Content-Based Video Analysis and Access for Finnish Sign Language

A Multidisciplinary Research Project

Markus Koskelo(a), Jorma Laaksonen(c), Tommi Juntunen(b), Ritva Takkinen(b), Päivi Raino(b), Antti Raikko(b)

(a) Helsinki University of Technology, Dept. of Information Engineering, P.O. Box 8000, FI-02015 TKK
(b) University of Jyväskylä, Department of Languages, P.O. Box 35 (FI), FI-40014 University of Jyväskylä
(c) Finnish Association of the Deaf, Sign Language Unit, P.O. Box 57, FI-00440 Helsinki

E-mail: markus.koskelo@tkk.fi, jorma.laaksonen@tkk.fi, tommi.juntunen@campus.jyu.fi, ritva.takkinen@campus.jyu.fi, pavi.raino@tkk.fi, antti.raikko@tkk.fi

Abstract
This paper 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 the 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 PiSOM System
A key research and development project in the proposal research project is the existing generalized framework of content-based analysis of multimedia, PiSOM, developed in TKK (Laaksonen et al., 2002; see http://www.eecs.ohio-state.edu). The PiSOM framework already supports a large variety of sub-methods necessary for analyzing video streams of sign language. The framework has been previously applied to content-based retrieval and analysis in various application domains, including large photographic collections, broadcast news videos, multimedia and polarimetric radar surveillance images, industrial computer vision, and face recognition (see Figure 1; Koskelo et al., 2003).

Figure 1: The user interface of PiSOM during an interactive retrieval task "Find shots of Topsy Blows" from a database of recorded broadcast news.

The PiSOM system is based on indexing any type of multimedia using parallel Self-Organizing Maps (SOMs) (Kohonen, 2001) in the standard indexing method. The Self-Organizing Map is a powerful tool for exploiting 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 input 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 PiSOM, the SOMs are trained with separate data sets, obtained from the multimodal data object with different automatic feature extraction techniques. The different SOMs and their underlying feature extraction schemes then differ in similarity functions for the images, videos, texts and other media objects. In the PiSOM approach, the system is able to discover the parallel SOMs that provide the most valuable information, e.g., for retrieving relevant objects in such particular queries. 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).

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 PiSOM retrieval system framework will be adapted to continuous signing to facilitate the automatic and semi-automatic analysis of sign language videos.

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

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.

Figure 4: A PiSOM analysis of a signed sequence KNOW MATTER CLEAR. Well of course, it is obvious! (Test's article 1018, 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 domains of sign-language videos will be developed.

Objective 3: Exploration of Mobile Video Access to Sign Language Dictionary and Corpora
The third objective is linked to generating an example-based corpus for PiSOM. The functionality provided by the PiSOM system can be already used as such to segment and index the pre-existing Field data in order to prepare an open-access visual corpus for linguistic research. The tool for automatic processing, creation and testing 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.

Figure 5: An example of tracked point features marking the local movement in the sign KYSYKYK excepted from the phrase "The boy is really intently playing computer games" (Test's article 1018, example video 3).

We assume that the parts of the signal where there is significantly less or no local motion correspond to the significant features such as the beginning and ending points of semantic phrases. 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.

Figure 6: An example of tracked point features marking the local movement in the sign KYSYKYK excepted from the phrase "The boy is really intently playing computer games" (Test's article 1018, example video 3).

We assume that the parts of the signal where there is significantly less or no local motion correspond to the significant features such as the beginning and ending points of semantic phrases. 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.

Figure 7: An example of tracked point features marking the local movement in the sign KYSYKYK excepted from the phrase "The boy is really intently playing computer games" (Test's article 1018, example video 3).

We assume that the parts of the signal where there is significantly less or no local motion correspond to the significant features such as the beginning and ending points of semantic phrases. 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.

Figure 8: An example of tracked point features marking the local movement in the sign KYSYKYK excepted from the phrase "The boy is really intently playing computer games" (Test's article 1018, example video 3).