Large-scale Learning of Sign Language by Watching TV (Using Co-occurrences)

Tomas Pfister[1]
tp@robots.ox.ac.uk
James Charles[2]
j.charles@leeds.ac.uk
Andrew Zisserman[1]
az@robots.ox.ac.uk

[1] Department of Engineering Science
University of Oxford
Oxford, UK

[2] School of Computing
University of Leeds
Leeds, UK

We present a framework that automatically and quickly learns a large number of signs from sign language-interpreted TV broadcasts by exploiting supervisory information available in the subtitles.

Our contributions are: (i) we show that, somewhat counter-intuitively, mouth patterns are highly informative for distinguishing words in a language for the Deaf, and their co-occurrence with signing can be used to significantly reduce the correspondence search space; and (ii) we develop a multiple instance learning method using an efficient discriminative search, which determines a candidate list for the sign with both high recall and precision.

The previous approach of Buehler *et al*. [2] for learning signs relies on complex features and a computationally expensive, application-specific learning framework. This has hindered the large scale application of this method. In this paper we describe a method that is much simpler and computationally lighter.

Motivation. TV programmes in many countries across the world are now routinely broadcast with both subtitles and an overlaid signer translating to the Deaf audience (Fig. 2). Our aim is to use this material to learn *signs* corresponding to English *words* in the subtitles [2, 3]. We use this continuous and rich source of training material to build a database of word-sign pairs for a large number of signs and signers. The vision is that this database can later be used to train a large-scale person-independent sign language to text translator.

Summary of method. We cast the problem as one of Multiple Instance Learning (MIL), where the training data are visual descriptors (hand trajectories) with weak supervision from subtitles. We proceed in three steps: (i) the search space for correspondences is significantly reduced by exploiting lip and hand motion co-occurrences to filter away irrelevant intervals of the temporal sequences; (ii) candidates for the signs are obtained using an efficient discriminative search over all remaining sequences by casting each candidate as a classifier for the positive and negative sequences; and finally (iii) these candidates are then selected or rejected using the MIL support vector machine framework (MI-SVM) [1]. Fig. 1 illustrates the processing steps and Fig. 3 shows example results.

We demonstrate the method on videos from BBC TV broadcasts, and achieve higher accuracy and recall than previous methods, despite using much simpler features.

[1] S. Andrews, I. Tsochantaridis, and T. Hofmann. Support vector machines for multiple-instance learning. In *NIPS*, 2002.

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

[3] H. Cooper and R. Bowden. Learning signs from subtitles: A weakly supervised approach to sign language recognition. In *Proc. CVPR*, 2009.

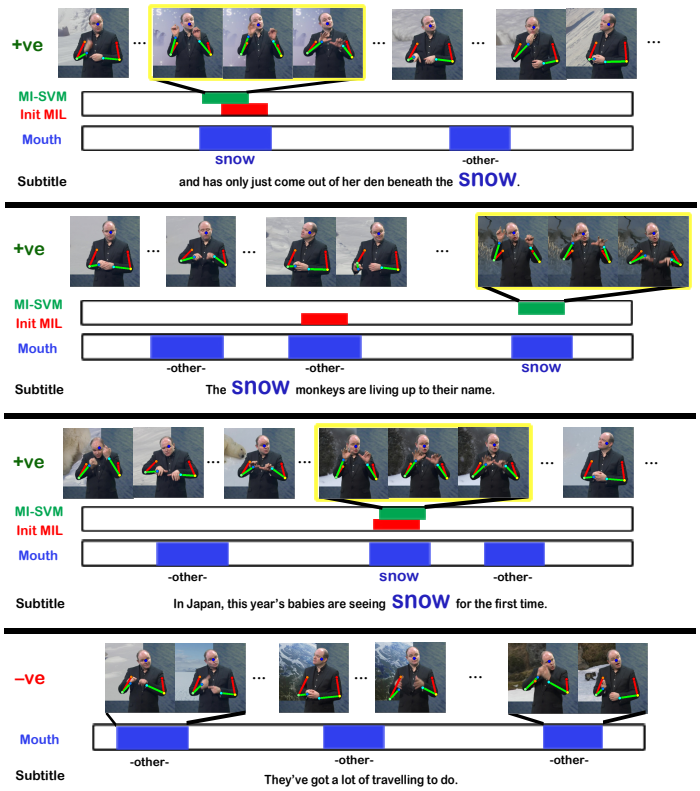

Figure 1: **Learning signs from co-occurrences of subtitle text, mouth and hand motion.** The top three rows are positive subtitle sequences which contain the text word and sign for 'snow'. The final row is an example of a negative subtitle sequence which does not contain 'snow'. Signs are learnt from this weakly aligned and noisy data. A fixed size temporal window is slid across the frames in which mouth motion occurs (blue). The rest of the sequence can be ignored, thus reducing the temporal search space. Candidate signs are proposed by a discriminative MIL search using temporal correlation. A subset of these candidates (red) are used to initialise a MI-SVM, resulting in the final correspondence matches (green). The red and green lines on the signer show the detected limbs and head.

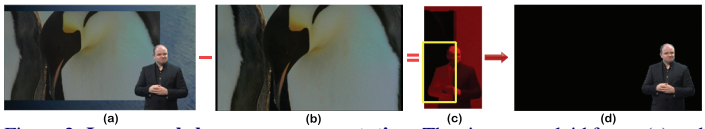

Figure 2: **Large-scale human co-segmentation.** The signer-overlaid frame (a) and original frame (b) are subtracted after alignment, resulting in a difference image (c) that is used as a constraint (yellow) in the segmentation. (d) shows the final result.

Figure 3: **Example output videos for the signs "snow" (top) and "vision" (bottom)** performed by two different signers and learnt automatically.