Enhancing a Sign Language Translation System with Vision-Based Features

Philippe Dreuw, Daniel Stein, and Hermann Ney

Human Language Technology and Pattern Recognition, RWTH Aachen University
surname@cs.rwth-aachen.de

Abstract. In automatic sign language translation, one of the main problems is the usage of spatial information in sign language and its proper representation and translation, e.g. the handling of spatial reference points in the signing space. Such locations are encoded at static points in signing space as spatial references for motion events.

We present a new approach starting from a large vocabulary speech recognition system which is able to recognize sentences of continuous sign language speaker independently. The manual features obtained from the tracking are passed to the statistical machine translation system to improve its accuracy. On a publicly available benchmark database, we achieve a competitive recognition performance and can similarly improve the translation performance by integrating the tracking features.

1 Introduction

Most of the current sign language recognition systems use specialized hardware [14] and are very person dependent [11]. Furthermore, most approaches focus on the recognition of isolated signs only [12], or on the simpler case of isolated gesture recognition [13] which often can be characterized just by their movement direction. In [8] a review on recent research in sign language and gesture recognition is presented.

In statistical machine translation for sign language, only a few groups reported works on corpus-based approaches. In [7], an example-based approach is used for the translation. A statistical phrase-based translation model for the translation of weather forecasting news is presented in [10], which uses additional morpho-syntactic linguistic knowledge derived from a parser to improve the translation performance. In [2], a novel approach is proposed to translate Chinese to Taiwanese sign language and to synthesize sign videos based on joint optimization of two-pass word alignment and intersign epenthesis generation.

Although deaf, hard of hearing and hearing signers can fully communicate among themselves by sign language, there is a large communication barrier between signers and hearing people without signing skills.

We use a vision-based approach for automatic continuous sign language recognition [5], which does not require special data acquisition devices (e.g. data gloves or motion capturing systems), and a statistical machine translation framework

M. Sales Dias et al. (Eds.): GW 2007, LNAI 5085, pp. 108–113, 2009.

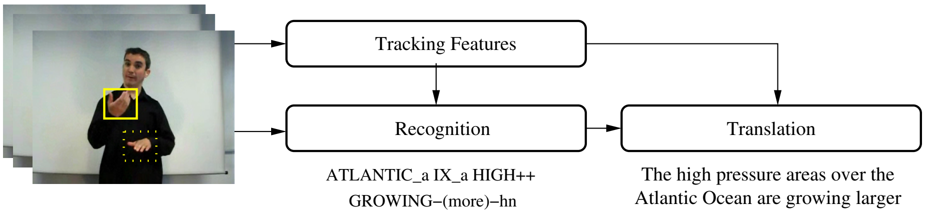

Fig. 1. Complete system setup with an example sentence: After automatically recognizing the input sign language video, the translation module has to convert the intermediate text format (glosses) into written text. We propose to use tracking based features also during translation.

for sign language translation [6]. Here, we propose to enhance a sign-to-text communication system with vision-based features. The system can aid the signing community with their everyday communication problems with the non-signing community. Fig. 1 illustrates the various components necessary for such a system.

2 Translation System: Overview

As mentioned above, recognition is only the first step of a sign-language to spoken-language system. The intermediate representation of the recognized signs is further processed to create a spoken language translation.

Statistical machine translation (SMT) is a data-based translation method that was initially inspired by the so-called noisy-channel approach: the source language is interpreted as an encryption of the target language, and thus the translation algorithm is typically called a decoder. In practice, statistical machine translation often outperforms rule-based translation significantly on international translation challenges, given a sufficient amount of training data.

A statistical machine translation system is used here to automatically transfer the meaning of a source language sentence into a target language sentence. Following the notation convention, we denote the source language with J words as $f_1^J = f_1 \ldots f_J$, a target language sentence, with I words, as $e_1^I = e_1 \ldots e_I$ and their correspondence as the a-posteriori probability $\Pr(e_1^I|f_1^J)$. The sentence $\hat{e}_1^I$ that maximizes this probability is chosen as the translation sentence as shown in Eq. 1. The machine translation system accounts for the different grammar and vocabulary of the sign language.

$$\hat{e}_1^I = \arg\max_{e_1^I} \left\{ \Pr(e_1^I) \cdot \Pr(f_1^J|e_1^J) \right\} \tag{1}$$

This classical source-channel model is generalized into a log-linear model, which allows the easy integration of additional models into the system. Each model's weighting factors are trained according to the maximum entropy principle. For a complete overview of the translation system, see [6].

To enhance translation quality, we propose to use visual features from the recognition process and include them into the translation as an additional

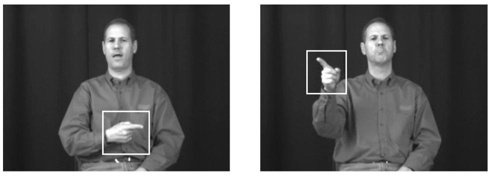

Fig. 2. Sample frames for pointing near and far used in the translation

knowledge source. The tracking positions of the 161 training sentences of the RWTH-Boston-104 database were clustered and their mean calculated. Then, for deictic signs, the nearest cluster according to the Euclidean distance was added as additional word information for the translation model (see Fig. 2).

3 Tracking System: Overview

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, the dominant hand is tracked in each image sequence. Therefore, 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. Our head and hand tracking framework is based on the algorithm described in [3,4]. This tracking algorithm is based on dynamic programming and is inspired by the time alignment algorithm in speech recognition and which guarantees to find the optimal path w.r.t. a given criterion and prevents taking possibly wrong local decisions.

These hand features, which are usually used within the recognition framework, can also be used within the translation framework, in order to decrease the translation error rate, too.

4 Experimental Results

All recognition experiments are measured similar to that done for speech recognition in terms of word error rates (WER) which is composed of errors that are due to deletion, insertion, or substitution of words.

To overcome the problem of incorrect WER due to the dependency on the perfect word order, we use the position independent word error rate (PER) as an additional measure in the translation experiments, which ignores the order of the words when comparing the words of the produced translation and the reference translation.

Sign Language Recognition: Some results concerning the recognition performance of our sign language recognition system on the RWTH-Boston-104 database and detailed information about the database itself are presented in [5]. A WER of 17.9% (i.e. 17 del., 3 ins., and 12 subst.) is achieved for a log-linear combination of two independently trained models accounting for long words and

Table 1. Recognition examples

Recognition reference	`JOHN IX GIVE MAN IX NEW COAT`
Recognition recognized as	`JOHN __ GIVE ___ IX NEW COAT`

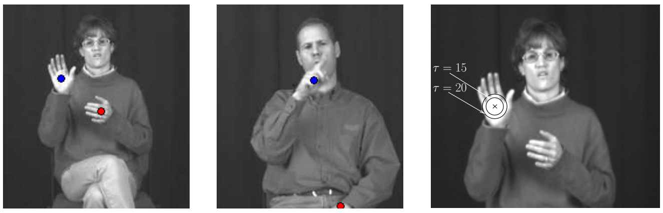

Fig. 3. Sample frames of the RWTH-Boston-Hands database with annotated hand positions. Left and right hand are marked with red and blue circles respectively. The last image shows different tolerance radii for $\tau = 15$ and $\tau = 20$ pixels.

short words respectively. The model weights have been optimized empirically. An example of a recognition result is shown in Tab. 1.

Tracking: A database for the evaluation of hand tracking methods in sign language recognition systems has been prepared. The RWTH-Boston-Hands database[1], which is freely available for further research, consists of a subset of the RWTH-Boston-104 videos. The positions of both hands have been annotated manually in 15 videos. A total of 1119 frames have been annotated.

For an image sequence $X_1^T = X_1, \ldots, X_T$ and corresponding annotated hand positions $u_1^T = u_1, \ldots, u_T$, the tracking error rate (TER) of tracked positions $\hat{u}_1^T$ is defined 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:

$$TER = \frac{1}{T} \sum_{t=1}^{T} \delta_\tau(u_t, \hat{u}_t) \qquad \text{with} \qquad \delta_\tau(u, v) := \begin{cases} 0 & \|u - v\| < \tau \\ 1 & \text{otherwise} \end{cases} \tag{2}$$

For $\tau = 20$, we achieve 2.30% TER for a 20×20 search window, where frames, in which the hands are not visible, are disregarded. Examples of annotated frames and the tolerance τ are shown in Fig. 3.

Sign-To-Text Translation: On the best recognition result of the 40 test sentences presented in [5], we achieve an overall system baseline performance of a signed-video-to-written-English translation of 27.6% WER, and 23.6% PER (computed on the written language) which is a very reasonable quality and, in spite of glosses, is intelligible for most people.

In another set of experiments, for incorporation of the tracking data, the tracking positions of the dominant-hand were clustered and their mean calculated. Then, for deictic signs, the nearest cluster according to the Euclidean distance

[1] `http://www-i6.informatik.rwth-aachen.de/~dreuw/database.php`

Table 2. Examples for different translation output while integrating hand tracking features from the sign language recognition system

Translation without tracking features	Translation with tracking features
John gives that man a coat .	John gives the man over there a coat .
John buy a car in the future .	John will buy a car soon .
Sue buys the blue car .	the blue car over there .

was added as additional word information for the translation model. The reference point was taken from the average positions in the training material, and so far only from the frontal perspective, thus without including 3-dimensional distance information. For example, the sentence `JOHN GIVE WOMAN IX COAT` might be translated into *John gives the woman the coat* or *John gives the woman over there the coat* depending on the nature of the pointing gesture `IX`. For temporal signs, we also measured the velocity, for example for the sign `FUTURE`, where it depends on the speed of the movement if it just marks future tense, refers to some event in the very near future ("soon") or in the general future. In Tab. 2, the first sentence corrects the meaning of the deictic sign from a distinctive article, "that", to a location reference "over there". For the second sentence, the temporal sign is refering to the very near future, "soon", instead of the more general "in the future", and is also corrected with the additional input. The last example corrects the deictic sign, but the overall translation deteriorates since the subject was not translated – it seems that, with different alignments in the training phase, some errors are possible. In general, however, the newly introduced words derived from the tracking feature, which affect roughly 18% of all sentences, helped the translation system to discriminate between deixis as distinctive article, locative or discourse entity reference function, and improved translation quality by 3% in WER and 2.2% PER.

5 Summary and Conclusions

We presented an approach to improve an automatic sign language translation system using visual features which is strongly inspired by state-of-the-art approaches in automatic sign language recognition. The used tracking system achieves a very good TER of 2.30% on the publicly available RWTH-Boston-Hands database.

For the translation step, preliminary experiments have shown that the incorporation of the tracking data for deixis words helps to properly interpret the meaning of the deictic gestures. By combining different data sources, the translation error rate decreases about 3% in WER and 2.2% PER.

The results suggest that hand tracking information is an important feature for sign language translation, especially for grammatically complex sentences where discourse entities and deixis occur a lot in signing space.

Other features that are likely to improve the error rates include tilt of the head, shifts of the upper body, or features describing the hand and body configuration as e.g. in [1,9]. These features should be analyzed and combined with the existing feature set in the recognition and translation framework.

Furthermore, a thorough analysis of the entities used in a discourse is required to properly handle pronouns.

References

1. Agarwal, A., Triggs, B.: Recovering 3D Human Pose from Monocular Images. IEEE Trans. PAMI 28(1), 44–58 (2006)
2. Chiu, Y.-H., Wu, C.-H., Su, H.-Y., Cheng, C.-J.: Joint Optimization of Word Alignment and Epenthesis Generation for Chinese to Taiwanese Sign Synthesis. IEEE Trans. PAMI 29(1), 28–39 (2007)
3. Dreuw, P., Deselaers, T., Rybach, D., Keysers, D., Ney, H.: Tracking Using Dynamic Programming for Appearance-Based Sign Language Recognition. In: 7th Intl. Conference on Automatic Face and Gesture Recognition, pp. 293–298. IEEE, Southampton (2006)
4. Dreuw, P., Forster, J., Deselaers, T., Ney, H.: Efficient Approximations to Model-based Joint Tracking and Recognition of Continuous Sign Language. In: IEEE Face and Gesture Recognition, Amsterdam, The Netherlands (September 2008)
5. Dreuw, P., Rybach, D., Deselaers, T., Zahedi, M., Ney, H.: Speech Recognition Techniques for a Sign Language Recognition System. In: Interspeech 2007 - Eurospeech, Antwerp, Belgium, August 2007, pp. 2513–2516 (2007)
6. Mauser, A., Zens, R., Matusov, E., Hasan, S., Ney, H.: The RWTH Statistical Machine Translation System for the IWSLT 2006 Evaluation. In: IWSLT, Kyoto, Japan, November 2006, pp. 103–110 (2006) (Best paper award)
7. Morrissey, S., Way, A.: An Example-Based Approach to Translating Sign Language. In: Workshop Example-Based Machine Translation (MT X 2005), Phuket, Thailand, September 2005, pp. 109–116 (2005)
8. Ong, S.C., Ranganath, S.: Automatic Sign Language Analysis: A Survey and the Future beyond Lexical Meaning. IEEE Trans. PAMI 27(6), 873–891 (2005)
9. Ramanan, D., Forsyth, D.A., Zisserman, A.: Tracking People by Learning Their Appearance. IEEE Trans. PAMI 29(1), 65–81 (2007)
10. Stein, D., Bungeroth, J., Ney, H.: Morpho-Syntax Based Statistical Methods for Sign Language Translation. In: 11th Annual conference of the European Association for Machine Translation, Oslo, Norway, June 2006, pp. 169–177 (2006)
11. Vogler, C., Metaxas, D.: A Framework for Recognizing the Simultaneous Aspects of ASL. CVIU 81(3), 358–384 (2001)
12. von Agris, U., Schneider, D., Zieren, J., Kraiss, K.-F.: Rapid Signer Adaptation for Isolated Sign Language Recognition. In: CVPR Workshop V4HCI, New York, USA, June 2006, p. 159 (2006)
13. Wang, S.B., Quattoni, A., Morency, L.-P., Demirdjian, D., Darrell, T.: Hidden Conditional Random Fields for Gesture Recognition. In: CVPR, June 2006, vol. 2, pp. 1521–1527 (2006)
14. Yao, G., Yao, H., Liu, X., Jiang, F.: Real Time Large Vocabulary Continuous Sign Language Recognition Based on OP/Viterbi Algorithm. In: 18th ICPR, August 2006, vol. 3, pp. 312–315 (2006)