valletta, Malta, May 2010.

[21] Philippe Dreuw, Daniel Stein, Thomas Deselaers, David Rybach, Morteza Zahedi, Jan Bungeroth, and Hermann Ney. Spoken language processing techniques for sign language recognition and translation. *Technology and Dissability*, 20(2):121–133, June 2008.

[22] Gaolin Fang, Wen Gao, and Debin Zhao. Large-vocabulary continuous sign language recognition based on transition-movement models. *IEEE Trans. on Systems, Man, and Cybernetics*, 37(1), January 2007.

[23] Jens Forster, Daniel Stein, Ellen Ormel, Onno Crasborn, and Hermann Ney. Best practice for sign language data collections regarding the needs of data-driven recognition and translation. In *4th LREC Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies (CSLT)*, Malta, May 2010.

[24] Dariu Gavrila. The visual analysis of human movement: A survey. *Computer Vision and Image Understanding*, 73(1):82–98, 1999.

[25] H. Grabner, P. M. Roth, and H. Bischof. Is pedestrian detection really a hard task? In *Tenth IEEE International Workshop on Performance Evaluation of Tracking and Surveillance*, October 2007.

[26] E.-J. Holden, G. Lee, and R. Owens. Australian sign language recognition. In *Machine Vision and Applications*, volume 16, pages 312–320, 2005.

[27] Stephan Kanthak, Achim Sixtus, Sirko Molau, Ralf Schlüter, and Hermann Ney. *Fast Search for Large Vocabulary Speech Recognition*, chapter "From Speech Input to Augmented Word Lattices", pages 63–78. Springer Verlag, Berlin, Heidelberg, New York, July 2000.

[28] Arne Mauser, Richard Zens, Evgeny Matusov, Saša Hasan, and Hermann Ney. The RWTH Statistical Machine Translation System for the IWSLT 2006 evaluation. In *IWSLT*, pages 103–110, Kyoto, Japan, November 2006. Best Paper Award.

[29] C. Neidle, J. Kegl, D. MacLaughlin, B. Bahan, and R.G. Lee. *The Syntax of American Sign Language*. MIT Press, 1999.

[30] Franz Josef Och. Minimum Error Rate Training for Statistical Machine Translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 160–167, Sapporo, Japan, July 2003.

[31] S. Ong and S. Ranganath. Automatic sign language analysis: A survey and the future beyond lexical meaning. *IEEE Trans. PAMI*, 27(6):873–891, June 2005.

[32] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002.

[33] Justus Piater, Thomas Hoyoux, and Wei Du. Video analysis for continuous sign language recognition. In *4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies*, pages 192–195, Valletta, Malta, May 2010.

[34] S. Sarkar, P.J. Phillips, Z.Y. Liu, I.R. Vega, P.J. Grother, and K.W. Bowyer. The humanid gait challenge problem: Data sets, performance, and analysis. *PAMI*, 27(2):162–177, February 2005.

[35] Bernt Schiele. Model-free tracking of cars and people based on color regions. *Image Vision Computing*, 24(11):1172–1178, 2006.

[36] Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas*, pages 223–231, Cambridge, Massachusetts, USA, August 2006.

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

[38] D. Stein, J. Bungeroth, and H. Ney. Morpho-Syntax Based Statistical Methods for Sign Language Translation. In *11th EAMT*, pages 169–177, Oslo, Norway, June 2006.

[39] D. Stein, P. Dreuw, H. Ney, S. Morrissey, and A. Way. Hand in Hand: Automatic Sign Language to Speech Translation. In *The 11th Conference on Theoretical and Methodological Issues in Machine Translation*, Skoevde, Sweden, September 2007.

[40] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1):71–86, January 1991.

[41] David Vilar, Daniel Stein, Matthias Huck, and Hermann Ney. Jane: Open source hierarchical translation, extended with reordering and lexicon models. In *ACL 2010 Joint Fifth Workshop on Statistical Machine Translation and Metrics MATR*, pages 262–270, Uppsala, Sweden, July 2010.

[42] P. Viola and M.J. Jones. Robust real-time face detection. *International Journal of Computer Vision*, 57(2):137–154, 2004.

[43] C. Vogler and D. Metaxas. A framework for recognizing the simultaneous aspects of american sign language. *CVIU*, 81(3):358–384, March 2001.

[44] U. von Agris and K.-F. Kraiss. Towards a video corpus for signer-independent continuous sign language recognition. In *Gesture in Human-Computer Interaction and Simulation*, Lisbon, Portugal, May 2007.

[45] S. B. Wang, A. Quattoni, L.-P. Morency, D. Demirdjian, and T. Darrell. Hidden conditional random fields for gesture recognition. In *CVPR*, volume 2, pages 1521–1527, New York, USA, June 2006.

[46] G. Yao, H. Yao, X. Liu, and F. Jiang. Real time large vocabulary continuous sign language recognition based on op/viterbi algorithm. In *ICPR*, volume 3, pages 312–315, Hong Kong, August 2006.

[47] Morteza Zahedi, Philippe Dreuw, David Rybach, Thomas Deselaers, Jan Bungeroth, and Hermann Ney. Continuous sign language recognition - approaches from speech recognition and available data resources. In *LREC Workshop on the Representation and Processing of Sign Languages: Lexicographic Matters and Didactic Scenarios*, pages 21–24, Genoa, Italy, May 2006.

5 Glossary

ASLR	automatic sign language recognition
GMM	Gaussian mixture model
GSL	German Sign Language
HMM	hidden Markov model
NGT	Nederlandse Gebaren Taal
OOV	out of vocabulary
PCA	principal components analysis
PP	language model perplexity
TER	tracking error rate
WER	word error rate

Acknowledgments.

This work received funding from the European Community's Seventh Framework Programme under grant agreement number 231424 (FP7-ICT-2007-3).