Multimedia Content Analysis and Indexing for Filtering and Retrieval Applications

N. Dimitrova
Philips Research
nvd@philabs.research.philips.com

Abstract

Today's multimedia landscape depends on advances in data compression technology coupled with high-bandwidth networks and storage capacity. Given its size and nature, how can multimedia content be analyzed and indexed? This paper surveys techniques for content-based analysis, retrieval and filtering of digital images, audio and video and focuses on basic methods for extracting features that will enable indexing and search applications. [ed.]

Keywords: data compression, multimedia, image analysis, video analysis audio analysis. [ed.]

Introduction

Data compression coupled with the availability of high-bandwidth networks and storage capacity have created the overwhelming production of multimedia content. In addition, the introduction of digital video will completely change the landscape of the entire video value chain. For content producers, advertisers, and consumers, there will be increased availability and increased challenges to manage the data. Users in the consumer and corporate domains will be given an overwhelming and confusing number of traditional and Internet media viewing choices and will look for ways to help them manage these choices [2,14,15,24,26,34,49,92]. Image and video archives in broadcast studios, corporate archives of multimedia collaborative sessions, video conferencing sessions, and educational content require tools for providing quick overview and transparent access. From a content production point of view, a broadcast studio archive will produce video at a rate of 19.2 Mb/s which translates into 207GB storage per day, assuming that only new content is broadcast. At that rate, the studio archive will require 75TB per year which means that in 14 years digital broadcast studios will have to cope with data in the petabyte range. The sheer size of the stored video data will pose serious issues for content owners to find and reuse some of the archived material. Content management for studio archives is just one of the many applications that incorporate tools for content analysis and retrieval of multimedia data.

Content management tools will aid in applications that will facilitate effective access, interaction, browsing and display of complex and inhomogeneous information consisting of images, video and audio. Such tools are important in various cases of professional and consumer applications such as education, digital libraries, entertainment, content authoring tools, geographical information systems, bio-medical systems, investigation services, surveillance and many others [64].

In this paper, we survey the techniques for content-based analysis, retrieval and filtering of digital images, audio and video. We will focus on basic methods for extracting features that will enable indexing and search applications. The deployment of a variety of these methods will enable powerful tools for both professionals and consumers to cope with multimedia data. Although the goal of these methods is content understanding which stems from computer vision systems, the methods surveyed here would be more similar to database methods for indexing. This is because a gap exists between these two aspects of the retrieval problem: databases do not provide content analysis and segmentation, and vision systems do not provide database query capabilities. The data acquisition in traditional databases relies primarily on the user to type in the data. Similarly, in the past image and video databases provided keyword descriptions of the visual descriptions of the visual data. However, the annotation based description is being augmented or replaced by automatic methods for feature extraction, indexing and content understanding [11,17].

This paper is organized as follows. Section 2 provides a survey of the methods for image analysis and retrieval. Section 3 describes techniques for video analysis, retrieval and filtering. Section 4 presents methods for audio based analysis and retrieval. In section 5 we provide a high level survey of systems

and standardization efforts in content description and retrieval.

Methods for Image Analysis and Retrieval

Image analysis is concerned with extraction of features for content representation. Often, extraction of low-level features such as color, texture and shape are used for this purpose. On the other hand, the intent of most image retrieval systems is to give the user tools to search for images using higher-level semantic descriptions. For example, the user may want to find all the "city" pictures or the "beach" pictures from the last vacations [26,29,54]. Currently one of the most challenging research topics is to provide the link between low level and higher-level features to offer meaningful content based image retrieval. Researchers have recognized that most objects and high-level concepts cannot be automatically extracted in a reliable manner. For this reason semiautomatic methods are developed where the user supplies a categorization label or annotation.

In the following sections, we describe various methods for feature extraction, analysis, and retrieval.

Analysis and indexing based on color, shape and texture

Image color plays a very prominent role in image analysis and retrieval. Many algorithms use a specific color space such as RGB (Red, Green and Blue), or HSB (Hue Saturation and Brightness). Computers use RGB color scheme, however, human perception of colors is closer to HSB. Color histograms are used to quantify the number of pixels for each color value. As such, it can be represented using a vector notation. The histograms of different images can be compared using distance measures such as L1 and L2. If the images are compared using the entire histograms then variations in position are neglected. For example when we compare the image with its flipped version the comparison will result in 100% similarity. An alternative to this "global color" image retrieval is to use location information and to compare images based on a subdivision of the image into smaller images. The simplest method is the rectangle subdivision of the image. However, significant color regions are extracted and compared using their color and location information. Niblack et al. [48] have used color cross correlation for color indexing and retrieval.

An alternative approach to color histograms is to use transformation such as Discrete Fourier transform or Wavelet transform. For example, Jacobs et al. [33] have used multiresolution wavelet decomposition for image representation and indexing. The coefficients from the decomposition are truncated, quantized, and normalized to produce "signatures." These image representations are then compared using a modified L1 distance measure.

Texture is a characteristic of a similarly patterned region in an image [39,41]. *Statistical* approach uses measures for variation of intensity in a texture window. Example measures use contrast, coarseness and directionality. *Structural* texture analysis extracts connectivity, density and regularity [73] in the image regions.

Shape based representation and retrieval is used to find images with objects of interest. An example is automatic classification and labeling of images with faces [79]. Shapes can be characterized with features such as color, edges, and texture. One approach is to represent these features in a multidimensional space and use multidimensional analysis and clustering techniques. Another approach for object detection is to start with a segmentation step and to merge the smaller regions using connected components algorithms to find objects with certain characteristics. For example in face detection, first, skin tone is used to find potential regions of interest, followed by higher level analysis to merge regions that might comprise a face [10,58]. A special case of shape retrieval is sketch-based retrieval [34]. The normalized image is first subjected to an edge detection step. An edge map is then extracted either from the full image or from regions comprising the image.

Methods for Video Analysis, Retrieval and Filtering

Video is a content rich medium in which actions and events in time and space comprise stories or convey particular information. Some of the methods for image and audio analysis can be used in video analysis. For example, image analysis and retrieval methods can be applied to selected representative frames extracted from video clips [50]. Image retrieval by color can be used to cluster shots by clustering their representative frames [19]. For example, Figure 1 depicts clustered shots using similarity between representative histograms for shots. However, there are unique characteristics that make video a more challenging medium than both image and audio to analyze and understand [56]. In many cases, researchers agree that indexing video aid in providing access and search methods for the current application needs without achieving full content understanding. In general, we can distinguish different ways of analyzing and searching video*: video summarization, video parsing, motion and event analysis*. Each of these methods has its own challenges and approaches to "fake" the video understanding problem. In what follows, we briefly review these challenging research issues, and the algorithms developed so far to address them.

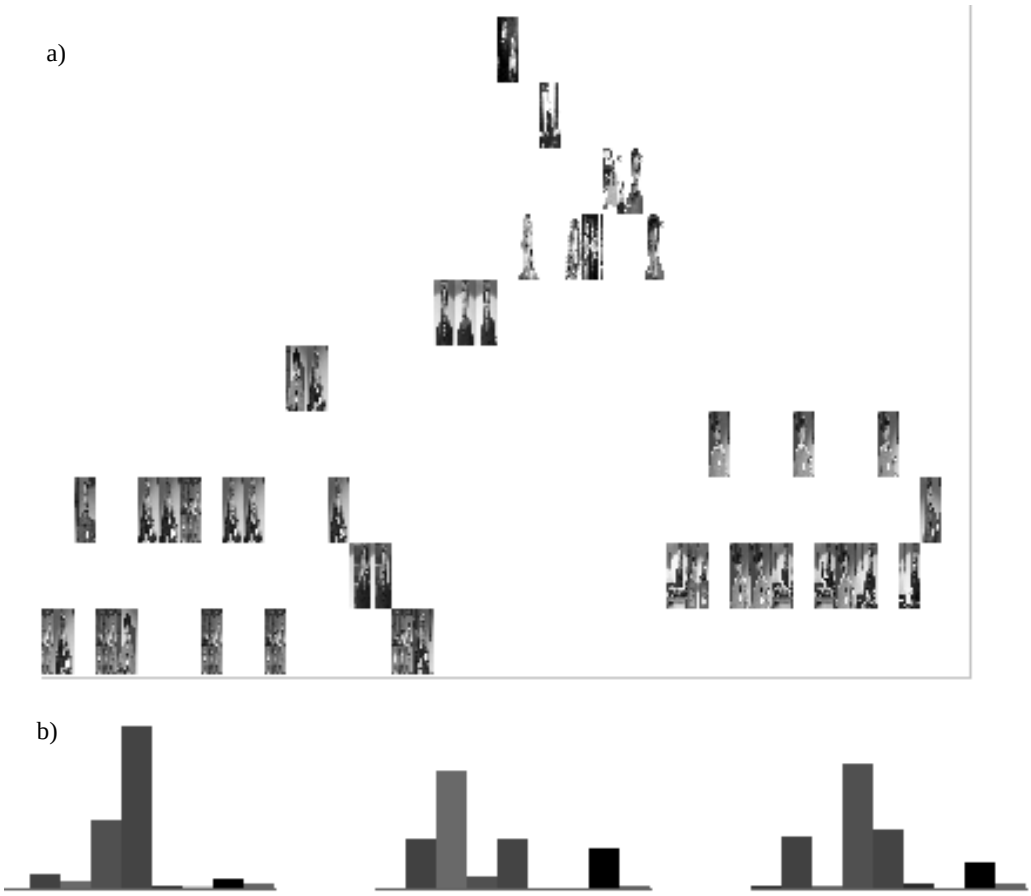


Figure 1. Shot clustering based on histogram representation:
a) Video shots from a game show clustered into nine families.
b) Three top histograms representing the largest three clustered keyframe families

Video summarization

Video summarization is the process of extracting abstract representation that will compress the essence of the video in a meaningful manner. This process enables organization of video data according to its temporal structure.

The simplest video summarization is pictorial summarization built from selected frames from the video [18]. Full video abstraction is the process of creating a presentation of visual, audio, and textual information, which should be much shorter than the original video [12,45]. This abstraction process is similar to extracting summaries from text documents. That is, we need to extract a subset of video data from the original video that has key-frames or highlights as entries for shots, scenes or stories. The result of the abstraction process forms the basis not only for video content representation, but also for content-based video browsing. Automatic summarization of video content in terms of extracting video highlights is an even more challenging research topic since it requires more high-level content analysis. A successful approach is to utilize information from multiple sources, including sound, speech, transcript and image analysis of video. The InforMedia project is a good example of this approach, which automatically skims documentaries and news videos with textual transcriptions by first abstracting the text using classical text skimming techniques and then looking for the corresponding parts in the video [13]. However, using such a text (keyword) driven approach may not yield satisfactory results in other categories of video where soundtracks contain music and other sound effects in addition to speech.

To summarize video, many research papers suggest video shot detection methods. Some of these methods are based on comparing pixel differences between frames [44, 47, 60,64], histograms, edge content or DCT coefficients. Other techniques have been applied in the compressed domain [3,25,66,82,87,], some of which take advantage of the compression process. Performing cut detection using only the DCT coefficients represents a midway approach because it does not require full decompression.

One of the first approaches proposed by Arman et al. [3] uses a DCT approach on both JPEG and MPEG streams. For MPEG streams, only I-frames are analyzed. This implementation employed a two-step approach. Video frames are compared based on their representation using a vector of subsets of DCT coefficients. Then the normalized inner product is subtracted from one and compared to a threshold. If a potential cut is detected, the images can be decompressed for further processing.

A multi-pass approach has been used by Zhang et al. but their technique also analyzes the B- and P-frames in an MPEG stream [89,91]. The first two passes compare the images based on DCT coefficients with different skip factors on I-frames. In another pass, the number of motion vectors is compared to a threshold. If there are fewer motion vectors than some threshold, a scene break is determined.

Kobla et al. also reported on video segmentation using DCT coefficients [35]. Their method is similar to Zhang's method in that it counts the motion vectors for the predicted blocks if it is an MPEG stream. If they determine that it is a Motion JPEG stream, they switch to DCT comparison and sum the square of differences of the DC coefficients between successive I frames.

Yeo et al. investigated using only the DC values of the DCT coefficients for frame comparison in the compressed domain [82,83]. They sum the DC differences between successive frames. If the difference is the maximum in a temporally sliding window and if it is n times larger than the next largest peak in the same window of frames, then it is a cut. They also detect "gradual transitions" (e.g. dissolves, fade in and fade out) by comparing each frame to the following kth frame over some time interval. The value for k should be larger than the time interval.

Zabih et al. [88] developed a method for detecting cuts using edge detection. Canny's algorithm is used as the basis for the edge detector. Originally, they checked the spatial distribution of entering and exiting edge pixels. Shen et al. have recently used edge detection in the compressed domain. The edges are extracted directly from the compressed image [66]. Then, Hausdorff distance histograms are obtained for each region by comparing edge points extracted from successive I frames. The histogram of the whole frame is obtained by merging the histograms of subregions in multiple passes. The merging algorithm is designed to increase the SNR of true motion during each pass while suppressing the mismatch information introduced by the noise.

Figure 2. Extraction of visual summary from video by automatically removing repeating and unicolor frames.

Hampapur and his colleagues [27] present a model driven approach to digital video segmentation. The paper deals with extracting features that correspond to cuts, spatial edits, and chromatic edits. The authors present extensive formal treatment of shot boundary identification based on models of video edit effects.

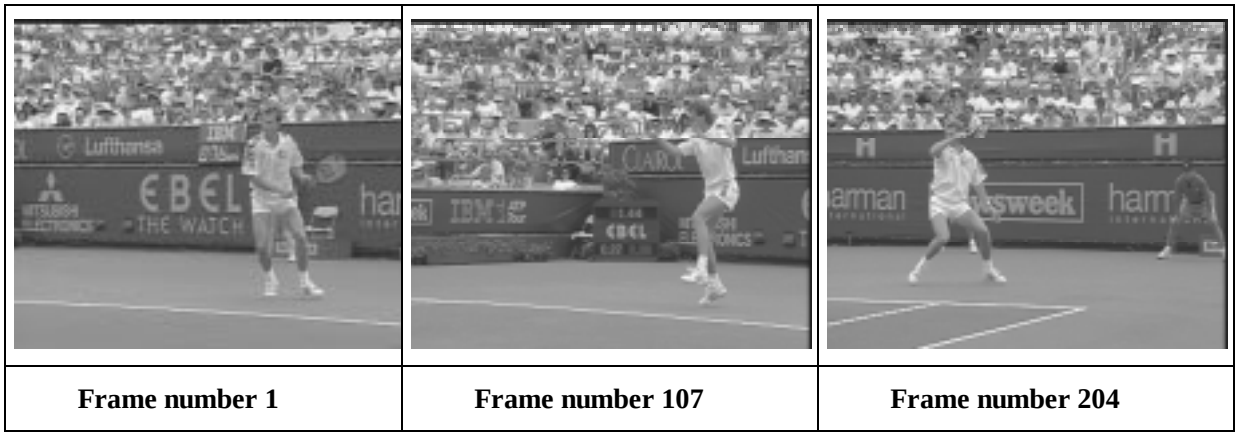
All of the above techniques have reported good results for cut detection. However, a comparison of algorithms to detect shots boundaries has been performed by Boreczky and Rowe [7]. They selected and implemented some of the above algorithms. Their results showed that DCT based algorithms had the lowest precision for a given recall. This result was expected due to a large number of false positives generated because of random noise in the black frames between commercials.

Visual summarization can be performed automatically using the cut detection methods described above. The frames that best represent the content of shots (i.e. video segment between two consecutive cuts) are called keyframes (see Figure 2.) The representational power of a set of keyframes depends on how they are chosen from all frames of a sequence because the same frame could not be a representative frame under different context [18,42,46].

An alternative approach to selection of keyframes is extraction of composite images from video shots. These images, called "mosaic" images or panoramic overviews, are derived using camera motion detection methods as well as image composition methods [5,6,70,72]. The mosaic representation is a good spatio-temporal synthetic representative visual representation for a video shot (see Figure 3). Mosaic extraction consists of two steps: motion estimation and motion accretion. At a given instant of the generation process, we need to merge the incoming frame with the current mosaic composed from previous frames. A global motion estimation algorithm is used to find the motion parameters, to merge it correctly. After motion estimation the current mosaic is computed by using a weighting function to reject parts of the image that do not belong to the background before blending the current mosaic and the warped incoming frame.

Video structure parsing

Video structure parsing is an involved process in video content analysis, which extracts temporal structural information of video segments [23,42,69,91]. This process enables us to organize video data according to their temporal structures and relations and thus build an abstracted view of a video program. It involves not only detection of temporal boundaries but also identification of meaningful composition of temporal



Frame number 1	Frame number 107	Frame number 204

Figure 3. Mosaic extracted from 239 frames. Three frames from the sports input sequence are shown.

elements in video. Ideally, these composition primitives should be categorized in a representation similar to storyboards used in filmmaking. This representation can lay out the frames in a hierarchy where the top-level consists of sequences or stories, which are composed of sets of scenes. Scenes are further partitioned into shots. Each shot contains a sequence of frames recorded contiguously and representing a continuous action in time or space.

Swanberg et al. [69] proposed one of the first approaches to content parsing. They proposed a new set of tools that could be used to semi-automatically segment video data into domain objects; process the video segments to extract features from the video frames; represent desired domains as models; and compare the extracted features and domain objects with the representative models. For example, using the TV news as a domain we can represent the anchor shots using an appropriate model. The model would have a "talking head" describing the anchor shot. The article suggests the representation of episodes with finite automatons, where the alphabet consists of the possible shots (i.e. events) making up the continuous video stream and the states contain a list of arcs, i.e., a pointer to a shot model and a pointer to the next state.

In contrast, M. Yeung et al. describe content characterization by a two step process of labeling, i.e., assigning shots that are visually similar and temporally close to each other the same label, and model identification in terms of the resulting label sequence [78,79]. Three fundamental models are proposed: dialogues, action and story unit models. Each of these models has a corresponding recognition algorithm.

Object, motion and event analysis

Often in queries of video clips, it is desirable to identify and recognize objects and sub-regions within the viewable image. Description of objects would enable more complex representation than the previously described overall color or texture representation. This will provide basis for interesting queries about objects and their behavior [76,77]. The object-oriented compression scheme standardized by MPEG-4 provides an ideal data representation for supporting such indexing and retrieval schemes. This will also simplify the task of video structure parsing and keyframe extraction since much of the content features needed in these processes such as object motion are readily available. A framework to utilize such content information in video content representation, abstraction, indexing and browsing was proposed in VideoQ [10].

Features that are easily extractable from video, such as color, texture, shape, structure, layout, and motion, cannot be easily mapped into semantic concepts, such as "Jimmy's birthday." However, all these features can be extracted automatically from the visual domain and used to build higher level descriptions of the video used for retrieval and filtering applications. For example, color segmentation is used as a basic step in face detection [59,71,76]. Motion information on the other hand could be extracted directly from uncompressed or compressed data [18,39,51,59,61,72] and used independently or in concert with other methods for retrieval.

Detecting Prominent Objects

In this section, we discuss methods for detecting prominent objects in video such as faces and superimposed text and motion information

There are many methods in the literature for face detection (see for example [11,56,60,71,74,76].) These methods coupled with the algorithms for transcript analysis and text detection can be used for finding names of people starting with an image and vice versa find the video related to a person given the name [59]. There are two classes of face detection methods: feature-based methods and classification-based methods. The feature-based methods locate different facial features and use their relationship to detect faces. For example, Yow et al. [85] use a set of spatial filters to detect possible feature points. The detected feature points are grouped using geometric and gray level constraints to form face candidates. A probabilistic network evaluates the likelihood of a candidate to be a face. Leung et al [36] use a set of Gaussian derivative filters to extract locations for facial features, such as the eyes, nose, and nostrils. The spatial arrangement of the located features is considered a random graph and the detection of a face is treated as a random graph-matching problem. Tankus et al, [71] consider the faces as three-dimensional objects consisting of convex and concave regions to develop an attentional operator that can extract regions of the eyes and hairs in images.

The face detection system proposed by Yang and Huang consists of three levels [81]. The first level uses a set of rules to locate face candidate regions in the input image. The next level uses another set of rules to operate on face candidate regions to perform further screening. Finally, the valid face regions are established at the third level by performing facial feature extraction.

The examples of classification-based methods for face detection are the systems by Rowley [59] and Sung [68]. The system developed by Rowley et al is a neural network-based system. It consists of two stages. The first stage uses a feed-forward neural network to classify every possible sub-image of a certain size as a face or non-face region. Neural networks use the sub-images at several scales to detect faces of different size. The second stage of the system consists of an arbitrator unit that merges the output of the neural network to eliminate overlapping detections. The most interesting aspect this system is the use of a bootstrapping technique to collect non-face training examples for neural network learning. Sung and Poggio also use a learning approach to detect faces [68]. In their

method, partitioned face pattern sub-images are grouped into few face and non-face model clusters. Sub-images at each position are matched against each cluster to determine the presence of faces. The advantage of the above two approaches is obvious; these systems require less a-priori domain knowledge and are relatively easier to adapt to detect other targets. However, the need to classify images at different resolutions to obtain invariance to scale brings in extra computation effort.

An important source of semantic information in the video track is the overlaid text in the video frame. This information is complementary to the visual, audio and transcript information (closed captioning or teletext) [21]. This text can be used in conjunction with shot detection algorithms for video indexing to generate important keyframes with anchor's name or with scores from a football game. On the other hand, scene text gives us a clue about the video content. This is very attractive, as it is much cheaper computationally to analyze text, rather than to analyze the visual content of the video. By keeping track of text patterns, we can find out if the text is scrolling, static or flying. eg. presence of scrolling text means beginning or ending of programs.

Ohya et al [51], perform character extraction by local thresholding and detect character candidate regions by evaluating gray level difference between adjacent regions. They merge detected regions that exist close to each other with similar gray levels to generate character pattern candidates. Hauptmann and Smith [28] use the spatial context of text and high contrast of text regions in scene images to merge large numbers of horizontal and vertical edges in spatial proximity to detect text. Lienhart [38] use a non-linear RGB color system to reduce the number of colors. A subsequent split-and-merge produces homogeneous segments having similar color. Further, they use multiple heuristics that characters are in the foreground, monochrome, rigid with size restrictions. Shim et al., use a generalized region labeling algorithm to find homogeneous regions for text segmentation and extraction [67]. The foreground images are clustered to find the color and location of text.

Extraction of object descriptions and trajectories was proposed in using an algebraic framework for data modeling and formulation of a query language [16,17]. This approach relies on the spatio-temporal nature of video streams represented in a dual structure which consists of object hierarchy and motion (temporal) hierarchy [17]. A clustering process is used to produce candidate trajectories in video. Each motion trajectory is described using chain code or spline-like representation which takes into account the temporal dimension. The objects of interest are represented by object-motion-video triplets. Meng et al. propose object indexing, camera motion, prominent moving objects and shape extraction and processing to retrieve semantic context of video [43]. Schonfeld et al [63] emphasize the concepts of working on compressed data and using the critical motion compensation information produced by the encoder's motion based prediction, specifically for immediate object tracking and video retrieval. In this case, only the objects of interest (intruder in a surveillance video) are detected and directly tracked in the bitstream without generating an object index. Sahouria [61] developed a system to analyze and index surveillance videos where the motion vectors from MPEG 1 compressed video formed the sole input to the system. From these the trajectories of objects in a fixed scene are extracted and represented by their wavelet transforms. Recently the idea of a motion trajectory descriptor was proposed to the MPEG 7 standard that embodies all these efforts for motion description [65].

While visual content is a major source of information in a video program, an effective strategy for recognizing events and understanding video is to use information carried in the other media components, such as text (superimposed on the images, or included as closed captions), audio, and speech [62]. A combined and cooperative analysis of these components would be far more effective in the characterization of the video segments.

Audio Analysis and Retrieval

Audio signal can be stored or delivered by itself, or as an accompanying medium to video. In both cases audio is a powerful medium for capturing important information. Speech is normally used to convey the meaning of an activity or a story and in general, music is very good at conveying moods or emotions. As a result, the interest in exploring retrieval in audio archives and analysis of soundtracks for video indexing is growing.

The methods developed for audio analysis and retrieval can be categorized into three different categories: audio characterization, speaker identification and keyword spotting.

Audio characterization

In *audio characterization*, audio information is classified into different categories: silence, speech, music, and other voice [80]. The classification method is based on a number of audio features extracted from each audio segment. These methods use audio features computed from one or more of the input data *packets*. The features computed could be average amplitude, average energy, candidate pitch, bandwidth, frequency spectrum, and Mel Fourier Cepstral Coefficients. These features combined with a variety of pattern classification methods can be used to detect almost all of the above events with reasonably good accuracy.

Speaker identification

The speaker identification can be performed only on those segments labeled as speech segments during audio characterization and classification. This stage includes extraction of mel-frequency cepstrum coefficients (MFCC), nearest neighbor classifier, and a pooling process. The MFCC are extracted by using a sliding window of 30ms with an overlapping of 20ms. A set of MFCC is extracted from each windowed data. The MFCC are then fed to the nearest neighbor classifier to generate k classification labels for each moving window. Finally, a pooling operation is performed to combine the k classification decisions from each moving window to generate a single decision for each segment. The decision thus generated represents the identity of the speaker.

Keyword/topic spotting

In addition to audio classification and speaker identification, keyword/topic-spotting techniques give important information for more detailed description. Current speech recognition systems perform well on a limited vocabulary for a single speaker. However, for the general conversational speech, such as multi-speaker, unconstrained context, large vocabulary, and continuous speech, existing speech recognition systems perform poorly especially for audio tracks coming from video segments. In the case of keyword/topic spotting, not all words have to be recognized correctly from the speech. By making use of the recurrence of certain words or structures, we can obtain satisfactory performance. To further improve the performance of keyword spotter in an environment with multi-speaker or background noise, we can also combine the keyword spotting technique with audio source separation approach mentioned above.

Audio-Video Analysis

The audio features extraction and analysis can also be used to detect the genre of the video, detect speakers, generate transcripts etc. The storyboard can then be hierarchically organized to facilitate easy browsing. By including additional features such as the audio track and closed-captioning along with the video stream, a more accurate and easily searchable representation can be created. Some digital media systems also incorporate this approach to video classification. Sethi et al. have been investigating audio-based methods for video analysis. In their earlier work, they showed the effectiveness of audio characterization and speaker identification for video indexing and classification. They used features such as short time energy, band energy ratio, pause rate, and pitch. Further, they have also established a framework for performing audio analysis directly on compressed video and audio bit streams.

Wold et al. have also developed an approach for audio classification. The static properties such as mean, variance, duration, and correlation for several acoustic features are utilized [80]. These features include loudness, pitch, brightness, bandwidth, and normalized harmonicity. Pfeifer et al have presented audio analysis operators such as volume analysis, frequency analysis, pitch analysis, frequency transition maps, fundamental frequency analysis and beat analysis, and introduced applications such as music indexing, and retrieval, and violence detection in movies [56]. Ghias et al. have developed a music indexing method that uses the melodic “contour” defined as the sequence of relative differences in pitch between successive notes [22]. They used a pitch tracking system and then encoded three possible relationships between pitches (up, down, same) representing situations where a note is above, below, or same level as the previous note. The paper considers three types of algorithms for pitch tracking that use autocorrelation, maximum likelihood and cepstrum analysis. For finding similar tunes, they used comparison between the two melodies using a string-matching algorithm that accounts for error tolerance. Beyerlein et al. have used a transcription system for recognition of speech in radio and television broadcasts [4]. The system uses continuous mixture density crossword HMM system based on MFCC features and Laplacian densities. A segmentation is first performed to obtain sentence-like partitions of the full broadcast. These segments are then clustered using data driven clustering method. They also perform channel and speaker normalization. The final transcript is then produced by using an adaptive multipass decoder starting with phrase bigram decoding using word-internal triphones and finishing with a phrase-trigram decoding using crossword models.

Emerging Systems, Standards and Applications

In the past few years, video retrieval systems evolved towards systematic integration of the visual, auditory and textual cues. Examples of such systems include QBIC, Photobook, Informedia, Vabstract, VisualSeek, VisualGrep, and CONIVAS. In addition there are commercial systems from companies such as Virage, ISLIP, BullDog [8], Excalibur [20] and Magnifi that provide advanced solutions for different application domains.

Systems

The Query-By-Image-and Video-Content (QBIC) [48, 58] system developed at IBM’s Almaden Research Center uses a variety of features for retrieving images from image/video database. The system is described as a set of technologies and associated software that allows a user to search, browse and retrieve image, graphic and video data from large on-line collections. The system allows image and video databases to be queried using visual features such as color, layout and texture. In QBIC the queries are matched pictorially so that users can match their perception of the visual features without using

words. The query is matched against a database of pre-computed features clustered meaningfully. With Query-by-Example (QBE) type queries, the user can select any thumbnail from the list of images within the database or specify an image and request retrieval of similar images. The user can also sketch an image or parts thereof for describing the query image.

The feature of generating video storyboards has also been added to QBIC. A storyboard consists of representative frames selected from subsequences within the video. Each subsequence is separated from the other by significant changes such as scene cuts or gradual transitions. Once the storyboard has been generated for the MPEG-1 compressed video sequences, the methods discussed above can be applied to these representative frames to retrieve video clips by content.

The Photobook System developed at the MIT's Media Labs is described as a set of interactive tools for browsing and searching images and image sequences [55]. Direct search on the image content is made possible through its semantics preserving image compression techniques using Karhunen-Loeve Transform (KLT) and the Wold Decomposition Methods. The Photobook allows search based on 2-D shape, gray level appearance and textural properties. The focus of this system is the semantics preserving image compression that replaces the image in the database with a set of parameters that can be used to reconstruct the image in its entirety. This differs from the other methods that find features from the image that can be used to perform similarity matching where the features are extracted from selected parts of the image and cannot be used to reconstruct the image.

The MoCA project, developed at University of Mannheim, is designed to provide content-based access to a movie database

Figure 4. A snapshot of the CONIVAS image retrieval by example.

[57]. Besides segmenting movies into salient shots and generating a digital abstract of the movie, the system also detects and recognizes the title credits and performs audio analysis on the audio track. The text detection component tracks moving text and performs OCR on the text. The audio analysis component detects silence, human speech, music and noise. The latter is further analyzed to detect violence in the scenes. This is done in conjunction with the visual analysis component. A related work from the same group can also detect the presence of commercials (advertisements) in the video sequence.

CONIVAS (CONtent-based Image and Video Access System) is a client-server based system developed at Philips Research [1] (see figure 4) The system employs cut detection for extraction of a storyboard used for browsing and retrieval from a digital studio archive. Features extracted from the keyframes are used for building an index of the content. Segmentation can be applied either in the compressed domain or the uncompressed domain. Feature extraction is performed either using low level visual features such as global or local color, shape, and texture, or using full text retrieval. In case of retrieval by visual features, images in the database are archived by extracting the relevant features and storing them in a database. At search time, the user enters an image as an example or composes a sketch and the system analyzes the input by extracting the visual features, then finds the closest images in the database based on these features. In addition, video segments can be retrieved using example query segments.

The VideoQ system developed by Columbia University classifies video characteristics in the compressed domain [9]. The system consists of three modules: parsing, visualization and authoring. The parsing module segments the video into shots. The shots are then analyzed for camera motion, object motion, shape and trajectory. The visualization module then extracts the key frames or objects from these shots. The authoring module allows the user to add special effects such as dissolves to the video sequence and cut-and-paste operations on the compressed sequence.

The InforMedia Digital Library developed at Carnegie Mellon University (CMU) is a full fledged digital video library under development [12]. The digital library user's interests will lie in short video clips and content-specific segments. These segments have been called skims by the authors. The system contains methods to create a short synopsis of each video. Language understanding is applied to the audio track to extract meaningful keywords. Each video in the database is then represented as a group of representative frames extracted from the video at points of significant activity. This activity may be abrupt scene breaks, some form of rapid camera movement, gradual changes from one scene to another, and points in the video where some keywords appear. Caption text is also extracted from these frames, which adds to the set of indices for the video.

The VisualGREP system [26] is developed as a follow-up to the ImageGREP project developed at the University of California at San Diego. The work is also done in conjunction with the MoCA project at the University of Mannheim. The system analyzes the video sequence at the frame level as well as at the shot level. It uses color and motion characteristics for classifying the video and also detects the presence of frontal human faces. The attempt is to use a normalized similarity distance measure based on human psychological studies. The system classifies the video into non-aggregated and partially aggregated and complete sequences based on the importance of the scene as well as the presence of visually important features. The key frames (and objects) can then be retrieved with the VisualSEEk image retrieval system. This system is specifically designed for image retrieval and can retrieve images based on color and spatial locations of objects / regions.

CAETI Internet Multimedia Library is a project at the Princeton Unversity to develop new tools and techniques for browsing and annotating single video clips and for navigating through large collections of video [86,87]. These techniques take into account the educational needs of students and teachers, the psychology of human-computer interaction, and the feasibility of implementing those techniques on low-cost computer systems in educational setting. IML is also developing cost-effective architectures which can supply new video and multimedia material to the classroom from content providers and can provide video within the classroom video cluster. In addition they have developed hypervideo tools in a new learner-centered curriculum on News Media and Politics, which examines the relationship between political events, politicians, and television and other news media.

MARS (Multimedia Analysis and Retrieval System) is a system developed at the University of Urbana Champaign in 1994 [53]. It is a Java-written web-based system supporting content-based image retrieval based on color, texture, shape, and any Boolean combinations of them. The novel part is the integration of DBMS techniques (query processing), IR techniques (Boolean retrieval model), and Image Processing techniques (image features). MARS supports image retrieval using relevance feedback and image composition. The system presents the users with a list of images and the user selects the closest images to the desired image. The process is repeated and as more images are selected the MARS engine dynamically refines the initial query as a better approximation to user's information need.

The ImageSearch engine, developed at the Imaging and Multimedia Group at Leiden University, searches an index of over 25 million images over the Web [37]. The system creates an abstract similarity of an image using sketches and representative icons for objects, such as faces, trees or stones. The object recognition in the ImageSearch engine uses a form of visual learning from positive and negative examples.

ViBE is the video indexing and browsing environment developed at the Purdue University [11]. The system uses shot detection method based on extraction of a high dimensionality feature vector and the use of a binary regression tree to estimate the conditional probability of a shot boundary for each frame. In order to capture the salient aspects of complex shots, they introduce the idea of a shot tree. The shot tree is a binary tree, which contains a set of representative frames from the shot. Furthermore the system provides pseudo-semantic labeling method for higher level semantic features (such as faces).

Research in content-based retrieval has also matured to the point of commercial introduction. Recently there has been proliferation of many commercial systems for content-based image, video and audio retrieval by companies such as Virage [78], ISLIP [32], Imagine [30], Magnifi [40], Excalibur [20]. Content management tools as an addition to a database management system are offered by Informix [31], Oracle [52], Cinebase [13] and BullDog [8].

Standardization

In order to make the content search ubiquitous and available across different domains and applications by providing a framework for interoperability, MPEG started the activity called 'Multimedia Content Description Interface' [46]. The goal for MPEG-7 is to answer the need to uniformly describe content for future reuse. MPEG-7, formally will standardise:

- A set of description schemes and descriptors that will form an ontology to describe multimedia content
- A language to specify description schemes, i.e. a Description Definition Language (DDL).
- A scheme for coding the description

MPEG -7 will provide a standardized description of various types of multimedia information. The normative part of the standard will focus on a framework for encoding the descriptors and description schemes. The standard will not comprise the extraction of descriptors (features) and will not specify search engines that will use the descriptions. Instead, the standard will enable the exchange of content between different content providers along the media value chain. In addition, it will enable development of applications that will utilize the MPEG-7 descriptions without specific ties to a single content provider.

In the process of standardizing the descriptions, it is expected that the current research results will be absorbed in MPEG-7 by 2001 when it will become an international standard. However, this will have an impact on availability of additional information along with the images and video segments for many applications. This fact will help focus the needs for research on content analysis topics, which are not available in the MPEG-7 description schemes.

Conclusions

Content-based retrieval provides solutions to help users quickly and easily find images, video and audio based on the inherent characteristics of the content. For example in large image, video, and audio archives users can pose queries that involve visual features that describe the content of the data in addition to textual annotations [74,75,84,90]. In consumer domain, viewers who do not want to waste time watching commercials will no longer need to. Parents who are concerned about sex, violence, or profanity on TV can eliminate them without stopping their child from watching an entire program. People who take many pictures can automatically categorize them and then quickly search through a home image or video library. In this paper we have described some of the basic techniques and systems for content based retrieval. As the technology becomes stronger, there will be more research and commercial systems exposing the solutions to the content retrieval problem and allow users to make search retrieval and filtering of digital media manageable.

References

1. M. Abdel-Mottaleb, N. Dimitrova, R. Desai and J. Martino, "CONIVAS: Content-based Image and Video Access System," *Proc. of ACM Multimedia*, pp 427-428, Boston 1996.

2. Y. Alp Aslandogan and C.T. Yu, "Techniques and Systems for Image and Video retrieval", IEEE Transactions on Knowledge and Data Engineering, Vol. 11, No. 1, 1999, pp. 56-63.

3. F. Arman, A. Hsu, and M-Y. Chiu, "Image Processing on Encoded Video Sequences," *Multimedia Systems* vol. 1, no. 5, pp. 211-219, 1994.

4. P. Beyerlein, X. Aubert, R. Haeb-Umbach, D. Klakow, M. Ulrich, A. Wendemuth, P. Wilcox, "Automatic Transcription of English Broadcast News," DARPA Broadcast News Transcription and Understanding Workshop, VA, Feb 8-11, 1998.

5. Michel Bonnet, "Mosaic representation for video shot description", p636, MPEG-7 Evaluation Ad-Hoc Meeting, Lancaster, February 1999.

6. Michel Bonnet, Benoît Mory, "Global motion analysis for searching and browsing of audio-visual data", to be published in WIAMIS'99 proceedings, Berlin 31/05-06/01 1999.

7. J.S. Boreczky, and L.A. Rowe, "Comparison of video shot boundary detection techniques," *Proc. IS&T/SPIE 1996, Storage and Retrieval for Image and Video Databases IV*, San Jose, February, 1996.

8. URL: BullDog, Inc. (http://www.bulldog.ca/bd/)

9. S-F. Chang, W. Chen, H. E. Meng, H. Sundaram and D. Zong, "VideoQ: An Automated Content Based Video Search System Using Visual Cues," *Proc. ACM Multimedia*, pp. 313-324, Seattle, 1994.

10. R. Chellappa, C.L. Wilson and S. Sirohey, "Human and machine recognition of faces: A Survey", *Proceedings of the IEEE*, Vol. 83, No. 5, pp. 705-740, 1995.

11. Jau-Yuen Chen, Cuneyt Taskiran, Alberto Albiol, Edward J. Delp, and Charles A. Bouman, "Vibe: A Compressed Video Database Structured for Active Browsing and Search", Purdue University, 1999

12. M. Christel, T. Kanade, M. Mauldin, R. Reddy, M. Sirbu, S. Stevens and H. Wactlar, "Informedia Digital Video Library," *Comm. of the ACM*, Vol. 38, No. 4, pp. 57-58, 1995.

13. URL: Cinebase Software (http://www.cinebase.com)

14. M. Davis, Media Streams, "An Iconic Visual Language for Video Annotation", Proc. *Telektronikk* 4.93, pp.59-71, 1993.

15. Y.F. Day, S. Dagtas, M. Iino, A Khokhar and A. Ghafoor, "Spatio-Temporal Of Video Data For On-Line Object-Oriented Query Processing," *Proc. IEEE International conference on multimedia computing and systems*, pp. 98-105, May 15-18, 1995.

16. N. Dimitrova and M. F. Golshani, "Video and Image Content Representation and Retrieval," *Handbook of Multimedia Information Management, Prentice Hall*, pp. 95-138,1997.

17. Nevenka Dimitrova and Forouzan Golshani, "Motion Recovery for Video Content Classification," ACM Transactions on Information Systems Vol 13, No. 4, Oct. 1995, pp 408-439.

18. N. Dimitrova, T.McGee, H. Elenbaas, "Video Keyframe Extraction and Filtering: A Keyframe is not a Keyframe to Everyone," *Proc. ACM Conf. on Knowledge and Information Management*, pp.113-120, 1997.

19. N. Dimitrova, J. Martino, L. Agnihotri and H. Elenbaas, "Color Super-histograms for Video Representation," IEEE Conference on Image Processing, Kobe, Japan, 1999.

20. URL: Excalibur Technologies (http://www.excalibur.com)

21. U. Gargi, S. Antani and R. Kasturi, "Indexing text events in digital video databases", *in International Conference on Pattern Recognition*, Brisbane, August 1998, pp. 916-918.

22. A. Ghias, J. Logan, D. Chamberlin and B.C. Smith, "Query by Humming," ACM Multimedia '95, pp. 231-236, 1995.

23. Y. Gong, *et al*, "Automatic Parsing of TV Soccer Programs," *Proc. Second IEEE International Conference on Multimedia Computing and Systems*, Washington DC, 15-18 May 1995, pp167-174.

24. W. Grosky, "Managing Multimedia Information in Database Systems," *Proc. Communications of the ACM*, Vol. 40, No. 12, pp. 73-80, 1997.

25. B. Gunsel, A. Ferman, and A. Tekalp, *Video Indexing Through Integration of Syntactic and Semantic Features*, IEEE Workshop on Applications of Computer Vision, 1996.

26. *A.* Gupta and R. Jain, "Visual Information Retrieval," *Proc. Communications of the ACM*, Vol. 40 No. 5, pp. 70-79, May 1997.

27. A. Hampapur, R. Jain, and T. Weymouth, "Digital Video Segmentation," *Proc. ACM Multimedia '94*, pp.357-364, San Francisco, CA, October 1994.

28. A. Hauptmann and M.Smith, Text, speech, and vision for video segmentation: The informedia project. *In AAAI Fall 1995 Symposium on Computational Models for Integrating Language and Vision 1995*.

29. W. Hsu, T.S. Chua, and H.K. Pung, "An Integrated Color-Spatial Approach to Content-Based Image retrieval, ACM Multimedia, pp. 305-313, 1995.

30. URL: Imagine Products, Inc. (http://www.imagineproducts.com)

31. URL: Informix Corporation (http://www.informix.com)

32. URL: ISLIP Media, Inc. (http://www.islip.com)

33. C. Jacobs, A. Finkelstein, and D. H. Salesin, "Fast Multiresolution Image Querying," ACM SIGGRAPH 1995, pp. 277-286, 1995.

34. T. Kato, T. Kurita, N. Otsu and K. Hirata, "A Sketch Retrieval Method for Full Color Image Database, Query by Visual Example," Proc. IEEE-IAPR-11, pp. 530-533, Sept. 1992.

35. V. Kobla, D. Doermann, K. Lin, and C. Faloutsos, "Compressed Domain Video Indexing Techniques using DCT and Motion Vector Information in MPEG Video," *Proc. IS&T SPIE, Storage and Retrieval for Image and Video Databases V*, Volume 3022, pp.200-211, San Jose, 1997.

36. T.K. Leung, M.C. Burl and P.Perona, "Finding faces in cluttered scenes using random labeled graph matching", *Proceedings Fifth Intl. Conference on Computer Vision*, pp. 637-644, Cambridge, MA, June 1995.

37. M.S. Lew, K. Lempinen, N. Huijsmans, Webcrawling Using Sketches, Proceedings VISUAL97 conference, San Diego 15-17 Dec 1997, pp 77-84.

38. R. Lienhart and F. Suber, "Automatic recognition for video indexing", SPIE conference on image and video processing, January 1996.

39. F. Liu and R. Picard, "Periodicity directionality and randomness: Wold features for image modeling and Retrieval," IEEE Transactions on Pattern Analysis and machine Intelligence, Vol. 18, No. 7, pp. 722-733, 1996.

40. URL: Magnifi Inc. (http://www.magnifi.com)

41. B.S. Manjunath and WY Ma, Texture Features for Browsing and Retrieval of Image Data, IEEE Transactions on Pattern Analysis and machine Intelligence, Vol. 18, No. 8, pp. 837-842, 1996.

42. *T. McGee and N. Dimitrova,* "Parsing TV programs for identification and removal of non-story segments", *Proc. of SPIE Conf. on Storage and Retrieval for Image and Video Databases*, San Jose, CA, USA, January, 1999.

43. J. Meng and S. Chang, "Tools for compressed domain video Indexing and Editing," *Proc. IS&T SPIE, Storage and Retrieval for Image and Video Databases V*, Volume 2670, San Jose, 1996.

44. C. Merlino, D. Morey and M. Maybury, "Broadcast navigation using story segmentation," Proc. of ACM Multimedia'97, Seattle, November 1997, pp.381-388.

45. M. Mills, J. Cohen, Y.Y. Wong, "A magnifier tool for video data." *Proc. CHI 92, Monterey, ACM Press*, pp. 93-98, 1992.

46. The MPEG-7 standardization (http://drogo.cselt.stet.it/MPEG)

47. A. Nagasaka and Y. Tanaka, "Automatic Video Indexing and Full-Search for Video Appearances," in E. Knuth and I.M. Wegener editors, *Visual database Systems*, Elsevier Science Publishers, Vol.II, Amsterdam, 1992, pp.113-127.

48. W. Niblack, X. Zhu, J.L. Hafner, T. Bruel, D. B. Ponceleon, D. Petkovic, M. Flickner, E. Upfal, S.I. Nin, S. Sull, B.E. Dom, "Updates to the QBIC System," *Proc. IS&T SPIE, Storage and Retrieval for Image and Video Databases VI*, Volume 3312, pp. 150-161, San Jose, 1998.

49. K.C. Nwosu, B. Thuraisingham and P.B. Berra, "Multimedia database systems a new frontier," *Proc. IEEE Multimedia* ,Vol. 4. No 3, pp. 21-23, July- September 1997.

50. C. O'Connor, "Selecting Key Frames of Moving Image Documents: A Digital Environment for Analysis and Navigation", *Microcomputers for Information Management*, 8(2), pp. 119-133, 1991.

51. J. Ohya, A. Shio, and S. Akamatsu. Recognizing characters in scene images, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol 16, 214-224, 1994.

52. URL: Oracle Corporation (http://www.oracle.com)

53. Michael Ortega, Yong Rui, Kaushik Chakrabarti, Kriengkrai Porkaew, Sharad Mehrotra, and Thomas S. Huang, Supporting Ranked Boolean Similarity Queries in MARS, IEEE Tran on Knowledge and Data Engineering , Vol. 10, No. 6, pp905-925, Dec. 1998.

54. Y.C. Park, F. Golshani, S. Panchanathan, K.S. Candan, "Interactive Classification of still and motion Pictures in VideoRoadMap," Proc. of SPIE Conference on Multimedia Storage and Archiving Systems III, Vol. 3527, November 1998, pp. 122-133.

55. A.P. Pentland, R. W. Picard, and S. Sclaroff, "Photobook: Content-based manipulation of image databases," International Journal of Computer Vision, Vol. 18, No. 3, pp. 233-254, 1996.

56. S. Pfeifer, S. Fischer, W. Effelsberg, "Automatic Audio Content Analysis," ACM Multimedia '96, pp. 21-30, Boston, MA, 1996.

57. S. Pfeiffer, R. Lienhart, S. Fischer, and W. Effelsberg, "Abstracting Digital Movies Automatically," Proc. Journal on Visual Communications and Image Representation, Vol. 7, No. 4, pp. 345-353, 1996.

58. URL: QBIC IBM (http://wwwqbic.almaden.ibm.com)

59. H.A. Rowley, S. Baluja and T. Kanade, "Neural network-based face detection", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 20, No. 1, p23-37, 1998.

60. Eli Saber and A. Murat Tekalp, "Frontal-view face detection and facial feature extraction using color, shape and symmetry based cost functions", *Pattern Recognition Letters*, Vol. 19, No. 8, pp. 669-680, 1998.

61. Emile Sahouria and Avideh Zakhor, "Motion Indexing of Video", International Conference on Image Processing 97, Santa Barbara 1997.

62. S. Satoh, Yuichi Nakamura, and Takeo Kanade, "Name-It: naming and Detecting Faces in News Videos," IEEE Multimedia, Vol. 6. No. 1., pp. 22-35, 1999.

63. Dan Schonfeld and D. Lelescu, "VORTEX: Video Retrieval and Tracking from Compressed Multimedia Databases: Affine Transformation and Occlusion Invariant Tracking from MPEG-2 Video", *Proc. IS&T SPIE, Storage and Retrieval for Image and Video Databases VII*, Volume 3656, pp. 131-143, San Jose, 1999.

64. B. Shahraray, Scene Change Detection and Content-Based Sampling of Video Sequences, *IS\&T/SPIE'95 Digital Video Compression: Algorithm and Technologies*, San Jose, 1995, February, Vol.2419, pp.2--13.

65. Alfred She, Sylvie Jeannin, Ali Tabatabai, Thumpundi Naveen, "Object Motion Trajectory", ISO/IEC JTC1/SC29/WG11/MPEG99/M4476, March 1999.

66. B. Shen, and I. Sethi, "Convolusion-based edge detection for image/video in block DCT domain," Journal of Visual Communications and Image Representation.

67. J-C. Shim, C. Dorai, and R. Bolle, "Automatic Text Extraction from Video for Content-Based Annotation and Retrieval," In Proc. of the International Conference on Pattern Recognition, pp. 618-620, 1998.

68. K-K. Sung and T. Poggio, "Example-based learning for view-based human face detection", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 20, No. 1, pp. 39-50, 1998.

69. D. Swanberg, C. F. Shu and R. Jain, "Knowledge guided parsing in video databases," *Proc. of SPIE Conf. on Storage and Retrieval for Image and Video Databases*, San Jose, CA, USA, February, 1993.

70. Y. Taniguchi, A. Akutsu, and Y. Tonomura, “PanoramaExcepts: Extracting and packing panoramas for video browsing,” *Proc. ACM Multimedia Conference*, Seattle, December 1997, pp.427-436.

71. A. Tankus, Y. Yeshurun and N. Intrator, "Face detection by convexity estimation", *Pattern Recognition Letters*, Vol. 18, pp. 913-922, 1997.

72. L. Teodosio, and Bender, W., “Salient Video Stills: Content and Context Preserved,” *Proc. ACM Multimedia 93*, pp. 39-46, Anaheim, CA, August 1993.

73. F. Tomita and T. Saburo, “Computer Analysis of Visual Textures,” Kluwer, 1990.

74. Y, Tonomura, Abe, S. “Content oriented visual interface using video icons for visual database systems,” *Proc. IEEE workshop on visual languages*, pp. 68-37, Rome, IEEE Computer Society Press, 1989.

75. Y. Tonomura, *et al*, “VideoMAP and VideoSpaceIcon: Tools for Anatomizing Video Content,” *Proc. InterChi'93*, ACM, 1994, pp.131--136.

76. V. Vilaplana, F. Marqués, P. Salembier, L. Garrido, "Region-based segmentation and tracking of human faces", Proceedings of the Ninth European Signal Processing Conference EUSIPCO-98, vol I, pp 311-315, Rodas, 1998.

77. V. Vinod and H. Murase, “Video Shot Analysis Using Efficient Multiple Object Tracking,” Proceedings of the IEEE Conference on Multimedia Computing and Systems, pp. 501-508, June 1997.

78. URL: Virage Inc. (http://www.virage.com)

79. Gang Wei and Ishwar K. Sethi, “Face Detection for Image Annotation,” Workshop on pattern Recognition in Practice, The Netherlands, 1999.

80. E. Wold and T. Blum “Content-based Classification, search, and retrieval of audio, IEEE Multimedia, pp. 27-36, Fall 1996.

81. G. Yang, and T. S. *Huang, Human Face Detection in a Complex Background*, Pattern Recognition, Vol. 27, No. 1, pp. 53-63, 1994.

82. B-L, Yeo, and B. Liu, “A Unified Approach to Temporal Segmentation of Motion JPEG and MPEG Compressed Video,” *Proc. Multimedia Tools and Applications*, Vol. 1, No. 1, pp. 81-88, 1995.

83. M.M. Yeung and B. Yeo, “Video Content Characterization and Compaction for Digital Library Applications, in Storage and Retrieval for Image and Video Databases V”, *Proc. SPIE 3022*, pp. 45– 58, 1997.

84. M.M. Yeung, *et al*, Video Browsing using Clustering and Scene Transitions on Compressed Sequences, *Proc. of IS\&T/SPIE'95 Multimedia Computing and Networking*, San Jose, Vol.2417, February 1995, pp.399-413.

85. K.C. Yow and R. Cipolla, "Feature-based human face detection", *Image and Vision Computing*, Vol. 15, pp. 713-735, 1997.

86. H. Yu, W. Wolf, "A visual search system for video and image databases", in Proceedings, IEEE multimedia computing and systems 97, June 1997

87. H. Yu, W. Wolf, " A hierarchical multi-resolution video transition detection scheme", to appear, Journal of Computer Vision and Image Understanding, Academic Press.

88. R. Zabih, J. Miller, and K. Mai, “ A Feature-Based Algorithm for Detecting and Classifying Scene Breaks,” *Proc. ACM Multimedia*, pp. 189-200, San Francisco, 1995.

89. H.J. Zhang A. Kankanhalli and S. W. Smoliar, “Automatic Partitioning of Full-Motion Video,” *Multimedia Systems*, ACM-Springer, Vol.1, No.1, 1993, pp.10-28.

90. H.J. Zhang and S. W. Smoliar, “Developing Power Tools for Video Indexing and Retrieval,” *Proc. SPIE'94 Storage and Retrieval for Video Databases*, San Jose, CA, USA, February, 1994.

91. H.J. Zhang, Y.L. Chien, and S.W. Smoliar, “Video Parsing and Browsing Using Compressed Data,” *Proc. Multimedia Tools and Applications*, Vol. 1, No. 1, pp. 89-111, 1995

92. W. Zhou, Y. Shen, A. Vellaikal, C.-C. J. Kuo, “Online Scene Change Detection of Multicast (Mbone) Video,” Proc. of SPIE Conference on Multimedia Storage and Archiving Systems III, Vol. 3527, November 1998, pp. 271-282.