Smart Media Management

Rainer Lienhart

Intel Corporation
Microprocessor Research Lab
2200 Mission College Blvd.
Santa Clara, CA 95052 - 8119
Rainer.Lienhart@intel.com

Introduction

A system is called 'smart' if a user perceives the actions and reactions of the system as being smart. Media is therefore managed smartly if a computer system helps a user to perform a large set of operations efficiently, fast and conveniently on a large media database. Such operations are

- Searching,
- Browsing,
- Manipulating,
- Sharing, and
- Re-using.

The term media usually denotes images, videos and audio data.

Automatic media content analysis

The more the computer knows about the media it manages, the 'smarter' it can be. Thus, algorithms which are capable of extracting semantic information automatically from media are an important part of a smart media management systems. In our lab, we have directed our attention towards

- Reliable shot detection,
- Text localization and text segmentation in images, web pages and videos, and
- Automatic semantic labeling of images.

Many other automatic content analysis algorithms have been developed by the research community, about which a good overview can be found at the *Video Content Analysis Home Page*[5] and in the past proceedings of *SPIE Storage and Retrieval for Media Databases*.

RELIABLE SHOT DETECTION. A shot is commonly defined as the uninterrupted recording of an event or locale. Any video sequence in this world consists of one or more shots concatenated by some kind of transition effects. Therefore, shots are generally considered as the elementary units constituting a video. Detecting shot boundaries thus means recovering those elementary video units, which in turn provide the ground for nearly all existing video abstraction and high-level video segmentation algorithms. In addition, during video production each transition type is chosen carefully in order to support the content and context of the video sequences. Therefore, automatically recovering all their positions and types may help the computer to deduce high-level semantics. For instance, in feature films dissolves are often used to convey a passage of time. Also dissolves occur much more often in features films, documentaries, biographical and scenic video material than in newscasts, sports, comedies and shows. The opposite is true for wipes. Therefore, automatic detection of transitions and their type can be used for automatic recognition of the video genre.

The vast amount of research in automatic shot boundary detection techniques in recent years documents the importance of reliable shot detection. A recent survey and practitioner's guide about the current state of the art in automatic shot boundary detection can be found in[1]. The survey emphasizes algorithms specialized in detecting specific types of transitions such as hard cuts, fades and dissolves. Representative of each concept a few sound and thoroughly tested approaches are present in detail, while others are just listed. It is state of the hart to detect hard cuts and fades at a high hit rate of 99% and 82% and at a low false alarm rate of 1% and 18%, respectively. Dissolves are more difficult to detect and the best approaches report hit and false alarm rates of 75% and 16% on a representative video test set.

TEXT LOCALIZATION AND TEXT SEGMENTATION IN IMAGES, WEB PAGES AND VIDEOS.

Extracting truly high-level semantics from images and videos in most cases is still an unsolved problem. One of the few exceptions is the extraction of text in complex backgrounds and cluttered scenes. For such text occurrences novel algorithms for detecting, segmenting and recognizing text have been developed recently[5]. These extracted text occurrences provide a valuable source of high-level semantics for indexing and retrieval. For instance, it enable users of a video database to query for all movies featuring John Wayne or produced by Steven Spielberg. Or it can be used to jump to news stories about a specific topic since captions in newscasts often provide a condensation of the underlying news story.

Figure 1: Example of visual text extraction in videos.

Detecting, segmenting and recognizing text in non-text parts of web pages is also a very important issue. More and more web pages present text in images. Existing document-based text segmentation and text recognition algorithms cannot extract such text occurrences due to their potentially difficult background and due to the large variety of text color used. The new algorithms allow to index the content of image-rich web pages properly. Automatic text segmentation and text recognition might also help in automatic conversion of web pages designed for large monitors to small LCD displays of appliances, since the textual content in images can be retrieved.

Our latest text segmentation method is not only able to locate text occurrences and segment them into large binary images, but also to label each pixel within an image or video whether it belongs to text or not[2]. Thus, our text detection and text segmentation method can be used for object-based video encoding. Object-based video encoding is known to achieve a much better video quality at a fixed bit rate compared to existing compression technologies. However, in most cases the problem of extracting objects automatically is not solved yet. Our text localization and text segmentation algorithms solve this problem for text occurrences in videos. As can be seen in Figure 2, the multiple video object video (multiple VOP in Figure 2) achieved a PSNR about 1.5 db better than the single object encoded MPEG-4 video. Thus, encoding the text lines as rigid foreground objects and the rest of the video separately achieved a visual much better quality.

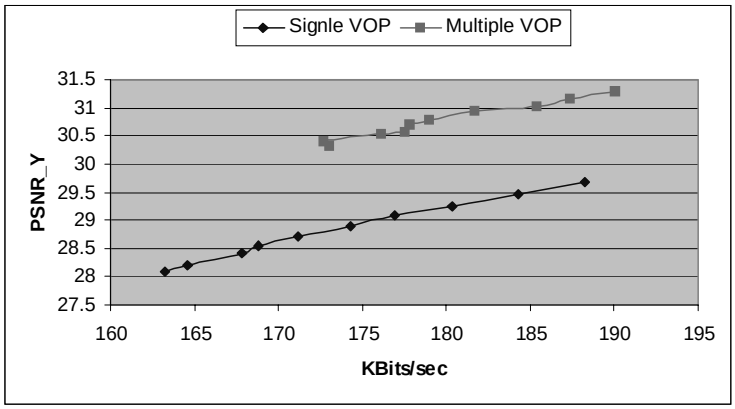
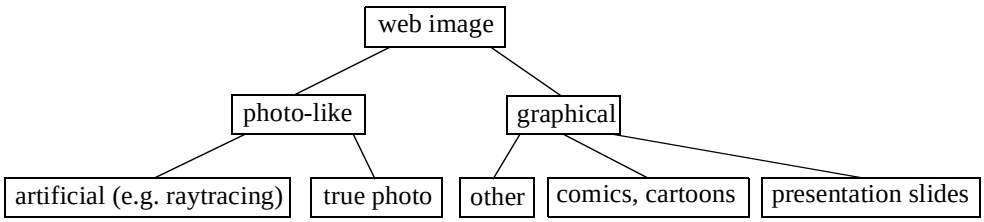


Figure 2: Example of the gain in PSNR by using object-based video encoding (MPEG-4) versus simple frame-based encoding

SEMANTIC LABELING OF IMAGES.

Numerous research work about the extraction of low-level features from images and videos has been published. However, only recently the focus has shifted to exploiting low-level features to classify images and videos automatically into semantically meaningful and broad categories. Examples of broad and general-purpose semantic classes are outdoor versus indoor scenes and city versus landscape scenes[3]. In one of our media indexing research

projects, we crawled about 300000 images from the web. After browsing carefully through those images, we came up with the following broad and general-purpose categories:



Although using only simple low-level features such as the overall color diversity in the image, the average noise level in the images, and the distribution of text line positions and sizes, our classification algorithm achieved an accuracy of 97.3% in separating photo-like images from graphical images on a large image database. In the subset of photo-like images, true photos could be separated from ray-traced/rendered image with an accuracy of 87.3%, while with an accuracy of 93.2% the subset of graphical images was successfully partitioned into presentation slides and comics. Sample images illustrating the chaos before and the order after their classification are shown in Figure 3[4]. Future research will be directed towards increasing the number of categories which can be classified automatically and will have to explore how joint classification can be done accurately and efficiently.

Figure 3: (a) Before classification, (b) after automatic classification

Media browsing/manipulation

Figure 4: Video browsing example

Although automatic media content analysis capabilities provide the basis of a smart media management system, efficient methods to browse a media database in a random but directed way are equally important. One potentially useful video browsing paradigm is shown in Figure 4. In the center, a normal video player allows to navigate through the currently selected video. While playing, every 3 seconds the whole video database is queried for shots which are most similar to the currently visible video sequence. The result of the query is shown as a decorative border around the main video player (see Figure 4). At any time a user can select any of those similar shots as the current video. In the example, similarity is based on color, however, any similarity measure can be applied. For instance, similarity based on the text visually occurring in a video sequence can be quite useful to browse through a database of newscasts recorded from a diverse set of broadcast channels.

Another equally important task is automatic video abstraction: A video abstract is a sequence of still or moving

images (with or without audio) presenting the content of a video in such a way that the respective target group is rapidly provided with the concise information about the content while the essential message of the original is preserved. Different abstraction algorithms for edited video (newscasts, feature films) and raw video (home video and raw news footage) have been developed in the past, but even better methods are needed for the future.

OUTLOOK. Many interesting challenges are still open and wait for being attacked by researches. SPIE's conference on 'Storage and Retrieval of Media Databases' is one of the major research conference, where people gather and exchange their latest research results. A new special feature track is on peer to peer media sharing and distributed media searching and indexing (see http://www.spie.org/Conferences/Calls/02/pw/confs/ei23.html). The interested reader is also encouraged to check out www.videoanalysis.org for more information and related work about smart media management.

References

[1] Rainer Lienhart. Reliable Transition Detection In Videos: A Survey and Practitioner's Guide. MRL technical report MRL_VIG000002-01, Intel Corporation, 2001. to appear in International Journal of Image and Graphics (IJIG).

[2] Axel Wernicke and Rainer Lienhart. On the Segmentation of Text in Videos. IEEE Int. Conference of Multimedia and Expo (ICME2000), Vol. 3, pp. 1511-1514, July 2000.

[3] Aditya Vailaya. Semantic Classification in Image Databases, PhD thesis, Department of Computer Science, Michigan State University, 2000, http://www.cse.msu.edu/~vailayaa/publications.html.

[4] Alexander Hartmann. Automatic Classification of Images on the Web. Master thesis. University of Mannheim, August 2000.

[5] *Video Content Analysis Homepage* at www.videoanalysis.org or www.videoanalysis.de