THE STONE AGE

Thomas Sikora

Technical University Berlin

SEGMENTATION AND INDEXING

Segmentation usually refers to the identification and labeling of segments of related pixels in images, indexing attempts to describe content of segments in a meaningful way. What is related and what is meaningful depends on the applications requirements.

Where are we with video segmentation and indexing? The answer: depending on the application at hand the field is either mature and settled or completely in its infancy.

Concentrate on TV type video applications and ask the above question in the context of video storage and retrieval. The purpose then is the description and indexing of video sequences for storage in a very large database – to allow users or software agents to identify shots of interest very efficiently. This is a typical MPEG-7 scenario [1]. Let us further assume that no text description is available – thus the descriptions will need to be extracted from the audiovisual information.

What kind of descriptions are required for indexing purpose? How is this linked to segmentation? Where are we with the state-of-the-art technology?

A CASE SCENARIO

Consider the following scenario: the user has a sample soccer game video containing person X - and would like to retrieve from the database:

a. the exact same video shot (also copies, possibly with different coding artifacts)
b. all soccer videos
c. all videos that contain a ball
d. all videos that contain person X, a greyhound dog and a jackknife.

Alternatively we might want to develop a system that is able to process and index all videos on a database, such that above retrieval tasks will be successful. This can be seen as the translation from pixels to text index keywords (at least for tasks c and d).

It is important to recognize that we obviously want the retrieval machine to perform like a human being, who is capable to perform all the above retrieval tasks without any problem.

Task a is the most simple one and can be successfully solved with today's technology - by analyzing each video's color and motion parameters - and comparing these parameters to the one in the user's sample video. There is no need to identify semantic objects in a scene nor any segments of particular interest. This is a typical low-level computer vision approach and the only one of the above tasks where a retrieval system might perform faster and more accuracy than a human being (i.e. if we want to measure whether two videos are pixel identical).

Task b: soccer videos contain distinct colors (mainly green), distinct camera motion and details. Thus also task b can be solved by analyzing the above parameters, plus additional analysis of patterns via edge distribution or texture analysis. In general, state-of-the art genre classification algorithms perform very well for sports scenes, music videos and TV news. However, compared to our human benchmark, state-of-the-art systems are far from competitive.

Task c concerns the recognition of a semantic object. Since a ball is a very simple one, it is sufficient to recognize circular segments in images containing the texture of a ball - which can be done using well established mid-level computer vision technology. However, since balls may vary drastically in texture, current technology is not very efficient and would return many false recognitions. The efficiency is much improved if only soccer games are being analyzed. The human observer will have no problem with this recognition task. The reason is that humans also analyze the context of a scene as well as depth clues from a monoscopic video.

Task d is the most difficult one for a retrieval system and reveals the general problem for indexing in retrieval systems. The system needs to recognize semantic objects. State-of-the-art computer vision systems need to employ very dedicated and specialized models for this task, one to identify people, one to identify dogs, and one to identify knifes. In addition to that it also needs to distinguish

person X from any other person, a jackknife from other knifes and a greyhound dog from all other dogs. Much research is focused towards recognizing whether people, even specific persons, are present in a scene – with still unsatisfactory success for most applications. For a human being all the above recognition tasks are easy.

Multimodal analysis can improve recognition rates of a retrieval system for the tasks above. State-of-the art audio analysis algorithms can recognize many events on the audio track, including a dog barking, a person speaking, which person speaking, applause, music, music genre [2,3].

WHERE ARE WE

Even with multimodal analysis the general problem persists: existing system only recognize very few semantic entities – with limited success. We are in the Stone Ages of information retrieval – this is particularly true for retrieval of audiovisual data.

Without doubt the case above scenario is one of great commercial relevance – far beyond retrieval applications. A system only capable of performing tasks a and b finds important applications, but is clearly of extremely limited value. These systems will simply not meet the demands of the information society, with its ever increasing flood of digital audiovisual information.

Retrieval systems will need to recognize large amounts of semantic entities in video to be commercially viable. The case study thus reveals the "semantic gap" between state-of-the-art technology and required solutions.

What we need is the "MediaGoogle", a search engine that translates text or spoken content queries into recognition tasks. Systems that understand context and content. Human language will need to be the ultimate description and index on the database.

The benchmark will need to be the human being – simply because it is mainly human beings that request feedback from a retrieval system. How many semantic entities a retrieval system needs to recognize? At least a good subset of what human beings might recognize. How many does the human being recognize? This depends on the age and experience. Humans learn meaning, appearance and context of semantic entities through continuous feedback from early on. Also retrieval search systems can learn audiovisual appearance and context through feedback. An example is the established relevance feedback strategy [4].

WHERE DO WE GO

It is vital that the retrieval research community recognizes the relevance of semantic entities for indexing. High-level MPEG-7 descriptors are only useful if they can be extracted automatically from the audiovisual data. This automatic extraction – the translation from pixels to semantic meaning – should be the prime research goal for the decades to come.

REFERENCES

[1] B. S. Manjunath*, Philippe Salembier**, Thomas Sikora (Editors), **"**Introduction to MPEG-7 - Multimedia Content Description Interface", *John Wiley & Sons Ltd*., 2002.

[2] Hyoung-Gook Kim, Juan José Burred, Thomas Sikora "How efficient is MPEG-7 for General Sound Recognition?", *25th International AES Conference "Metadata for Audio"*, London, UK, June 17-19, 2004.

[3] Hyoung-Gook Kim, Thomas Sikora, "Comparison of MPEG-7 Audio Spectrum Projection Features and MFCC applied to Speaker Recognition, Sound Classification and Audio Segmentation", *IEEE ICASSP* 2004, Montreal, Canada, May 17-21, 2004.

[4] T. Meiers, T. Sikora, I. Keller, "Hierarchical Image Database Browsing Environment with Embedded Relevance Feedback", *IEEE 2002 Int. Conf. on Image Processing*, Rochester, NJ, 2002.