Content-Based Video Analysis and Access for Finnish Sign Language – A Multidisciplinary Research Project

Markus Koskela[α], Jorma Laaksonen[β], Tommi Jantunen[γ], Ritva Takkinen[δ], Päivi Rainò[ε], Antti Raike[ζ]

[α, β] Helsinki University of Technology, Dept. of Information and Comp. Science, P.O. Box 5400, FI-02015 TKK
[γ, δ] University of Jyväskylä, Department of Languages, P.O. Box 35 (F), FI-40014 University of Jyväskylä
[ε] Finnish Association of the Deaf, Sign Language Unit, P.O. Box 57, FI-04001 Helsinki
[ζ] University of Art and Design, Media Lab, Hämeentie 135C, FI-00560 Helsinki
E-mail: markus.koskela@tkk.fi, jorma.laaksonen@tkk.fi, tommi.jantunen@campus.jyu.fi, ritva.takkinen@campus.jyu.fi, paivi.raino@kl-deaf.fi, antti.raike@taik.fi

Abstract

This paper outlines a multidisciplinary research project in which computer vision techniques for the recognition and analysis of gestures and facial expressions from video are developed and applied to the processing of sign language in general and Finnish Sign Language in particular. This is a collaborative project between four project partners: Helsinki University of Technology, University of Jyväskylä, University of Art and Design, and the Finnish Association of the Deaf. The project has several objectives of which the following four are in the focus of this paper: (i) to adapt the existing PicSOM framework developed by the Helsinki University of Technology regarding content-based analysis of multimedia data to content-based analysis of sign language videos containing continuous signing; (ii) to develop a computer system which can identify sign and gesture boundaries and indicate, from the video, the sequences that correspond to signs and gestures; (iii) to apply the studied and developed methods and computer system for automatic and semi-automatic indexing of sign language corpora; and (iv) to conduct a feasibility study for the implementation of mobile video access to sign language dictionaries and corpora. Methods for reaching the objectives are presented in the paper.

1. Introduction

This paper presents four key objectives of a research project that aims to develop computer vision techniques for the recognition and analysis of gestures and facial expressions from video in order to apply them to the processing of sign language, and especially Finnish Sign Language (FinSL). The project is a collaborative effort of four project partners, all representing the leading Finnish research units in their own fields: Helsinki University of Technology, University of Jyväskylä, University of Art and Design, and the Finnish Association of the Deaf. The composition of the consortium reflects the fact that the visual analysis and computerized study of sign language is a multidisciplinary challenge that calls for expertise in a large variety of scientific fields.

2. Objectives of the Project

2.1 Methods for Content-Based Processing and Analysis of Signed Videos

The first objective of the project is to develop novel methods for a content-based processing and analysis of sign language videos, recorded using a single camera. The PicSOM[1] retrieval system framework (Laaksonen et al., 2002), developed by the Helsinki University of Technology regarding content-based analysis of multimedia data, will be adapted to continuous signing, to facilitate the automatic and semi-automatic analysis of sign language videos. The framework has been previously applied to content-based retrieval and analysis in various application domains, including large photograph collections, broadcast news videos, multispectral and polarimetric radar satellite images, industrial computer vision, and face recognition. Figure 1 shows an example of the PicSOM user interface during interactive retrieval from a database of recorded broadcast news programs (Koskela et al., 2005).

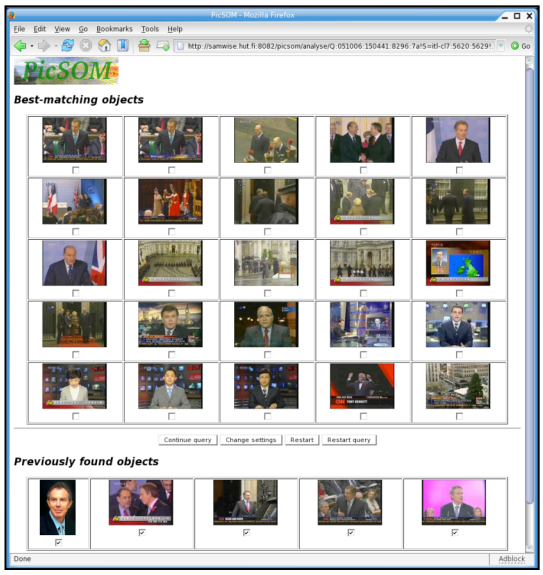

Figure 1: The user interface of PicSOM during an interactive retrieval task "Find shots of Tony Blair" from a database of recorded broadcast news.

[1] http://www.cis.hut.fi/picsom/

The PicSOM system is based on indexing any type of multimedia using parallel Self-Organizing Maps (SOMs) (Kohonen, 2001) as the standard indexing method. The Self-Organizing Map is a powerful tool for exploring huge amounts of high-dimensional data. It defines an elastic, topology-preserving grid of points that is fitted to the input space. It is often used for clustering or visualization, usually on a two-dimensional regular grid. The distribution of the data vectors over the map forms a two-dimensional discrete probability density. Even from the same data, qualitatively different distributions can be obtained by using different feature extraction methods.

During the training phase in PicSOM, the SOMs are trained with separate data sets, obtained from the multimodal object data with different automatic feature extraction techniques. The different SOMs and their underlying feature extraction schemes then impose different similarity functions on the images, videos, texts and other media objects. In the PicSOM approach, the system is able to discover the parallel SOMs that provide the most valuable information, e.g., for retrieving relevant objects in each particular query. Recently, the system has also been applied to other ways of analyzing video material, i.e. shot boundary detection and video summarization (Laaksonen et al., 2007).

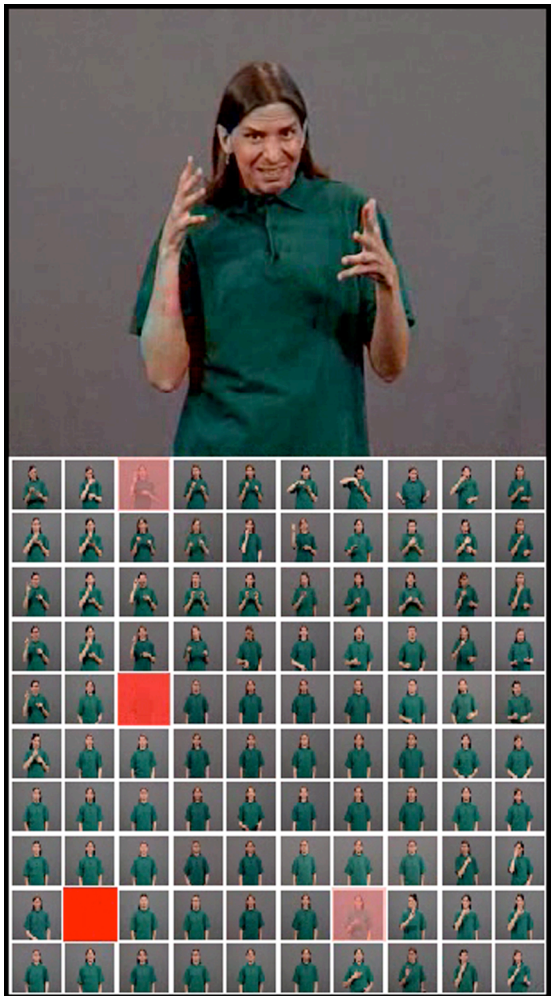

Figure 2: A PicSOM analysis of a signed sequence KNOW MATTER CLEAR 'Well of course, it is obvious!' (*Suvi*'s article 3, example video 6) using the standard MPEG-7 *Edge Histogram* feature.

The existing general-purpose video feature extraction methods will provide a starting point for the analysis of recorded sign-language videos in this project. At a later stage, more specific features for the domain of sign-language videos will be developed. Figure 2 shows an example of an analysis of a signed sequence with a SOM trained using a standard MPEG-7 *Edge Histogram* image feature. The sequence is from *Suvi*, the online dictionary of FinSL.[2]

2.2 Automatic Segmentation of Continuous Sign Language Videos

The second objective of the project is to develop a computer system which can both (i) automatically indicate meaningful signs and other gesture-like sequences from a video signal which contains natural sign language data, and (ii) disregard parts of the signal that do not count as such sequences. In other words, the goal is to develop an automatized mechanism that can identify sign and gesture boundaries and indicate, from the video, the sequences that correspond to signs and gestures.

An automatic segmentation of recorded continuous sign language data is an important first step in the automatic processing of sign language videos and online applications. Traditionally, the segmentation of sign language data has been done manually by using specific video annotation programs such as ELAN [3] (e.g. Crasborn et al., 2007) or SignStream[4] (Neidle, 2001). However, identifying signs and gestures from the video this way is extremely time consuming, a preliminary segmentation of one hour of data requiring two weeks of active working time from one person at the minimum. Automating or even semi-automating this preliminary and mechanical step in the data-handling phase would facilitate the workflow considerably.

So far there have been no real attempts to identify *only* sign and gesture-like *forms* from the stream of natural signed language video. Projects dealing with sign recognition (e.g. Ong & Rangarath, 2005) have all included the semantic recognition of signs' content as one of their goals. Also, most of the research done until now has dealt only with the recognition of isolated signs from data produced specially for research purposes. In this project the semantics of signs are not directly dealt with; the objective being data and signer independent identification of signs/gestures and their boundaries.

[2] http://suvi.viittomat.net/
[3] http://www.lat-mpi.eu/tools/elan/
[4] http://www.bu.edu/asllrp/SignStream/

Linguistically, the automatic identification of signs and gestures and their boundaries will be grounded as far as possible on prosodic information. For example, linguistic boundaries in sign languages are typically indicated by changes in the facial prosody, i.e. by the changes in the posture and movement of the mouth, eyes, eyebrows, and head (e.g. Wilbur, 2000). For the automatic detection of these changes, we shall apply our existing face detection algorithm (cf. Figure 3), which is capable of detecting the eyes, nose, and mouth separately (Yang & Laaksonen, 2005).

Figure 3: An example of face detection from a recorded sign language video. The detected eyes, nose, and mouth are also shown with separate bounding boxes.

In addition to still image features extracted from single video frames, an essential feature in the analysis of recorded continuous-signing sign language is that of motion. For tracking local motion in the video stream, we apply a standard algorithm based on detecting distinctive pixel neighborhoods and then minimizing the sum of squared intensity differences in small image windows between two successive video frames (Tomasi & Kanade, 1991). An example of detected local motion is illustrated in Figure 4. The tracked points that remain stationary are not shown.

We assume that the parts of the signal where there is significantly less or no local motion correspond to the significant junctures such as the beginning and ending points of lexematic signs. However, the exact relation between motion and sign boundaries is an open research question that is essential to this objective and will be studied extensively within the research project. It can be assumed that a combination of a hand detector, still image feature extraction, and motion analysis are needed for a successful detection of sign and gesture boundaries. The PicSOM system inherently supports such fusion of different features extracted from different modalities.

During the project, the analysis of motion tracked interest points will be further developed to test the general assumption in the current signed syllable

Figure 4: An example of tracked point features marking the local movement in the sign JOYSTICK excerpted from the phrase 'The boy is really interested in playing computer games' (*Suvi*'s article 1038, example video 3).

research, according to which sign internal phonological movements function as syllables' sonority peaks, that is, as the most salient parts of the signed signal (e.g. Jantunen, 2007; Jantunen & Takkinen, in press). We hypothesize that if the sonority assumption is correct, the motion tracked interest points should cumulate relatively more to the parts of the signal within the signs, not to the parts outside them.

2.3 Testing Methods for Indexing Existing Sign Language Material

The third objective is linked to generating an example-based corpus for FinSL. There exist increasing amounts of recorded video data of the language, but almost no means for utilizing it efficiently due to missing indexing and lack of methods for content-based access. The studied methods could facilitate a leap forward in founding the corpus. The tool for automatic processing, created and tested in this project, will be further applied to segmenting and indexing the pre-existing FinSL data in order to prepare an open-access visual corpus for linguistic research. Lacking content-based indexing and retrieval tools, the digitized data found in video magazines and online publications in FinSL covering the last 25 years has up to now been scarcely utilized within the FinSL research. It should be emphasized, however, that the functionality provided by the PicSOM system can be already used as such to analyze and index the visual content of signed video material and to construct a nonverbal index of recurrent images and video clips.

2.4 Implementation of Mobile Video Access to Sign Language Dictionaries and Corpora

The fourth objective is a feasibility study for the implementation of mobile video access to sign language dictionaries and corpora. Currently an existing dictionary can be searched by giving a rough description of the location, motion and hand form of the sign. The

automatic content-based analysis methods could be applied to online mobile phone videos, thus enabling sign language access to dictionaries and corpora.

In this application, it will be more essential than in the previous ones that the speed and robustness of the implementation can be optimized. We do not expect that the quality of mobile sign language videos could be good enough for accurate classification. However, we believe that by combining the automatic video analysis methods with novel interaction and interface techniques, we can take a substantial step towards a mobile sign language dictionary.

3. Conclusion

In this paper we have outlined the key objectives of the research project that aims to develop computer vision techniques for recognition, indexing, and analysis of sign language data. We believe that the PicSOM system developed by the Helsinki University of Technology provides an excellent basis for this task. As the project proceeds, we will explore more methods and apply the PicSOM system to the massive video data that will be the foundation of the new FinSL corpus.

References

Crasborn, O., Mesch, J., Waters, D., Nonhebel, A., van der Kooij, E., Woll, B., Bergman, B. (2007). Sharing Sign Language Data Online. Experiences from the ECHO Project. *International Journal of Corpus Linguistics* 12(4), pp. 537–564.

Jantunen, T. (2007). Tavu suomalaisessa viittomakielessä. [The Syllable in Finnish Sign Language]. *Puhe ja kieli* 27(3), pp. 109–126.

Jantunen, T., Takkinen, R. (in press). Syllable Structure in SL Phonology. In D. Brentari (Ed.), *Sign Languages*. Cambridge, UK: Cambridge University Press.

Kohonen, T. (2001). Self-Organizing Maps. Third edn. Springer-Verlag.

Koskela, M., Laaksonen, J., Sjöberg M., Muurinen, H. (2005). PicSOM Experiments in TRECVID 2005. Online Proceedings of the TRECVID 2005 Workshop. Gaithersburg, MD, USA. November 2005.

Laaksonen, J., Koskela, M., Oja, E. (2002). PicSOM – Self-Organizing Image Retrieval with MPEG-7 Content Descriptions. *IEEE Transactions on Neural Networks*, 13(4), pp. 841–853.

Laaksonen, J., Koskela, M., Sjöberg. M., Viitaniemi, V., Muurinen, H. (2007) Video Summarization with SOMs. Proceedings of 6th International Workshop on Self-Organizing Maps (WSOM 2007). Bielefeld, Germany. September 2007.

Neidle, C. (2001). SignStream™. A Database Tool for Research on Visual-Gestural Language. *Sign Language & Linguistics* 4(1/2), pp. 203–214.

Ong, S. C. W, Ranganath S. (2005). Automatic Sign Language Analysis: A Survey and the Future beyond Lexical Meaning. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 27(6), pp. 873–891.

Suvi = Suvi – Suomalaisen viittomakielen verkkosanakirja [Online Dictionary of Finnish Sign Language]. [Helsinki]: Kuurojen Liitto ry [The Finnish Association of the Deaf], 2003. Online publication: http://suvi.viittomat.net.

Tomasi, C., Kanade, T. (1991). Detection and Tracking of Point Features. Carnegio-Mellon University Technical Report CMU-CS-91-132. April 1991.

Wilbur, R. B. (2000). Phonological and Prosodic Layering of Nonmanuals in American Sign Language. In K. Emmorey, H. Lane (eds.), *The Signs of Language Revisited. An Anthology to Honor Ursula Bellugi and Edward Klima*, pp. 215–244. Mahwah, NJ, London: Lawrence Erlbaum Associates.

Yang, R., Laaksonen, J. (2005). Partial Relevance in Interactive Facial Image Retrieval. Proceedings of 3rd International Conference in Pattern Recognition (ICAPR 2005). Bath, UK. August 2005.