International Workshop on Spoken Language Translation (IWSLT), pp. 197-203, December 2013

Using Viseme Recognition to Improve a Sign Language Translation System

Christoph Schmidt, Oscar Koller, Hermann Ney [1]
Thomas Hoyoux, Justus Piater [2]

[1]Human Language Technology and Pattern Recognition Group,
RWTH Aachen University, Aachen, Germany
{surname}@cs.rwth-aachen.de

[2]Intelligent and Interactive Systems,
University of Innsbruck, Austria
{firstname}.{surname}@uibk.ac.at

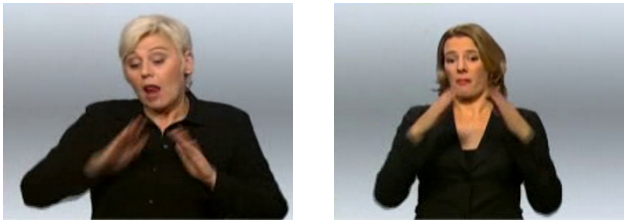
Abstract

Sign language-to-text translation systems are similar to spoken language translation systems in that they consist of a recognition phase and a translation phase. First, the video of a person signing is transformed into a transcription of the signs, which is then translated into the text of a spoken language. One distinctive feature of sign languages is their multi-modal nature, as they can express meaning simultaneously via hand movements, body posture and facial expressions. In some sign languages, certain signs are accompanied by mouthings, i.e. the person silently pronounces the word while signing. In this work, we closely integrate a recognition and translation framework by adding a viseme recognizer ("lip reading system") based on an active appearance model and by optimizing the recognition system to improve the translation output. The system outperforms the standard approach of separate recognition and translation.

1. Introduction

The aim of a sign language-to-text translation system is to translate a video of a person signing into a text in a spoken language. Similar to spoken language translation systems, such a system consists of a recognition component in which the individual signs are recognized, and a translation component in which the sequence of signs is translated into a text of the spoken language. The translation step is necessary as signed languages, if evolved naturally, differ at great length from spoken languages, having a unique grammar and vocabulary.

Sign languages are multi-modal in the sense that they express meaning simultaneously via different communication channels. Besides the manual information such as hand shape, orientation and movements, non-manual aspects such as body posture and facial expressions play a vital role in expressing meaning. In countries which have a strong oral education tradition, e.g. Germany, some signs are accompanied by mouthings, i.e. the signer pronounces the spoken language word with his lips while signing with his hands. These mouthings are particularly used to derive new signs by using the hand movements of a similar or more general sign and changing only the mouthing. In the example in Figure 1, the particular sign for the Alps is derived by depicting the form of a mountain while silently pronouncing the word Alps (German: "Alpen").



ALPS (mouthing "Alpen") MOUNTAIN (mouthing "Berg")

Figure 1: Two signs with the same manual component, differing only in the mouthing. At the time of the snapshots, the underlined letters are pronounced.

In this work, we want to use a mouthing recognition system, which is often also referred to as a visual speech recognition system, to improve the quality of the translation system by providing the mouthing as an additional input to the translation system and by exploiting the correspondence between the mouthings and spoken language words. Moreover, we achieve a close integration of recognition and translation by optimizing the recognition system with respect to the translation output. The approach is depicted in Figure 2.

This paper is structured as follows: First, we present related work in Section 2. The RWTH-Phoenix-Weather corpus, which is used in our experiments, is described in Section 3. In Section 4, we outline the technique of active appearance models, which we apply to track the face and the mouth region. The mouth shape and opening is then used to recognize viseme sequences in Section 5. We present our experimental results in Section 6. Conclusions and an outlook are given in

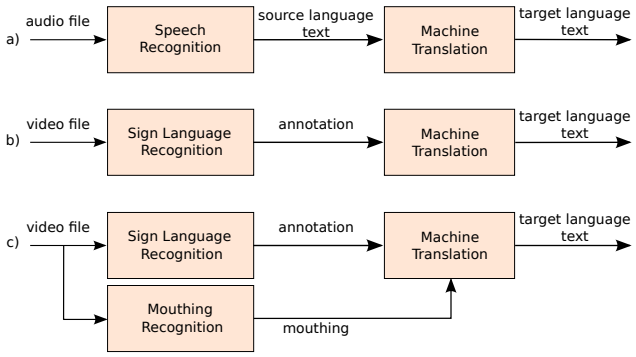


Figure 2: System architectures of a) spoken language translation, b) sign language translation and c) the system proposed in this work including a mouthing recognition

Section 7.

2. Related Work

To track the position of the face and the mouth, we apply an active appearance model. From the resulting locations of the mouth corners, we calculate high-level features such as the degree of opening, which we use to train a viseme recognition system.

[1] and [2] use active appearance models to recognize a predefined set of facial expressions. [3] and [4] provide facial features for the use in a sign language recognition framework, i.e. they integrate low-level facial features into their system to improve the recognition of the signs.

There are several approaches to viseme recognition. We follow the geometric approach, using distances between lips, chin and nose to train a recognition system. This approach is similar to [5], who use active contours ("snakes") to detect the lips. Their approach is more sophisticated, as they calculate histograms of the area inside the lips to detect tongues and teeth, while our approach is more general in that the active appearance models which are trained on the whole face also detect other facial features such as eyebrow raise and cheek movements.

Sign language machine translation faces the challenge that corpus resources are particularly sparse. A thorough overview of sign language translation is given in [6].

3. The RWTH-Phoenix-Weather Corpus

The RWTH-Phoenix-Weather corpus is a video-based, large vocabulary corpus of German Sign Language recorded and annotated for the use in statistical pattern recognition and statistical machine translation. The public TV broadcasting station Phoenix regularly broadcasts the major public news programs with an additional interpretation into German Sign Language using an overlay window which shows the interpreter.

The RWTH-Phoenix-Weather corpus contains the weather forecast portions of these news programs, which were manually annotated by a deaf expert and revised by a hard-of-hearing expert. The weather forecasts were chosen because weather forecasting forms a rather compact domain with a limited vocabulary. A complex domain such as news programs would require a much larger corpus to reliably estimate statistical models, but annotating such a corpus was infeasible due to time and budget constraints.

Since sign language expresses meaning simultaneously via hand movements, body posture, facial expressions, mouthing, etc., one open question in the sign language research community is how to capture this multi-modal nature of sign languages in a comprehensive annotation system. A simple annotation method is gloss annotation, where a sign is annotated by one or several words which roughly correspond to its meaning, usually written in the stem form in upper case. Since the same sign can have several meanings in different contexts, it can be transcribed differently depending on its context. In contrast to this, the term ID-gloss [7] is used if one sign is always annotated with the same gloss, independent of its meaning in a particular context. In our corpus and experiments, we use ID-glosses.

The annotation of a sign language video corpus highly depends on the task at hand. For example, if a linguist wants to study certain linguistic patterns, the annotation should be detailed with respect to these patterns. In the same way, an annotation suitable for an automatic sign language recognition system should be tailored according to the features which the system can actually recognize. Since the RWTH-Phoenix-Weather corpus was originally developed for the recognition of hand-based features, both the time boundaries of the ID-glosses and their label were mainly based on the signing hands. This means that signs which are identical in the hand components but differ in their mouthing received the same label. For example, the sign of a specific mountain is formed by mouthing its name and performing the general sign for mountain with the hands (see Figure 1). In the corpus, both variants were glossed as "MOUNTAIN", because they could not be distinguished by the hand features used at the time.

In addition to the annotation of the ID-glosses, the RWTH-Phoenix-Weather corpus has been marked with time boundaries on the sentence as well as the gloss level. The spoken German weather forecast has been transcribed semi-automatically using a state-of-the-art automatic speech recognition system. To train active appearance models on this corpus, facial landmarks have been manually labeled on a small set of images.

In the following, we will briefly describe the corpus setup and statistics. For a more thorough description see [8]. Note that in this setup, we use only the portions of RWTH-Phoenix-Weather for which time boundaries for individual glosses are annotated. These are necessary to extract the features for the viseme recognizer.

	DGS	German
# signers	7	
# editions	190	
duration[h]	3.25	
# frames	293,077	
# sentences	2,552	
# running glosses	14,771	30,860
vocabulary size	911	1,452
# singletons	120	337

Table 1: Statistics of the RWTH-Phoenix-Weather corpus for DGS and announcements in spoken German

Figure 3: Visualization of facial annotations

The corpus statistics for the RWTH-Phoenix-Weather corpus with time boundaries for individual glosses can be found in Table 1. The database features a total of seven interpreters, consists of 2640 sentences and a total of 14,771 running glosses. Baseline translation results both from German to German Sign Language and in the opposite direction can be found in [9]. Sign language corpora are much smaller than spoken language corpora for two reasons. Since there is no standard writing system for sign languages, sign language corpora containing a written notation do not exist by themselves but have to be produced by experts who define a suitable annotation scheme for the task at hand. Moreover, annotating a video corpus is quite time consuming, because the annotators have to mark time boundaries of individual signs and have to use a canonical notation for sign variants which are frequent.

To train active appearance models on this corpus, 38 facial landmarks for all seven interpreters have been labeled in a total of 369 images (that is, about 50 images per interpreter). Care was taken in selecting a set of images which contain many different expressions, including extreme ones, such that the trained models can approximately represent a large span of expressions for each interpreter. Two examples of the facial annotations are shown in Figure 3.

4. Active Appearance Models

The facial features which are used for recognizing the signer's mouthing consist of continuous measurements of some quantities related to mouthing, such as horizontal and vertical mouth openness, and other facial cues such as eye brow raise. As shown in Table 2, these measurements are based on lower-level facial features which are defined as a set of consistent, salient point locations on the interpreter's face. As illustrated in Figure 3, these fiducial points – also called landmarks – correspond to key locations on the cheeks and chin outlines, the nose ridge and nose base, the eyelids and eye corners, the eyebrow outlines and the lip and mouth corners. We wish to track those point features accurately in the sign language videos in order to extract the higher-level facial features which will in turn be used to recognizing the words pronounced by the signer. Since the structure of the human face as described by a set of such point features exhibits a lot of variability due to changes in pose and expression, we chose to base our tracking strategy on the deformable model registration method known as active appearance models.

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

Table 2: High-level facial features used in the proposed clustering approach and the related lower-level point features (Figure 3)

Active appearance models (AAMs), first proposed in [10] and notably reformulated in [11], are a popular instance of the family of deformable model methods for image interpretation. Such model-based methods attempt to recover an object's structure as it appears in an image by registering a deformable shape model of the object to the image data. Mathematically, the shape $\mathbf{s}$ of an object is defined as the vector of stacked coordinates of its v landmark points:

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

assuming here that each landmark is a 2-dimensional point representing a semantically meaningful part of the object, such as an eye corner in the human face.

AAMs model shape deformation using a so-called point density model (PDM), which is a parametric linear subspace model learned statistically by principal component analysis (PCA) on a set of training shape examples. These examples are given as expert annotations of images of the object of interest, such as shown in Figure 3 for the human face. In such a representation, any shape $\mathbf{s}$ of the deformable object can be expressed by the generative model as a base shape $\mathbf{s}_0$ plus a linear combination of n shape vectors $\mathbf{s}_i$:

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

Registering a PDM to the image data then reduces to finding the optimal coefficient values p_i of this linear combination,

i.e. the optimal PDM's parameters. AAMs propose to model the coupling between the PDM and the image data, i.e. the predictions on the PDM's landmarks locations given a target image, via a holistic appearance model of the pixel intensity values of the object's image. This appearance model is again a parametric linear subspace model, obtained by applying PCA to shape-normalized training example images of the object of interest. This shape normalization involves the warping of every example image to a reference frame, which is typically done by piecewise affine warping functions defined between each example shape and the base shape $\mathbf{s}_0$ of the PDM. The generative appearance model is then used to express any object's appearance $A(\mathbf{x})$ as a base appearance $A_0(\mathbf{x})$ plus a linear combination of m appearance images $A_i(\mathbf{x})$:

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

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

Given these two generative models and following the so-called "independent" AAMs formulation proposed in [11], registration can be seen as an image matching problem between the synthetic model image and the shape-normalized target image; the fitting goal can therefore be expressed as finding the parameters $\mathbf{p} = (p_1, p_2, \ldots, p_n)^\intercal$ and $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_m)^\intercal$ that minimize the following sum of squared differences:

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

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

Variants met in the AAM-related literature mostly differ in the way they parameterize this linear approximation to derive the parameters update equation. In this work, we chose to use the efficient version of the simultaneous inverse-compositional AAM (SICAAM) proposed in [12]. This variant is more robust than others to large variations in shape and appearance, which typically occur when dealing with facial expressions in the context of sign language. Moreover, in order to cope with large off-plane head rotations, which are also common in sign language and can lead a 2D AAM to failure, we used the refinement proposed in [13]. In this work, a 3D PDM is estimated using a non-rigid structure-from-motion algorithm on the training shapes, and is then involved in the optimization process which incorporates a regularization term encouraging the 2D shape controlled by the 2D PDM to be a valid projection of the 3D PDM. Similar to the 2D PDM, the 3D PDM expresses any 3D shape $\mathbf{S}$ as a 3D base shape $\mathbf{S}_0$ plus a linear combination of $\bar{n}$ 3D shape vectors $\mathbf{S}_i$:

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

Notice that the 3D PDM is also involved in the calculation of the high-level facial features described below.

The procedure for the production of the high-level facial features includes a training stage:

1. Extrude the set of 2D training shape examples to 3D by means of the 3D PDM.
2. Remove global translations and rotations by aligning every extruded shape to the base shape $\mathbf{S}_0$ of the 3D PDM.
3. Project the aligned extruded shapes to 2D and, for each, estimate local area-based measurements corresponding to the point features subsets given in Table 2.
4. For each point features subset, store as the training output the minimum and maximum values of the corresponding local area-based measurements.

Extracting high-level facial features from the tracked lower-level point features is then done in the following way:

1. Extrude the registered shape and remove its global translation and rotation by means of the 3D PDM
2. Project the aligned extruded shape to 2D and, for each point features subset given in Table 2, estimate the corresponding local area-based measurement.
3. Normalize each local area-based measurement between 0 and 1 according to the minimum and maximum values obtained during training for the corresponding point features subset.
4. Each registered shape is then associated with a vector of D (in our work $D = 7$) continuous values in the range $[0, 1]$, corresponding to our high-level facial features.

Seven SICAAMs specific to the seven interpreters of RWTH-Phoenix-Weather have been trained for the end purpose of extracting high-level facial features from the gloss-annotated videos as shown in Figure 4. Training and tracking with one single SICAAM for all seven interpreters would have been a viable choice as well because of the enhanced robustness of this AAM variant to variability in identity. However, we wanted to obtain the best possible accuracy in the tracking of the low-level point features. On the other hand, the calculation of our high-level features is rather sensitive to identity changes and as such had to be designed in an identity-dependent fashion. The extraction of reliable

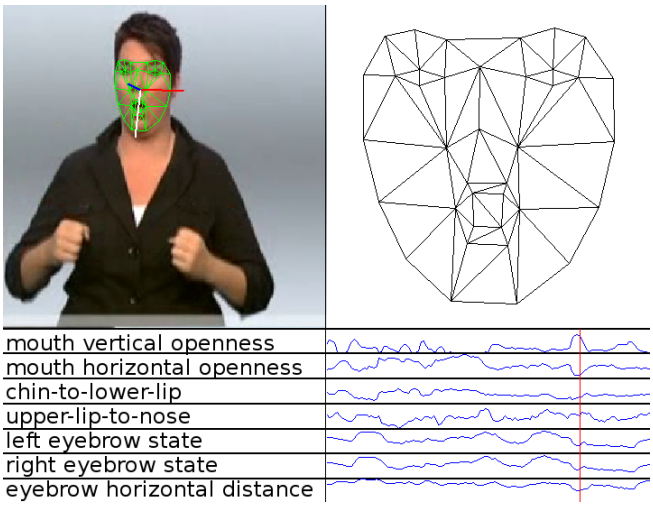


Figure 4: High-level feature extraction
Top left: the grid of fitted AAM points
Top right: rotated and normalized AAM points
Bottom: high-level feature values over time

identity-*independent* facial features similar to those used in this work is part of the advanced computer vision research topic known as "expression transfer" and is beyond the scope of this paper, where our primary goal is to give a proof of concept that including the mouthing information from a viseme-based mouthing recognizer can improve a sign language translation system. The mouthing recognizer will be described in more detail in the next section.

5. Viseme Recognition

Since the RWTH-Phoenix-Weather corpus was mainly annotated for the use in sign language recognition of hand-based features, mouthings have not been annotated for the whole corpus. To obtain possible candidates for the words the signer has pronounced while signing, we align the glosses denoting the signs with their translation in the spoken language. We use the open-source toolkit GIZA++ to align each gloss to at most one word. However, not all signs are accompanied by mouthing. We therefore include a silence model representing no mouth movement and a garbage model for mouthing gestures not representing specific viseme sequences in the viseme recognizer. To train a viseme recognizer on the videos, we need a viseme transcription of the spoken words. We first use a lexicon from our speech recognition system trained on German to lookup each German word which is aligned to a gloss and to find its corresponding sequence of phonemes. As many phonemes cannot be visually distinguished, for example the phonemes P and B differ only in the aspiration which is not visible, we further map the set of phonemes to a set of visemes, i.e. visually distinguishable phonemes. We follow the suggestion of [14] and map the set of phonemes to a set of 15 visemes. A list of the used visemes can be found in Table 3.

Phoneme	Viseme	Examples
p, b	P	Pause, Bitte
t, d, k, g	T	Tonne, Dach, König, Gier
n, @n, l, @l	N	Nadel, raten, Liebe, Igel
m	M	Mutter
f, v	F	Finder, Vase
s, z	S	Fass, Stein
S, Z, tS, dZ	Z	Schein, Garage, Tscheche
h, r, x, N	R	Hase, Reden, Dach, Wange
j, C	C	Junge, Wicht
i:, I, e:, E:, E	E	Bier, Tisch, Weg, Räte, Menge
a:, a	A	Wagen, Watte
o:, O	O	Wolle, Wogen
u:, U	U	Buch, Runde
@, 6	Q	Bitte, Weiher
y:, Y, 2:, 9	Y	Tür, Mütter, Goethe, Götter

Table 3: Phoneme-viseme mapping (taken from [14])

Statistics on the aligned gloss translation pairs allow to exclude noisy alignments. Specifically, this is done by using an empirically set threshold of at least four occurences per gloss translation pair and considering only translation alignments that represent at least 10% of all translations for a specific gloss. Gloss translation alignments which do not meet these requirements are put into the garbage model.

We then train our state-of-the art speech recognition system RASR [15] using 15 viseme hidden Markov models (HMMs) and the garbage model, each containing three states with single Gaussian densities, a globally pooled covariance matrix and global time distortion penalties. Silence visemes are represented by an additional single state HMM. The models are fed with the seven high-level facial features. A modelling lexicon defining possible pronounciation variants for each gloss is provided to the system. It is generated based on the statistics on the aligned gloss translation pairs. The system is initialized with a linear segmentation on the RWTH-Phoenix-Weather data providing gloss time boundaries. The EM-algorithm with viterbi approximation iteratively accumulates the HMMs and uses them to re-estimate the state-frame-alignment, while choosing the most likely pronounciation variants representing different sequences of visemes. This process can be considered as weakly supervised clustering. After 10 iterations the algorithm converges to a stable optimum, yielding the hypothesized viseme sequences for each gloss. In order to remove outliers we chose the RANSAC algorithm [16] to further refine the state-frame-alignment and hence the models.

Table 5 shows the achieved performance of the viseme recognizer after each of its training and refinement steps. The Character Error Rate (CER) compares the hypothesized viseme sequence on the character level to 640 manually annotated mouthings.

Subsequently, the hypothesized viseme sequences are fil-

	CER	Recall
initial segmentation	40.5	82.5
10x EM-realignment	35.7	47.5
after RANSAC processing	32.2	45.5

Table 4: Character Error Rate (CER) and recall in [%] of viseme recognizer measured on 640 manual annotations.

tered by comparing them to the original GIZA++ alignment and estimating the relative error for a given gloss and viseme sequence. Viseme sequences that cause a high mismatch to the GIZA++ alignment are less likely to support the following translation step. We tested different error thresholds on the development set and obtained best results for a threshold of 30. Translation variants with a relative error higher were removed, that is, no gloss variant was generated.

6. Experiments

For our experiments, We use the open-source translation system JANE [17]. The training corpus is word-aligned using GIZA++, and phrase pairs consistent with this alignment are extracted. Previous experiments on this corpus ([9]) have shown that phrase-based systems outperform hierarchical systems, and consequently we choose a phrase-based system for machine translation. Since the corpus is very small, regular MERT training on a held-out development set leads to unstable optimization parameters. We therefore apply a technique similar to cross-validation where we train five different systems, each with a different portion of the training data used as the development set. In each optimization iteration, we concatenate the n-best lists of each individual system and optimize the parameters on this concatenated list.

The baseline system consists of a two-stage approach in which the glosses with no additional information are translated. This corresponds to part b) in Figure 2.

In the approach proposed in this work, which is depicted in part c) of the same figure, we add the mouthing information obtained from the viseme recognizer as an additional knowledge source to the translation system. This is done in the following way. In cases in which the viseme recognizer has a high confidence to recognize a word correctly, we split up the gloss into several variants. E.g., the gloss MOUNTAIN(="BERG") from Figure 1 could be split up into two gloss variants MOUNTAIN_alps and MOUNTAIN_mountain. The machine translation system is then trained on these gloss variants.

Since the mouthing usually corresponds to a word in the spoken language, we want to increase the probability of the gloss variants which are translated into their mouthing component. This can be done on the word and the phrase level.

On the word level, we increase the probability of the IBM1-like lexical smoothing of such pairs by a factor α. The factor is optimized on the development set.

System	Dev		Test	
	BLEU	TER	BLEU	TER
Baseline	35.5	58.8	23.8	66.5
Oracle	36.8	53.4	29.8	60.1
+ word level	39.8	45.3	31.7	52.7
+ phrase level	40.8	43.6	32.6	49.9
+ word + phrase level	41.1	44.4	33.6	48.7

Table 5: Oracle machine translation results, assuming all mouthings were recognized correctly

On the phrase level, we add binary as well as count features to the phrase table, indicating whether a gloss with a certain mouthing is translated into the corresponding spoken word (boolean feature) or counting the number of glosses in the phrase for which this is true (count feature).

Thus, the computer would e.g. learn to translate the gloss variant MOUNTAIN_alps (which consists of the manual sign for mountain, accompanied by the mouthing "Alps") into the German word for Alps. We refrained from hard-wiring these connection for two reasons. First, the viseme recognition also contains errors, which can partly be learnt by the machine translation system during training. Moreover, mouthings usually use the base form of the word without inflections, and thus the same mouthing can result in different inflections in the spoken language.

First we examine oracle translation results which assume that all mouthings have been recognized correctly. These results form an upper bound on the translation performance of the actual system and show the potential of adding the mouthing information to the system. The results can be seen in Table 5. Training a phrase-based system on the gloss-variants increases the system performance by 6 BLEU and 6.4 TER. Additional gains can be optained by increasing the probabilities of matching mouthings and translations on the word and phrase level. The best performance can be obtained by combining both of these models.

The translation result of the whole pipeline of viseme recognition and translation system is given in Table 6. Training the machine translation system on the gloss variants produced by the viseme recognizer leads to a degragation in BLEU, but TER is improved. Increasing the weight of corresponding mouthing and translation pairs either on the word or the phrase level leads to an improvement. Combining both models only slightly improves the BLEU score.

7. Conclusions / Outlook

In this paper, we propose the integration of a viseme recognizer into a sign language translation framework. Instead of using the facial features in the recognition phase, we opt for using the mouthing information as an additional knowledge source in the translation system. The system is able to out-

System	Dev		Test	
	BLEU	TER	BLEU	TER
Baseline System	35.5	58.8	23.8	66.5
Viseme + MT System	35.2	53.2	23.1	65.4
+ word level	36.1	54.3	24.1	65.5
+ phrase level	36.8	53.5	24.4	64.4
+ word + phrase level	37.5	52.6	24.8	64.4

Table 6: Machine translation results of systems including viseme recognition input

perform the baseline system which only translates the manual information of the signs. The use of mouthing information is especially useful in countries which have an oralist education tradition. In other countries, e.g. the US, fingerspelling is used more heavily.

In the future, we want to improve the quality of the viseme recognition by including a histogram of the mouth area. This can lead to improvements for visemes with distinct tongue or teeth configurations. Moreover, we want to incorporate other modalities besides the hands and the mouthing as well. One problem which we encountered during the experiments is the spreading of the mouthing, i.e. the mouthing is not synchronous to the hands but starts later. We want to address this issue using dynamic time alignment.

8. References

[1] I. Ari, A. Uyar, and L. Akarun, "Facial feature tracking and expression recognition for sign language," in *Computer and Information Sciences, 2008. ISCIS'08. 23rd International Symposium on*, 2008, pp. 1–6.

[2] I. Rodomagoulakis, S. Theodorakis, V. Pitsikalis, and P. Maragos, "Experiments on global and local active appearance models for analysis of sign language facial expressions," in *9th International Gesture Workshop on Gestures in Embodied Communication and Human-Computer Interaction*, 2011, pp. 96–99.

[3] J. Piater, T. Hoyoux, and W. Du, "Video analysis for continuous sign language recognition," in *Proceedings of 4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies*, 2010, pp. 22–23.

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

[5] S. Werda, W. Mahdi, and A. B. Hamadou, "Lip localization and viseme classification for visual speech recognition," *CoRR*, vol. abs/1301.4558, 2013.

[6] S. Morrissey and A. Way, "Manual labour: tackling machine translation for sign languages," *Machine Translation*, vol. 27, no. 1, pp. 25–64, 2013. [Online]. Available: http://dx.doi.org/10.1007/s10590-012-9133-1

[7] T. Johnston, "The lexical database of auslan (australian sign language)," *Sign Language and Linguistics*, vol. 4, pp. 145–169(25), 2001.

[8] J. Forster, C. Schmidt, T. Hoyoux, O. Koller, U. Zelle, J. Piater, and H. Ney, "Rwth-phoenix-weather: A large vocabulary sign language recognition and translation corpus," in *International Conference on Language Resources and Evaluation*, Istanbul, Turkey, May 2012. [Online]. Available: http://www.lrec-conf.org/proceedings/lrec2012/pdf/844_Paper.pdf

[9] 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. [Online]. Available: http://dx.doi.org/10.1007/s10590-012-9125-1

[10] G. J. Edwards, C. J. Taylor, and T. F. Cootes, "Interpreting face images using active appearance models," in *Proc. International Conference on Automatic Face and Gesture Recognition*, June 1998, pp. 300–305.

[11] I. Matthews and S. Baker, "Active appearance models revisited," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 135–164, 2004.

[12] R. Gross, I. Matthews, and S. Baker, "Generic vs. person specific active appearance models," *Image and Vision Computing*, vol. 23, no. 12, pp. 1080–1093, 2005.

[13] J. Xiao, S. Baker, I. Matthews, and T. Kanade, "Real-time combined 2d+ 3d active appearance models," in *Proc. Computer Vision and Pattern Recognition*, vol. 2, 2004, pp. II–535.

[14] B. Aschenberner and C. Weiss, "Phoneme-viseme mapping for german video-realistic audio-visual-speech-synthesis."

[15] D. Rybach, C. Gollan, G. Heigold, B. Hoffmeister, J. Lööf, R. Schlüter, and H. Ney, "The rwth aachen university open source speech recognition system," in *Interspeech*, Brighton, UK, Sept. 2009, pp. 2111–2114.

[16] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, p. 381395, 1981.

[17] D. Vilar, D. Stein, M. Huck, and H. 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*, Uppsala, Sweden, July 2010, pp. 262–270.