# Chapter 7

# Content-based information retrieval and analysis

**Erkki Oja, Jorma Laaksonen, Markus Koskela, Ville Viitaniemi, Mats Sjöberg, Xi Chen, Satoru Ishikawa, Matti Karppa, Mikko Kurimo, Ville Turunen**

## 7.1 Introduction

The Content-Based Information Retrieval Research Group studies and develops efficient methods for content-based and multimodal information retrieval and analysis tasks and implements them in the PicSOM[1] content-based information retrieval (CBIR) system. In the PicSOM CBIR system, parallel Self-Organizing Maps (SOMs) and Support Vector Machine (SVM) classifiers have been trained with separate data sets obtained from the multimodal object data with different feature extraction techniques. The different classifiers and their underlying feature extraction schemes impose different similarity measures and categorizations on the images, videos, texts and other media objects.

## 7.2 Semantic concept detection from images and videos

Extracting semantic concepts from multimedia data has been studied intensively in recent years. The aim of the research on the multimedia retrieval research community has been to facilitate semantic indexing and concept-based retrieval of unannotated multimedia content. The modeling of mid-level semantic concepts is often essential in supporting high-level indexing and querying on multimedia data as such concept models can be trained off-line with considerably more positive and negative examples than what are available at interactive query time.

In the course of previous years we have outlined and implemented our generic PicSOM system architecture for multimedia retrieval tasks. Detection of concepts from multimedia data—e.g. images and video shots—forms an important part of the architecture and we have formulated it as a standard supervised machine learning problem. Our concept detection technology is fundamentally based on fusion of a large number of elementary detections, each based on a different low-level audiovisual feature extracted from the multimedia data [1, 2].

During the period 2010–2011 we have continued our work in improving the bag of visual words (BoV) techniques for concept detection [3] and our participation in the annual TRECVID video analysis evaluations[2]. We have also applied our general-purpose algorithm for visual category recognition to the recognition of indoor locations [4]. Indoor localization is an important application in many emerging fields, such as mobile augmented reality and autonomous robots. A number of different approaches have been proposed, but arguably the prevailing method is to combine camera-based visual information to some additional input modalities, such as laser range sensors, depth cameras, sonar, stereo vision, temporal continuity, odometry, and the floorplan of the environment. We evaluated our method with other location recognition systems in the ImageCLEF 2010 RobotVision contest.

As a joint work together with the Speech Recognition Research Group, we have participated in the *Next Media* TIVIT ICT SHOK since 2010. We have applied our content-based video analysis and continuous speech recognition systems for the analysis of television broadcast material provided by the Finnish Broadcasting Company YLE. Figure 7.1 illustrates the results of the analysis for one regional news broadcast. On the left we can see how the temporal structure of the program has been revealed based on the clustering of

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

[2] `http://trecvid.nist.gov/`

visible human faces. In the right subfigure, the detected visual concepts are show on the top, the continuous speech recognition output on the bottom and the recognized name of the person on the right.

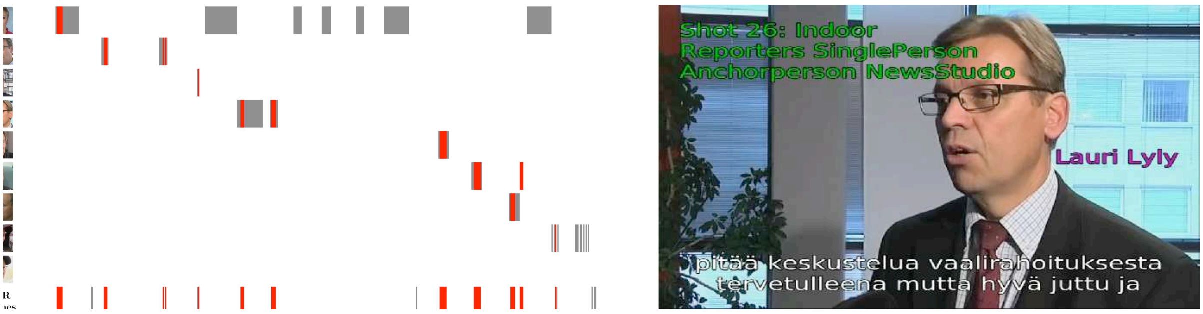

Figure 7.1: Temporal analysis of a news program based on clustering of facial images and the resulting content annotations.

## 7.3 Content-based video analysis and annotation of Finnish Sign Language

In January 2011 a new project *Content-based video analysis and annotation of Finnish Sign Language (CoBaSiL)*, funded by the Academy of Finland for four years, was started. In our joint work with the University of Jyväskylä and the Finnish Association of the Deaf, we are applying our methods of video content processing for the analysis, indexing and recognition of recorded Finnish Sign Language. In the project we study the use of computer vision techniques to recognize and analyze first the body parts of the signer and then his or her hand locations, shapes and gestures and facial expressions. Figure 7.2 illustrates the results of the stages of face detection, skin-color recognition and shape modelling with active shape models in the processing chain [5].

The linguistic goal of the project is to identify the sign and gesture boundaries and to indicate which video sequences correspond to specific signs and gestures [6]. This will facilitate indexing and construction of an example-based open-access visual corpus of the Finnish Sign Language for which there already exists large amounts of non-indexed video material. Currently we have concentrated our effort on studying the partially annotated material of the publicly available on-line dictionary of Finnish Sign Language, Suvi[3].

## 7.4 Image based linking

Augmenting the user's perception of her surroundings using a mobile device is a relatively new field of research which has been invigorated by the growth in number of capable mobile computing devices. These devices, while becoming increasingly small and inexpensive, allow us to use various computing facilities while roaming in the real world. In particular, ordinary mobile phones with integrated digital cameras are nowadays common, and even they can provide new ways to get access to digital information and services. Images or

[3] `http://suvi.viittomat.net/`

video captured by the mobile phone can be analyzed to recognize the object or scene appearing in the recording.

We studied new ways to get access to digital services for mobile phones in the Image Based Linking project in 2009–2011 [7, 8]. These kinds of methods can be used for various purposes linking digital information to the physical world. Possible application areas include outdoor advertising, additional digital material to magazine and newspaper articles, tourist applications, and shopping. Our focus in the project was on a use case with a magazine publisher as the content provider. Several target images can exist on the same page of a magazine, each linked to different extra information. Consequently, the target images may be rather small in print, and the captured photos may be highly blurred and out-of-focus. An example of matching such photos to the images in the magazine database is shown in Figure 7.3.

## References

[1] Ville Viitaniemi, Mats Sjöberg, Markus Koskela, and Jorma Laaksonen. Concept-based video search with the PicSOM multimedia retrieval system. Technical Report TKK-ICS-R39, Aalto University School of Science and Technology, Department of Information and Computer Science, Espoo, Finland, December 2010.

[2] Ville Viitaniemi, Mats Sjöberg, Markus Koskela, and Jorma Laaksonen. Automatic video search using semantic concepts. In *Proceedings of 8th European Conference on Interactive TV and Video (EuroITV 2010)*, Tampere, Finland, June 2010.

[3] Ville Viitaniemi and Jorma Laaksonen. Region matching techniques for spatial bag of visual words based image category recognition. In *Proceedings of 20th International Conference on Artificial Neural Networks (ICANN 2010)*, volume 6352 of *Lecture Notes in Computer Science*, pages 531–540, Thessaloniki, Greece, September 2010. Springer Verlag.

[4] Mats Sjöberg, Markus Koskela, Ville Viitaniemi, and Jorma Laaksonen. Indoor location recognition using fusion of SVM-based visual classifiers. In *Proceedings of 2010 IEEE International Workshop on Machine Learning for Signal Processing*, pages 343–348, Kittilä, Finland, August-September 2010.

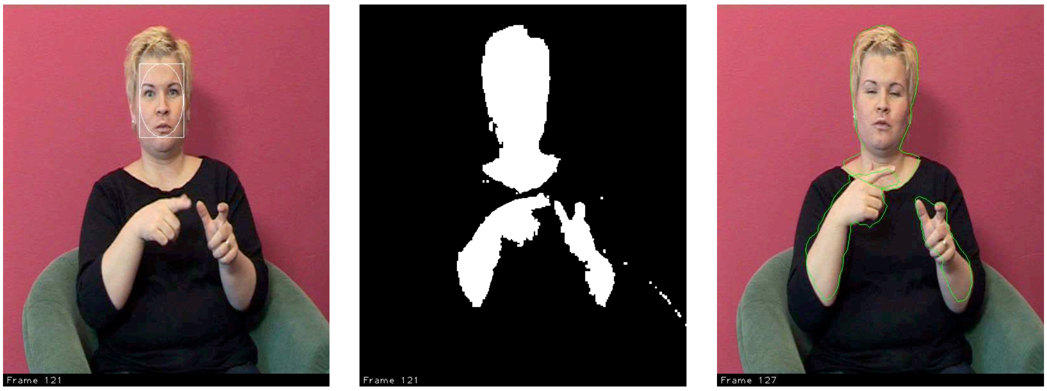

Figure 7.2: Example frames from the sign language video material. From left to right: Face detection, skin-color filtering, active shape models of skin regions.

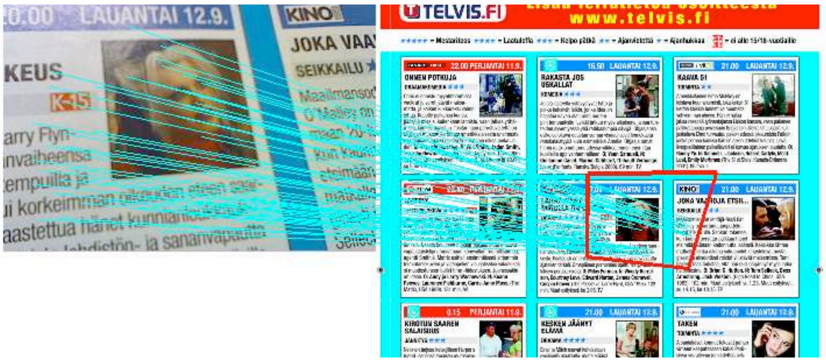

Figure 7.3: An example of matching an image captured with a mobile phone and the corresponding magazine page.

[5] Matti Karppa, Tommi Jantunen, Markus Koskela, Jorma Laaksonen, and Ville Viitaniemi. Method for visualisation and analysis of hand and head movements in sign language video. In C. Kirchhof, Z. Malisz, and P. Wagner, editors, *Proceedings of the 2nd Gesture and Speech in Interaction conference (GESPIN 2011)*, Bielefeld, Germany, 2011. Available online as `http://coral2.spectrum.uni-bielefeld.de/gespin2011/final/Jantunen.pdf`.

[6] Tommi Jantunen, Markus Koskela, Jorma Laaksonen, and Päivi Rainò. Towards automated visualization and analysis of signed language motion: Method and linguistic issues. In *Proceedings of 5th International Conference on Speech Prosody*, Chicago, Ill. (USA), May 2010.

[7] Xi Chen, Markus Koskela, and Jouko Hyväkkä. Image based information access for mobile phones. In *Proceedings of 8th International Workshop on Content-Based Multimedia Indexing*, Grenoble, France, June 2010.

[8] Xi Chen and Markus Koskela. Mobile visual search from dynamic image databases. In *Proceedings of Scandinavian Conference on Image Analysis (SCIA 2011)*, Ystad, Sweden, May 2011.