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

Markus Koskela(α), Jorma Laaksonen(α), Tommi Jantunen(β), Ritva Takkinen(β), Päivi Rainò(γ), Antti Raike(δ)

(α) Helsinki University of Technology, Dept. of Information and Comp. Science, P.O. Box 5400, FI-02015 TKK

(β) University of Jyväskylä, Department of Languages, P.O. Box 35 (F), FI-40014 University of Jyväskylä

(γ) Finnish Association of the Deaf, Sign Language Unit, P.O. Box 57, FI-04001 Helsinki

(δ) 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 poster presents the technology and outlines four key objectives of 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. The project is a collaborative effort between four project partners: Helsinki University of Technology, University of Jyväskylä, University of Art and Design, and the Finnish Association of the Deaf.

The PicSOM System

A key research method in the proposed research project is the existing general framework of content-based analysis of multimedia, PicSOM, developed at TKK (Laaksonen et al., 2002; see http://www.cis.hut.fi/picsom/). The PicSOM framework already supports a large variety of sub-methods necessary for analyzing video streams of sign language data. 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 (see Figure 1; Koskela et al., 2005).

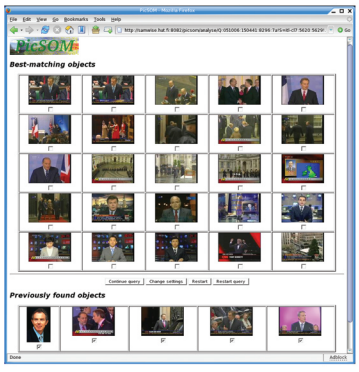

Figure 1: The user interface of PicSOM during an interactive retrieval task "Find shots of Tony Blair" from a database of recorded broadcast news.

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.

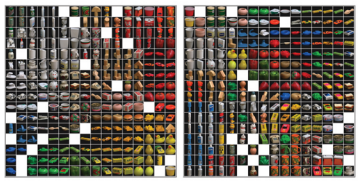

Figure 2: An object database organized with SOMs trained with color (left) and edge (right) features.

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).

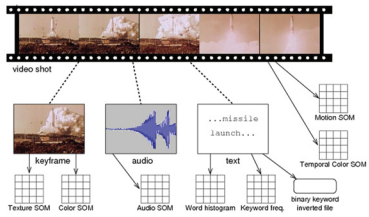

Figure 3: The general architecture for indexing video with multimodal SOMs in PicSOM.

Objective 1: Develop 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. The PicSOM retrieval system framework will be adapted to continuous signing to facilitate the automatic and semi-automatic analysis of sign language videos.

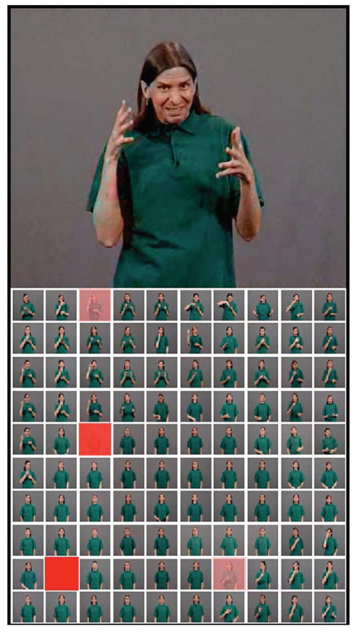

Figure 4: A PicSOM analysis of a signed sequence KNOW MATTER CLEAR 'Well of course, it is obvious!' (Suvi's article 3, example video 6) using the standard MPEG-7 Edge Histogram feature.

The existing general-purpose video feature extraction methods will provide a starting point for the analysis of recorded sign-language videos in this project (see Figure 4). At a later stage, more specific features for the domain of sign-language videos will be developed.

Objective 2: Automatic Segmentation of Continuous Sign Language Videos

The second objective of the project is to develop a computer system that can identify sign and gesture boundaries and indicate, from the video, the sequences that correspond to signs and gestures; the semantics of signs are not directly dealt with.

Linguistic boundaries in sign languages are typically indicated by changes in the posture and movement of the mouth, eyes, eyebrows, and head. For the automatic detection of these changes, we shall apply our existing face detection algorithm (cf. Figure 5), which is capable of detecting the eyes, nose, and mouth separately (Yang & Laaksonen, 2005).

An essential feature in the analysis of recorded sign language data is that of motion. For tracking local motion in the video stream, we apply a standard algorithm based on

Figure 5: An example of face detection from a recorded sign language video. The detected eyes, nose, and mouth are also shown with separate bounding boxes.

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; see Figure 6).

Figure 6: 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).

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.

Objective 3: Testing Methods for Indexing Existing Sign Language Material

The third objective is linked to generating an example-based corpus for FinSL. The functionality provided by the PicSOM system can be already used as such to segment and index the pre-existing FinSL data in order to prepare an open-access visual corpus for linguistic research. Also, the tool for automatic processing, created and tested in this project, will be further applied for this purpose.

Objective 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. 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.

References

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.

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.

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.

TKK JYVÄSKYLÄN YLIOPISTO UNIVERSITY OF ART AND DESIGN HELSINKI Kuurojen Liitto ry