Video Content Modeling: An Overview

Faisal I. Bashir, Ashfaq A. Khokhar
{fbashir,ashfaq}@ece.uic.edu
University of Illinois at Chicago
Chicago, IL 60607

Abstract

This paper provides an overview of different video content modeling techniques employed in existing content-based video indexing and retrieval (CBVIR) systems. Based on the modeling requirements of a hypothetical (somewhat ideal) CBVIR system, we analyze and categorize existing modeling approaches. Starting with a review of techniques to model raw video data, we study approaches used to describe physical objects, and conclude with a review on high-level semantic modeling of data with focus on the multimodal analysis. Based on the current status of research in CBVIR systems, we identify the growth potential, future directions, and open research issues. Finally, a hypothetical CBVIR system is outlined in the concluding remarks, which exploits object-based representation of MPEG-4 compressed bitstream and uses multimodal features based on high-level description of video.

Index Terms: Video Content Modeling, Content-Based Access of Video, Video Retrieval, Temporal Segmentation, MPEG-4, Multimedia, Semantic content modeling.

I: Introduction

Digital image and video are rapidly evolving as the modus operandi for information creation, exchange and storage in our modern era. Primarily, this is attributed to advances in the three major technologies that determine its growth: VLSI technology that is unleashing greater processing power; broad-band networks (ISDN, ATM etc) that are providing almost unlimited bandwidth for most practical purposes, and image/video compression standards (JPEG, H.263, MPEG etc) that are enabling efficient storage and communication. The combination of these three advances is spurring the creation and handling of increasing high-volume image/video data, along with its efficient compression and transmission over higher-bandwidth networks. This current trend towards the removal of every conceivable bottleneck in using multimedia and its impact on the whole spectrum of users from advanced research organizations to home users has led to an explosive growth of visual information available in the form of digital libraries or online multimedia archives. According to a press release by Google Inc. in December 2001, the search engine offers access to over 3 billion web documents and its Image search comprises more that 330 million images. AltaVista has been serving around 25 million search queries per day in more than 25 languages, with its multimedia search featuring over 45 million images, videos and audios. Although search engines such as above mainly use keyword based textual search techniques on text annotations for visual information [15], this manner of indexing and retrieving visual information is highly non-scalable as well as resource-hungry in terms of manual work and extra storage requirements [3]. A consequence of the growing consumer demand for visual information is that sophisticated technology is needed for representing, modeling, indexing, and retrieving multimedia data. In particular, we need robust techniques to develop semantically rich models to represent visual data, computationally efficient methods to index/retrieve and compress visual information, new scalable browsing algorithms allowing access to very large databases of images and videos, and semantic visual interfaces integrating the above components into a single concept of CBVIR systems.

Generally, CBVIR systems bear on modeling and extracting effective features describing visual media being indexed; the high-dimensional feature vector thus formed is stored in a database. Once a query is posed, the whole database of feature vectors for all visual media (or a subset cluster of the database) is searched to generate a similarity rank (not the exact match) with the given input query. The underlying features can be low-level (primitive) or high-level (semantic), but the extraction and matching process are predominantly automatic [1]. The modeling of the visual media plays perhaps the most dominant role in system's robustness and we look into this matter in section II in more detail. Given this schema, content-based visual information indexing and retrieval proves itself to be a multidisciplinary field, lying at the frontiers of many already mature engineering and computer science faculties.

This document primarily focuses on video content modeling techniques dealing with uncompressed and compressed domain representation of data. We avoid the description of domain specific content modeling such as those tailored for sports, news, or medical videos. The remainder of this paper is organized as follows: Section II presents the general CBVIR paradigm highlighting the important criteria for modeling video data along with a classification of these modeling techniques based on efforts stemming out from different research communities. Section III reviews modeling techniques for uncompressed video data. Video segmentation based on shot boundary detection and shot similarity is also described in this section. Given the success of its predecessors and growing availability of object-based compressed bitstreams, Section IV is devoted to the review of MPEG-4 based systems. Section V reviews the approaches that deal with low to high level semantic gap or high-level semantic representation of video data. Section VI identifies some open issues and future research directions in the field and Section VII ends up with some concluding remarks featuring a hypothetical content-based video indexing and retrieval system.

II: Content-Based Visual Information Retrieval Paradigm

Image retrieval has been an active research area since 1970's, with thrust from two major research communities, Database Management and Computer Vision. As recently as in 1990's, the major drawbacks of searching visual media based on textual annotations [10] were recognized to be unavoidable and content-based image retrieval was proposed [2] in order to overcome these drawbacks. In content-based retrieval, manual annotation of visual media is avoided and indexing is instead performed on the basis of content within the media itself modeled by features such as color, texture, spatial relationship of objects, motion etc. There have been extensive studies on the need and effectiveness of automatic retrieval by content [9]. Several successful research/commercial content-based image search engines such as QBIC[12], PhotoBook[13], VisualSeek[14] have been built.

A: Content Modeling and Feature Selection Criteria

Content modeling forms the basis of CBVIR systems. As pointed out in [63], the central goal of CBVIR research is first to understand how the structure of surrounding world is evidenced by regularities among the pixels of image array (or frames of video data), and then to figure out how these regularities are mapped onto predicates that constitute the primitive elements of cognition. To achieve this goal, one should identify desirable attributes of features that summarize the content of visual information captured from the world-view. The features selected should be able to capture and preserve enough information about intermediate or high-level description about visual media; e.g., description of some event, presence of some known individual, and representation of feelings and emotions. Feature space should also be able to distinguish between images and videos containing substantially different content and at the same time, should have no trouble in retrieving images/videos with similar content. To wrap it up efficiently, is another major concern in feature selection, and both computation time as well as storage overhead function as constraints to be fine-tuned here.

Different features that have been explored in literature for representation of visual media range from low-level to highly derived ones. There has been a wealth of survey of literature on feature selection for image indexing and retrieval systems [2, 4 – 7]. The most dominant features used to describe visual content have been color, shape, texture and spatial relationship of objects. There have been efforts towards processing of compressed domain visual data as first hand data, and an excellent review on this topic can be found in [3]. Features used to represent video data have conventionally been the same ones as for images, extracted from keyframes of the video sequence with an additional motion feature to capture temporal aspects of video data. In [51], Nephade et al. first segment the video spatio-temporally obtaining regions in each shot. Each region is then processed for feature extraction. They use linearized HSV histogram having 12 bins per channel as color feature, Gray Level Co-occurrence matrices at four orientations as texture feature, moment invariants as shape features, and inter-frame affine motion parameters of the region as motion feature. On similar lines, Shih-Fu Chang et al. [25] use quantized CIE-LUV space as color feature, three Tamura texture measures (coarseness, contrast and orientation) as texture feature, as well as shape components and motion vectors. All these features are extracted from objects detected and tracked in video sequence after spatio-temporal segmentation.

B: Classification of Content Modeling Schemes

Studies into human visual perception (or any sensual perception, in general) indicate that there exists a gradient of sophistication in human perception, ranging from seemingly primitive inferences of shapes, textures, colors, etc. to the sophisticated notions of structures such as chairs, trees, affordances, and to cognitive processes such as recognition of emotions and feelings [9]. Though there exist two different schools of thought as how to model these perceptions, but the existence of primitive and semantic clues in visual information is well understood. Marr [64] regards perception as a smooth series of progressively more sophisticated inferences while Pentland [63] agues that there are separate specialized mechanisms for primitive and sophisticated inferences. Given the vastly multidisciplinary nature of the techniques for modeling, indexing and retrieval of visual data, efforts from many different communities have come together in the advancement of CBVIR systems. Depending on the background of the research teams different levels of abstractions have been assumed to model the data. As shown in Figure 1, we classify these abstractions into three categories based on the gradient model of human visual perception. In this figure we also capture the mutual interaction of some of the mature engineering, computer science and cognitive sciences faculties.

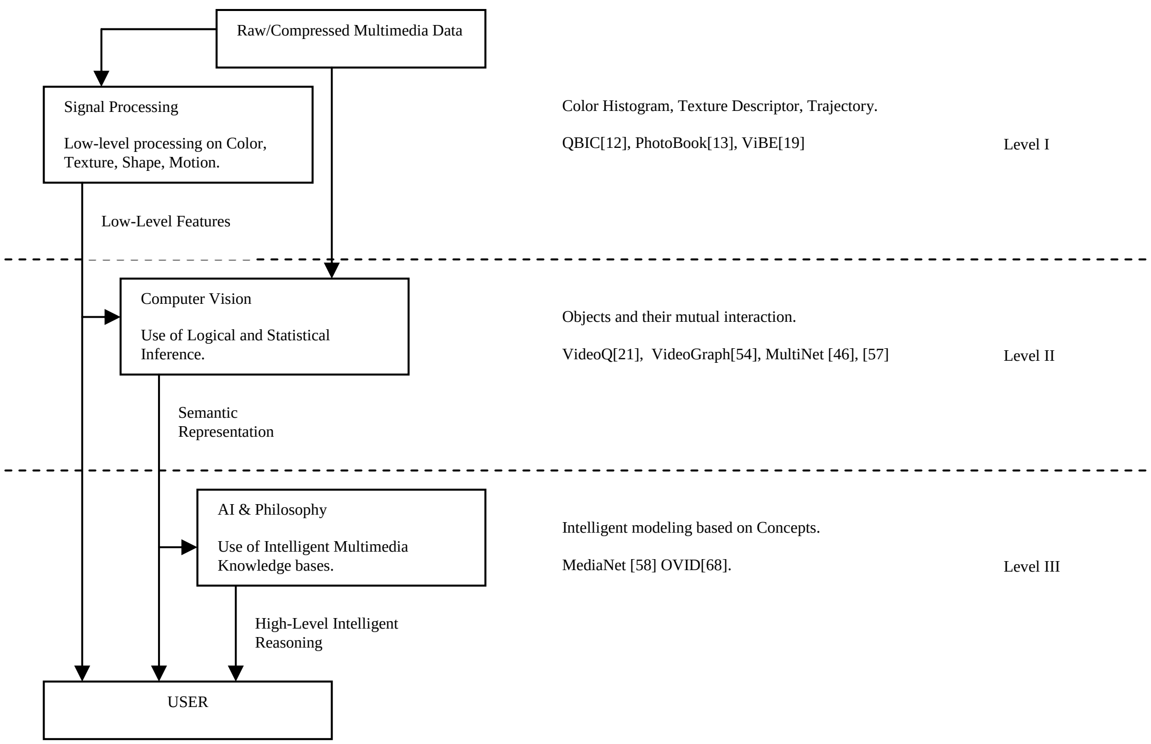

Figure 1: Classification of Content Modeling Techniques. Level I: Modeling of Raw Video Data. Level II: Representation of derived or logical features. Level III: Semantic level abstractions.

Level I represents systems that model raw video data using features such as color histogram, shape and texture descriptors, or trajectory of objects. It can serve the queries like “shots of object with dominant red color and moving from left corner to right”. CBVIR systems based on these models Systems at this operate directly on the data employing techniques mostly from signal processing domain. Level II consists of derived or logical features involving some degree of statistical and logical inference about the identity of objects depicted by visual media. An example query at this level can be “shots of Sears Tower”. At this level, systems normally operate on low-level feature representation though they can also use video data directly to reach at a semantic representation. Level III deals with semantic abstractions involving a significant amount of high-level reasoning about the meaning and purpose of the objects or scenes depicted. An example of a query at this level can be "shots depicting human suffering or sorrow". As the figure indicates at level III, AI community has been in the leading role and systems at this level can take semantic representation of level II as input and further process it to generate results. Shih-Fu Chang et al. [4] have provided a broad taxonomy of CBVIR systems based on functionality and mode of operation.

III: Video Data Modeling and Representation

Most existing video database systems start off with temporal segmentation of video into shots and scenes, a hierarchical video stream model first proposed by Swanberg et al. [64]. Once a temporal segmentation of video into shots is done, we have a smaller set of storyboards representing the whole video sequence. Most existing systems then proceed by selecting representative key-frames, one per shot. This concludes the process of video summarization and Table of Contents (TOC) generation. One important piece of information that is lost during temporal segmentation is the recognition and tracking of objects beyond shots in overlapping intervals; this issue as we will see in section V can be addressed by using appropriate video data model at the semantic level of video description. An important issue that most of the systems overlook is compact representation of shots so as to keep as much of information intact as possible. Similarity measure between two shots, represented by their corresponding multi-modal feature vectors, is an important building block of existing CBVIR systems. Query paradigm also plays an important role in the design of CBVIR systems. Video data, due to its rich information content, can be queried in diversely different ways; Query by Keyword (QBK) as in text-based search engines, Query by Example (QBE) being the one borrowed from content-based image retrieval systems, and Query by Sketch (QBS) being more suitable for video data search having the capability of capturing motion information from video sequences. A broader classification of different types of queries can be found in [43].

A: Temporal Segmentation

Video data can be viewed as a hierarchy, where at the lowest level, video data is made up of frames; a collection of frames that result from single camera operation, focusing on one object or depicting one event is called a *Shot*; A *Scene* is defined by a complete unit of narration which consists of a series of shots or a single shot that takes place in a single location and that deals with a single action [66]. Transitions or boundaries between shots can be abrupt (Cut) or they can be gradual (Fade, Dissolve, Wipe). Traditional temporal segmentation techniques have focused on cut detection, but there has been activity on gradual shot boundary detection as well. Most of the existing techniques reported in literature detect shot boundary by extracting some form of feature for each frame in the video sequence, then evaluating a similarity measure on features extracted from successive pairs of frames in the video sequence, and finally declaring shot boundary if the difference exceeds a fixed global threshold. One such approach is [23] in which two difference metrics, Histogram Distance Metric (HDM) and Spatial Distance Metric (SDM) are computed for every frame pair. HDM is defined in terms of 3-channel linearized histograms computed for successive frame pair f_i and f_{i+1} as follows:

$$D_h(f_i, f_{i+1}) = \frac{1}{M \times N} \sum_{j=1}^{256x3} \left| H_i(j) - H_{i+1}(j) \right|,$$

where H_i represents the histogram of frame f_i and MxN is the dimension of each frame. For each histogram, 256 uniform quantization levels for each channel are considered. SDM is defined in terms of the difference in intensity levels between successive frames at each pixel location. Let $I_{i,j}(f_k)$ denote the intensity of a pixel at location (i,j) in the frame f_k, then the spatial distance operator is defined as:

$$d_{i,j}(f_k, f_{k+1}) = \begin{cases} 1 \cdots\cdots if \left| I_{i,j}(f_k) - I_{i,j}(f_{k+1}) \right| > 0 \\ 0 \cdots\cdots else \end{cases}$$

SDM is then computed as follows:

$$D_s(f_k, f_{k+1}) = \frac{1}{M \times N} \sum_{i=1}^{M} \sum_{j=1}^{N} d_{i,j}(f_k, f_{k+1})$$

These two distances are then treated as a 2-D feature vector and an unsupervised K-Means clustering algorithm is used to group shot boundaries into one cluster. For a review of major conventional shot boundary detection techniques, refer to [26] which also provides a comparison between five different techniques based on pixel difference from raw data, DCT coefficients difference and motion compensated difference. Due to the huge amount of data to be processed in the case of full-frame pixel difference based methods and also their susceptibility to intensity differences caused by motion, illumination changes and noise, many novel techniques, outside the scope of above review in [26], have been proposed in compressed as well as uncompressed domain which we explore briefly in the next few paragraphs.

In [30], a frequency domain correlation approach is proposed. For each 32x32 block in one frame, the best matching block in corresponding neighborhood in the next frame is sought by calculating the normalized cross correlation in frequency domain as:

$$\rho(\varepsilon) = \frac{F^{-1}\left\{\hat{x}_1(\varpi)\times\overset{*}{\hat{x}}_2(\varpi)\right\}}{\sqrt{\int\left|\hat{x}_1(\varpi)\right|^2 d\varpi\cdot\int\left|\hat{x}_2(\varpi)\right|^2 d\varpi}},$$

where ε and ϖ are the spatial and frequency coordinate vectors respectively, $\hat{x}_i(\varpi)$ denotes the Fourier transform of frame $x_i(\varepsilon)$, F^{-1} denotes the inverse Fourier operation and * is the complex conjugate. Next the mean and standard deviation of the correlation peaks for each block in the whole image are calculated and the peaks one standard deviation away from the mean are discarded, making the technique more robust when majority of content remains the same from frame to frame. An average mean is then computed from this pruned data. This average match measure is then compared to average match of previous pair and a shot boundary is declared if there is a significant decrease in this similarity match feature.

A novel approach proposed in [29] argues that at the shot boundary, the contents of new shot differ from contents of the whole previous shot instead of just the previous frame. Their recursive natured Principal Component Analysis- based generic approach, which can be built upon any feature extracted from frames in a shot, generates a model of the shot trained from features in previous frames. Features from current frame are extracted and a shot boundary is declared if the features from current frame do not fit well in the existing model by projecting the current feature onto existing eigenspace.

In an effort to cut back on huge amount of data available for processing and emphasizing on the fact that in video shots while objects may appear or disappear, background stays much the same and follows camera motion within one shot, Oh et al [27] have proposed a background tracking (BGT) approach. A strip along top, left and right border of the frame covering around 20% of frame area is taken as fixed background area (FBA) out of which a signature, a 1-D vector called transformed background area (TBA) formed from gaussian pyramid of FBA, is computed. Background tracking is achieved by 1-D correlation kind of matching between two TBA’s from successive frames. This approach has been reported to detect and classify both abrupt and gradual scene changes.

Observing the fact that single features can’t be used accurately in a wide variety of situations, Delp et al [23] have proposed to construct a high-dimensional feature vector, called Generalized Trace (GT), by extracting a set of features from each DC frame. For each frame, GT contains the number of intra coded as well as forward- and backward- predicted macroblocks, histogram intersection of current and previous frames for Y, U and V color components and standard deviation of Y, U and V components for current frame. GT is then used in a binary regression tree to determine the probability that each frame is a shot boundary. These probabilities are then used to determine frames corresponding to the shot boundary.

Hanjalic [31] has put together a nice analysis of the shot boundary detection problem itself, identifying major issues that need to be considered, along with a conceptual solution to the problem in the form of a statistical detector based on minimization of average detection-error probability. The thresholds used in their system are defined at the lower level modules of detector system. The decision making about presence of a shot boundary is then left solely to parameter-free detector, where all the indications coming from different low-level modules are evaluated and combined.

Schonfeld et al. [32] present a scene change detection method for compressed domain videos using stochastic sequential analysis theory. The DC data from each frame of MPEG compressed video is processed using Principal Component Analysis to generate a very low dimensional feature vector Y_k corresponding to each frame. These feature vectors are assumed to form an i.i.d. sequence of multidimensional random vectors having Gaussian distribution. Scene change is then modeled as change in the mean parameter of this distribution. Scene change is declared at frame k when the maximum value of the parameter g_k evaluated over frame interval j to k, as:

$$g_k = \max_{1 \le j \le k} \left\{ \frac{k-j+1}{2} (X_j^k)^2 \right\},$$

exceeds a preset threshold. Here X_j^k is defined as:

$$X_j^k = \left[(\overline{Y}_j^k - \Theta_0)^T \Sigma^{-1} (\overline{Y}_j^k - \Theta_0) \right]^{1/2}.$$

In the expression of X_j^k, $\overline{Y}_j^k$ is the mean of feature vectors Y in the current frame interval j to k, and Θ_0 is the mean of Y in the initial training set frame interval consisting of M frames. This approach, which is free from human fine-tuning, has been reported to perform equally well for both abrupt and gradual scene changes.

B: Content Representation and Similarity Measures

The next logical step after temporal segmentation is compact representation and modeling of contents inside each shot, so as to facilitate the robust matching between any two shots at query time. Most existing systems represent the content by using one representative frame from the shot, called keyframe. Keyframe extraction has been recognized as one important research issue in content based video abstraction. One approach is to use first frame of each shot as a keyframe [16]. Although the approach is simple but each shot gets only one frame for its representation no matter how complex the shot contents be. To decide about the selection of more than one frame per shot, Zhang et al [17] use multiple different criteria like color content change and zoom-in type of effects in shot content. A technique for shot content representation and similarity measure using subshot extraction and representation is presented in [18]. They use two content descriptors, Dominant Color Histogram (DCH) and Spatial Structure Histogram (SSH), to measure content variation and to represent subshots. Delp et al [19] represent a shot using a tree structure called shot tree, formed by clustering frames in a shot. This approach unifies the problem of scene content representation for both browsing and similarity matching where for browsing only the root node of the tree (keyframe) is used, while for similarity matching two or three levels of tree can be used employing the standard tree matching algorithms. The ultimate goal for any CBVIR system is to retrieve video sequences visually most similar to input query, for which a measure of similarity between the contents of two shots is required that takes into account the real perceptual similarity between shots. Since video shots encapsulate spatial, temporal and high-level semantic information, more sources of information taken into account are likely to yield more accurate results. Delp et al [19] use four different distance measures between shots; Shot Tree dissimilarity measures the distance between two shots represented in their shot tree form; Temporal distance captures how far away two shots are based on difference between starting and ending frame numbers; Motion distance takes into account the number of non-intra coded macroblocks in each P or B frame; Pseudo-Semantic distance is the L_1 norm between two vectors one per shot containing the confidence measure (a number in range 0-1) of the shot being a member of all the pseudo-semantic classes. Individual distance measures are finally summed up with proper weighting factors. Shih-Fu Chang et al [21] segment video objects in each video sequence, tracking each object within shots, and finally building a database of features for each object. Color, texture, size, shape and motion are the features that are added up in proper weights to generate the distance between the query object and current target object in the database.

IV: Modeling in the Compressed Domain

The ISO SC29 WG11 "Moving Pictures Experts Group" (MPEG), responsible for "Coding of moving pictures and associated audio", established in 1988 has been working on a new standard MPEG-4 aimed at content-based interactivity. The specifications of intended codec have been approved in the form of verification models, the latest of which – as of this writing - is verification model 18.0 approved in January 2001. MPEG-4 video aims at providing standardized core technologies allowing efficient storage, transmission and manipulation of video data in multimedia environments [33]. The main motivations behind this task are proven success of digital video in three fields, namely digital television, interactive graphics applications (synthetic image content) and interactive multimedia (world wide web, distribution and access to image content) [34]. With reference to existing standards, at least seven new key video coding functionalities have been defined in the major areas of content-based interactivity, compression and universal access in heterogeneous networked environments [40].

A: Motivation and Background

In contrast to current state-of-the-art video coding techniques, in MPEG-4, a scene is viewed as a composition of Video Objects (VO) with intrinsic properties such as shape, motion, and texture. The attempt is to encode the sequence in a way that allows the separate decoding and reconstruction of objects and to allow the manipulation of original scene by simple operations on the bitstream. The bitstream is object-layered and the shape and transparency of each arbitrarily-shaped object – as well as the spatial coordinates and additional parameters describing object scaling, rotation, or related parameters – are described in the bitstream of each object layer. The Video Objects (VOs) correspond to entities in the bitstream that the user can access and manipulate (cut, paste, etc). Instances of Video Object at a given time are called Video Object Planes (VOPs). The VOPs are either known by construction of the video sequence (hybrid sequence based on blue screen composition or synthetic sequences) or are defined by semi-automatic segmentation [42]. The encoder sends together with the VOP, composition information (using composition layer syntax) to indicate where and when each VOP is to be displayed. The MPEG-4 encoder is therefore composed of two parts: *shape encoder* and the conventional *motion and texture encoder*. The principles used in encoding and decoding the binary alpha component are similar to those used for encoding and decoding texture data; same block size (16x16 pixels for both Macroblock and Binary Alpha Block), same modes of compression (Intra-Coded, Inter-Code and Not-Coded) are some of the major similarities in the two encoders [41]. Every VOP can be encoded either as I-VOP utilizing only intramode techniques, P-VOP using temporal prediction from past VOPs or B-VOP using bi-directional temporal prediction from past and future VOPs. A specific number of groups of consecutive macroblocks are put in as a Video Packet, each separated from the other by Resynchronization Markers – the purpose of this packetization is to support error detection, localization and bitstream resynchronization while decoding. At the decoder side, the user may be allowed to change the composition of the scene displayed by interacting with the composition information. All these new functionalities in MPEG-4 add up to providing a framework which facilitates feature extraction from video sequences using minimal decoding of bitstream and make MPEG-4 compressed bitstreams good candidates for CBVIR systems. Although MPEG-4 offers an edge in coding over existing video coding standards, the content based indexing and retrieval techniques based on MPEG-4 bitstream are still in its infancy. Below we review some of the recent research efforts based on object-based video coding principle of MPEG-4.

B: Object Based Video Summarization

As pointed out earlier, in conventional video abstraction and summarization schemes, keyframes are used to represent a shot. In object-based video coding paradigm, key VOPs are sought to facilitate object-based queries, such as querying a video object with a given shape and color, and moving in a given direction at a given speed. Berna et al [35] use shape of the VOP as a cue for selecting key VOP. At the encoder side, the shape data associated with each of 16x16 macroblocks is transmitted in the bitstream, along with texture information that corresponds to the same area. Shape-based features offer the edge over texture because shape information is coded losslessly and also that MPEG-4 bitstream structure is designed such that it is not possible to decode texture information without having to decode the shape information first. Hence shape-based feature extraction, at least in the domain of MPEG-4 encoding, is more reliable and computationally more efficient. The first VOP is selected as the key VOP, and a new key VOP is declared whenever a significant change occurs in the shape of video object. Instead of operating on each pixel of binary alpha planes of the VOPs, the shape is approximated by noting the mode of each 16x16 binary alpha block which are encoded as Transparent, Opaque or Intra depending on whether the block is outside, inside or at the boundary of the shape [36]. The process is made spatial-shift invariant by aligning the mass centers of current and candidate VOP before calculating distance between them.

C: Features from MPEG-4 Bitstream

Taking advantage of object-based representation built in MPEG-4, shape as well as texture has been used on arbitrarily shaped video objects. The work in [36] reports computing some of the commonly used shape features like Compactness, Eccentricity and Fourier Descriptors efficiently from MPEG-4 compressed bitstream. Along with these, two new features namely Pruned Intra Density ID_p and Inside Intra Density ID_I are also proposed which depend upon the statistics of intra-coded blocks in the VOP shape. ID_p is defined as:

$$ID_p = \frac{I_p}{\sqrt{w^2 + h^2}},$$

where I_p is the number of intra coded shape blocks that need to be pruned from VOP shape to obtain a closed contour, and w and h are the width and height of the bounding of box of the VOP respectively. This feature takes a high value if the boundary of the shape has a lot of irregularities and branches. ID_I is defined as:

$$ID_I = \frac{I_i}{\sqrt{w^2 + h^2}},$$

where I_i is the number of intra coded shape blocks that do not have any transparent coded neighbor blocks based on four-point connectivity. A large value of this feature is an indication of the presence of holes in the VOP shape. Based on these features, a feature vector is formed comprising of features from each key VOP. Weighted distance between the query and the target VOP is evaluated at each time instance to give a distance measure between the two VOPs; finally, the distance between two VOs is computed by summing up minimum distance between two VOPs at each time instance. One step further is taken in [38] where shape deformation features are also taken into account. They describe the changes in shape content based on variances of various still shape features, and the final distance between query and target VOs are computed by weighted sum of distance between the still shaped and shape deformation based distances.

In [39], Color is taken as the representative feature, which is approximated from DC coefficients of Y, Cb and Cr components of intra coded VOPs. Color space to represent color is taken as MTM (Mathematical Transformation to Munsell), which is a perceptually uniform color space. They use a bin-size of 128 with 8 bins for Hue and 4 each for Value and Chroma. Color histogram of each VOP is calculated by using the color components from the Opaque or Intra blocks only. The VO histogram is obtained by taking the average of histogram values from all the frames of individual VOPs.

Work in [37] uses motion as a cue for partitioning the VOs into temporal segments with uniform activity level. This motion information can be used as a feature in video objects retrieval as well as it can assist in shot boundary detection in a video sequence. The algorithm is based on changes in texture- and shape- coding modes of inter-coded VOPs in MPEG-4 bitstream. Since texture/color of the VOs normally stays constant throughout object's lifespan, use of texture- as well as shape- coding modes to detect local motion activity has been implemented.

V: Semantic Modeling

As pointed out earlier, higher level indexing and retrieval of visual information, as at level II or level III in Figure 1, asks for semantic analysis that is beyond the scope of many of the techniques described so far. One important consideration that many existing content modeling schemes overlook is the importance of multi-modal nature of video data comprising of sequence of images along with associated audio and in many cases, textual captions. Fusing data from multiple different modalities improves the overall performance of system [11]. Many of the content modeling schemes based on low-level features work on Query By Example (QBE) paradigm; at times this constraint is unreasonable when an example video clip of what the person is originally looking for is not at hand. Query By Keyword (QBK) offers an alternate to QBE, in the high-level semantic representation [50]. There has been a drive towards incorporating intelligence into CBVIR systems and we will look into some intelligence based ideas and systems in this section. Modeling video data and designing semantic reasoning based Video Database Management Systems (VDBMSs) facilitate high-level querying and manipulation of video data. A prominent issue associated with this domain is development of formal techniques for semantic modeling of multimedia information. Another problem in this context is the design of powerful indexing, searching and organization methods for multimedia data. In [44], a multimedia database system for content-based retrieval is presented. An object-oriented data model and a query language are used. The database schema is represented through a hierarchy with *is_a* and *part_of* relationship among classes. A class is associated with the domain knowledge to represent a certain concept. Retrieval is done by matching the query and the domain knowledge stored in classes. A three-level motion analysis methodology is proposed in [45]. Starting from the extraction of trajectory of a macro-block in an MPEG video, followed by averaging all trajectories of the macro-blocks of objects, and finally relative position and timing information among objects, a dual hierarchy (spatial + temporal) is established for representing video.

A: Multimodal Probabilistic Frameworks

Multimedia indexing and retrieval presents a challenging task of developing algorithms that fuse information from multiple media to support queries. Content modeling schemes operating in this domain have to bridge the gap between low-level features and high-level semantics. Nephade et al [48] have proposed the concept of Multiject, Multimedia Object, which has a semantic label, associated multi-modal features (including both audio and video features) and a probability of its occurrence in conjunction with other objects in the same domain (shot). Multijects for concepts from three main categories of Objects (e.g., Airplane), Sites (e.g., Indoor) and Events (e.g., Gunshot) have been experimented with. Given the multimodal feature vector $\vec{X}_j$ of the jth frame and assuming uniform priors on presence/absence of any concept in any region, the probability of occurrence of each concept in jth frame is obtained from Bayes' rule as:

$$P(R_{ij} = 1 \mid \vec{X}_j) = \frac{P(\vec{X}_j \mid R_{ij} = 1)}{P(\vec{X}_j \mid R_{ij} = 1) + P(\vec{X}_j \mid R_{ij} = 0)},$$

where R_{ij} is a binary random variable taking value 1 if the concept i is present in frame j. During training phase, the identified concepts are given labels and the corresponding Multiject consists of a label along with its probability of occurrence and multimodal feature vector. Multijects are then integrated at the frame level by defining frame level features $F_i, i \in \{1 \cdots N\}$ (N is the number of concepts the system is being trained for) the same way as for R_{ij}. If M is the number of regions in current frame, then given $\chi = \{\vec{X}_1, \cdots \vec{X}_M\}$ the conditional probability of Multiject i being present in any region in the current frame is:

$$P(F_i = 1 \mid \chi) = \max_{j \in \{1, \cdots, M\}} P(R_{ij} = 1 \mid \vec{X}_j).$$

Observing the fact that semantic concepts in videos do not appear in isolation, but they interact and appear in context, their interaction is modeled explicitly and a network of multijects, called Multinet is proposed [46]. A framework based on multinet takes into account the fact that presence of some multijects in a scene boosts the detection of some other semantically related multijects and reduces the chances for some others. Based on this multinet framework, spatio-temporal constraints can be imposed to enhance detection, support inference and impose a priori information. While the multiject model is built using Bayes decision theory, they propose a factor graph based approach for the computational modeling of multinets [52], [47]. A realization of multinet based on Bayesian nets is explored in [50], [51], [53]. In [49], an algorithm for matching multimodal (audio-visual) patterns based on a dynamic programming approach for efficient video retrieval is proposed.

B: Intelligence Based Systems

The next step towards future CBVIR systems will be marked by the full introduction of intelligence into the systems as they need to be capable of communicating with the user, understanding the audio-visual content at a higher semantic level and reasoning and planning at human level [6]. Intelligence is referred to as the capabilities of the system to build and maintain situational or world models, utilize dynamic knowledge representation, exploit context, and leverage advanced reasoning and learning capabilities. An insight into human intelligence can help better understand users of CBVIR systems and construct more intelligent systems. Ana et al [58] propose an intelligent information system framework MediaNet, which incorporates both perceptual and conceptual representations of knowledge based on multimedia information in a single framework by augmenting the standard knowledge representation frameworks with the capacity to include data from multiple media. It models the real world by concepts, which are real world entities and relationships between those concepts, which can be either semantic (e.g., Is-A-Subtype-Of) or perceptual (e.g., Is-Similar-To). In MediaNet, concepts can be as diverse-natured as living entities (Humans), inanimate objects (Car), events in the real world (Explosion), or certain property (Blue). Media representation of the concepts involves data from heterogeneous sources.

C: Semantic Modeling and Querying of Video Data

Owing to its distinguished characteristics from textual or image data – very rich information content, temporal as well as spatial dimensions, unstructured organization, massive volume and complex and ill-defined relationship among entities – robust video data modeling is an active area of research [43]. The most important issue that shows up in the design of Video Database Management Systems (VDBMSs) is the description of structure of video data in a form appropriate for querying, easy enough for updating, and compact enough to capture the rich information content of the video. The process of designing the high-level abstraction of raw video to facilitate various information retrieval and manipulation operations is the crux of VDBMSs. To this end, current semantic based approaches can be classified into *segmentation-based* and *stratification-based*. The drawback of former approaches is lack of flexibility and incapability of representing semantics residing in overlapped segments. The later models however segment contextual information of video instead of simply partitioning it. SemVideo [54] presents a video model in which semantic contents not having related time information are modeled as ones that do; also, not only the temporal feature of semantic descriptions, but also the temporal relationships among themselves are components of the model. The model encapsulates information about *Videos*, each being represented by a unique identifier; *Semantic Objects*, description of knowledge about video having a number of attribute-value pairs; *Entities*, any of above two; *Relationships*, an association between two entities. Many functions are also defined that help in organizing data and arranging relations between different objects in video. Tran et al [55] propose a graphical model VideoGraph that supports not only the Event Description, but also Inter-Event Description that describes the temporal relationship between two events – a functionality overlooked by most of the existing video data models. They also have provision for exploiting incomplete information by associating the temporal event with a Boolean-like expression. A query language based on their framework is proposed in which query processing involves only simple graph traversal routines.

Khokhar et al [57] introduce a multi-level architecture for video data in which semantics are shared among various levels. An object-oriented paradigm is proposed for management of information at higher levels of abstraction. For each video sequence to be indexed, they first identify objects inside it, their sizes and locations, their relative positions and movements and this information is finally encoded in a spatio-temporal model. Their approach integrates both intra- and inter-clip modeling and uses both bottom-up as well as top-down object-oriented data abstraction concepts. Decleir et al [56] develop a data model that goes one step beyond the existing stratification-based approaches using *Generalized intervals*. Here instead of a time segment to be associated with a description, a set of time segments is associated with a description – an approach that allows handling with a single object all occurrences of an entity in a video document. They also propose a declarative, rule-based, constraint query language that can be used to infer relationships from information represented in the model, and to intentionally specify relationships among objects.

VI: Open Issues in CBVIR Systems

From the above review, it should be clear that there have been advances in this relatively new field in the past few years at low-level as well as at high, semantic level of visual information representation. However, there are still many open research issues that need to be addressed to make the full use out of visual information retrieval systems. In the following, we identify several open problems and provide a brief summary.

High-Dimensional Indexing – Although the dimensionality of feature vectors employed in most systems for representing visual media is normally quite high, of the order of 10^2, but as suggested in [67], the embedded dimension in most of the cases is much lower. There is a need in practical CBVIR systems as to ascertain the intrinsic dimension of the visual information and use of dimensionality reduction techniques to represent the visual information with least-dimensional feature vectors.

Similarity Matching – Retrieval of visual information requires similarity matching using some metric for its evaluation. Most of the systems employ Euclidean distance measure, which may not successfully simulate human perception of visual similarity in certain visual content. Distance metric has to take into account the gradient in visual perception between monochrome- (sharp perception) and color- (blunt perception) information, and between edge- and smooth- information. Several distance metrics, like histogram intersection, correlation, mahalanobis distance, etc have been proposed in literature. The objective fidelity measures given by distance metrics should be consistent with subjective fidelity results produced by human subjects on similar visual media.

Relevance Feedback – Earlier systems in CBVIR used to emphasize on fully automatic realization, but since no promising results have been generated by these approaches, there has been a shifting trend towards more user inclusion in the process and human in the loop has been promoted. Since the similarity rank problem of CBVIR is different from exact matching of computer vision and pattern recognition problems, learning from user intervention and taking feedback from the user promises to be an important aspect of future systems. Feedback from user can either be directly taken as input from the user, or it can come from intelligent software agents capturing user's browsing trends and sending these patterns to server which can adapt its response to user queries accordingly.

Low- To High- Level Semantic Gap – Visual feature based techniques at the low-level of abstraction, mostly from the contribution of signal processing and computer vision communities have been explored in the literature. Current research efforts are more inclined towards high-level description and retrieval of visual content. Most of the techniques at high level of abstraction assume the availability of high-level representation and process that information for indexing. The techniques that bridge this semantic gap between pixels and predicates are a field of growing interest. Intelligent systems are needed that take low-level feature representation of the visual media and provide a model for the high-level object representation of the content.

Performance Evaluation and Standard Test-Bed – Probably the single area in a high need of standardization and vastly neglected is the evaluation criteria of a system based on which the users can judge how well the system is performing and a comparison between the performance of different systems can be made. SNR has been the performance evaluation metric used in data compression, whereas Precision and Recall have been used in text based information retrieval; there has been no standard bench marking system for CBVIR systems. Similarly, a standard test-bed acting as a common frame of reference is still missing. Lena image has been used as one in image compression, and MPEG-video test bitstreams have been provided for video compression systems, but no such universal test data exists for CBVIR systems testing and evaluation.

Multi-Disciplinary and Multi-Modal approach – As argued earlier in section II, successful realization of CBVIR systems calls for a fusion of efforts from different disciplines and a nice representation of visual information requires data from multiple different modalities to be integrated. The integration of Multidimensional Signal Processing, Computer Vision, Database Management, Artificial Intelligence and Information Retrieval can provide techniques and algorithms for future CBVIR systems.

Compact, Scalable Search Systems – Visual information being high dimensional and information rich in nature yields a good number of massive features to be indexed upon. An efficient and scalable storage of these features in the database presents an overwhelming task. A scalable search system ensures that search time does not increase exponentially with the number of visual media entities in the database, while compact storage of features ensures that precious disk space on the server side is conserved. Although there have been efforts in compact representation of visual features [60], [61], [62], [59], [23], they are all system specific and a universal compact and scalable representation of visual features is still far from realization.

VII: Concluding Remarks

Content-based indexing and retrieval of visual information is an emerging research area that has received growing attention in the research community over the past decade. Though modeling and indexing techniques content-based image indexing and retrieval domain have reached reasonable maturity, content-based access of video data did not receive attention of that level. We have observed that content representation through low-level features has been fairly addressed, and there is a growing trend towards bridging the semantic gap. Mono-modal approaches have proven successful to a certain level, and more efforts are being put for fusion of multiple media. As visual databases grow bigger with advancements in visual media creation, compaction and sharing, there is a growing need for storage-efficient, scalable search systems. In light of the observations in previous sections, we outline one possible realization of a CBVIR system here. As depicted in Figure 2, the system can operate on raw video data but the details of audio-visual feature extraction has been skipped because it is not the main contribution of this paper and can be looked up in abundant references in last sections. The other part of indexing system operates specifically on MPEG-4 compressed bitstream taking advantage of its object-based representation. Indexing process is done all off-line at the time when new videos are added into the database. From video bitstream, each Video Object (VO) is processed separately to extract visual features summarizing the object's shape, color and motion throughout its lifespan. In visual feature extraction, emphasis is put on minimal decoding of the bitstream and approximate representation of shape, trajectory and color based only on the key VOPs for each Video Object. Similarly, audio bitstream is processed to extract audio features, and the two sets of features are fused together for a single representation of the video clip. After the feature-based representation of video clip is done, a knowledge base is consulted which generates appropriate keywords given the audio-visual features. The multimedia features are then integrated with this metadata and this multidimensional feature vector is indexed into the database.

At query time, the user has option of submitting query in any of the three paradigms as mentioned in section V. For QBE, the user submits an example video clip to the system from which audio-visual feature are extracted for matching in the indexed database. In case of QBS, the user is provided with an interface to sketch the objects in video clip and/or their motion trajectories. For QBK search systems, the keywords entered by user are sent to a knowledge base, which interprets the meaning of each keyword and figures out the approximate shape, color and motion of objects the user is interested in. This information is then mapped into audio-visual features. Once the audio-visual features have been compiled, the knowledge base is consulted which generates the metadata corresponding to these features. The query processor then takes

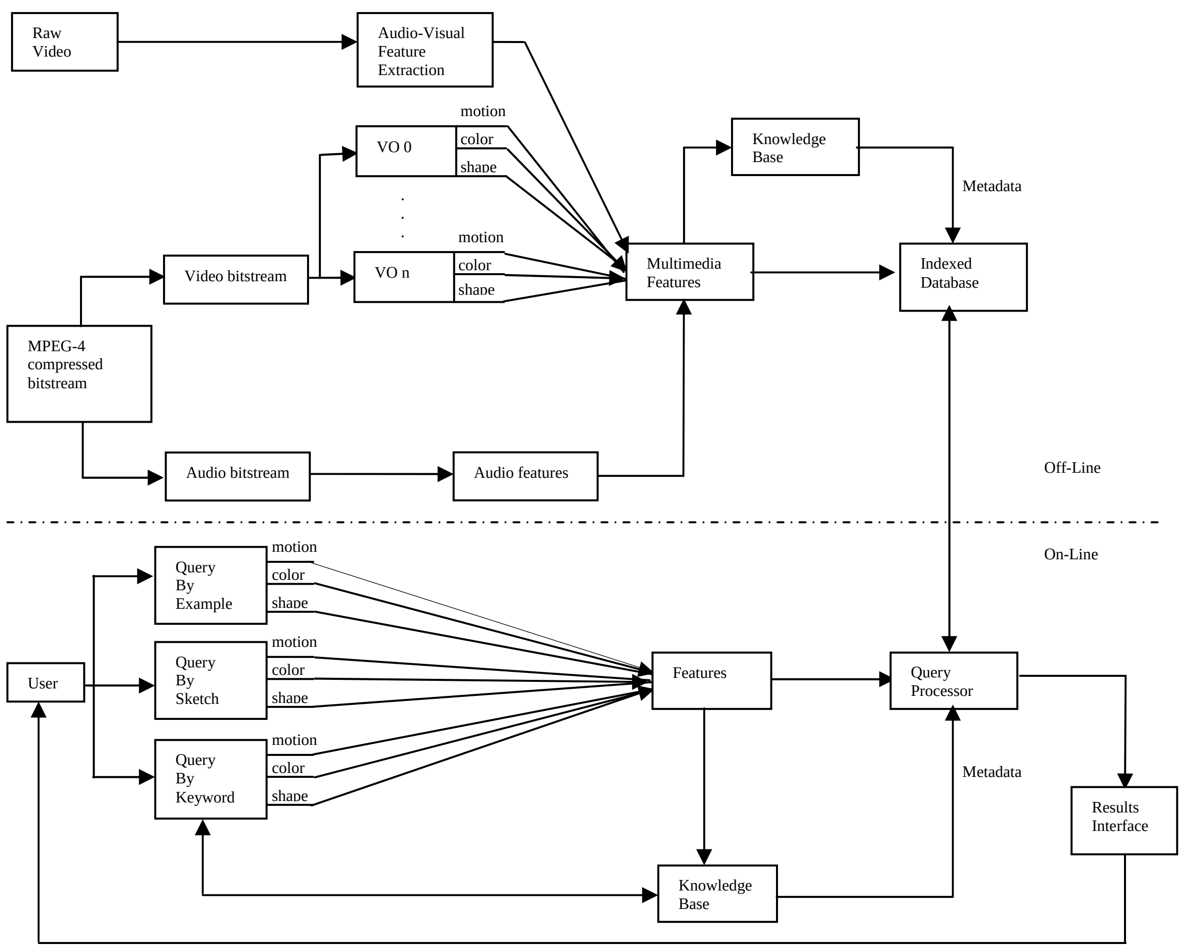

Figure 2: A hypothetical Video indexing and retrieval system operating on raw video or MPEG-4 compressed bitstream using multimedia features for indexing and retrieval.

up these features as well as the metadata and searches for similar feature vectors in indexed database with appropriate weights being assigned to different features, producing ranked results corresponding to user query, which are finally displayed to the user.

The main hallmarks of this system are that it outlines a system capable of exploiting the object-based representation incorporated in MPEG-4 compressed bitstream; integrates research efforts from many different disciplines and uses multimodal features for representation of video clip. The knowledge base module depicts semantic-level representation from multimedia features. The authors are of the view that a system with semantic representation capabilities, operating on multimodal feature representation of the video content is bound to generate promising results.

References

CBVIR – General Overview

[1] Eakins, J P and Graham, M E. “Content-based image retrieval: Report to JISC Technology Applications Programme, January 1999”.

[2] Yong Rui, Thomas S. Huang, and Shih-Fu Chang, “Image Retrieval: Current Techniques, Promising Directions and Open Issues”, Journal of Visual Communication and Image Representation, Vol. 10, 39-62, March, 1999.

[3] M. K. Mandal , F. Idris, and S. Panchanathan, “A Critical Evaluation of Image and Video Indexing Techniques in the Compressed Domain”, Image and Vision Computing Journal-special issue on Content Based Image Indexing, Vol. 17, Issue 7, pp. 513-529, May 1999.

[4] S-F.Chang, J. R. Smith, M. Beigi and A. B. Benitez, “Visual Information Retrieval from Large Distributed On-line Repositories”, Communications of the ACM, Vol. 40, No. 12, pp. 63-71, Dec 1997.

[5] Michael J. Swain, “Searching for Multimedia on the World Wide Web”, 99-1, (March, 1999), Technical Report Cambridge Research Lab.

[6] A. B. Benitez and J. R. Smith, “New Frontiers for Intelligent Content-Based Retrieval”, Proceedings of the SPIE 2001 Conference on Storage and Retrieval for Media Databases (IS&T/SPIE-2001), Vol. 4315, San Jose, CA, Jan 24-26, 2001.

[7] Colin C. Venters Dr. Matthew Cooper, “A Review of Content-Based Image Retrieval Systems”, University of Manchester, July 2000, Technical Report of JISC Technology Application Forum.

[8] Cees G.M. Snoek and Marcel Worring, “Multimodal Video Indexing: A Review of the State-of-the-art”, ISIS technical report series, Vol. 2001-20, December 2001.

[9] Eakins, J P. “Automatic image content retrieval - are we getting anywhere?”, In Proceedings of Third International Conference on Electronic Library and Visual Information Research, De Montfort University, Milton Keynes, May 1996, p123-135.

[10] Abby A. Goodrum, “Image Information Retrieval: An Overview of Current Research”, Informing Science, Vol 3 No 2, 2000.

[11] Yao Wang, Zhu Liu, Jin-Cheng Huang, “Multimedia Content Analysis using both Audio and Visual Clues”, IEEE Signal Processing Magazine, November 2000.

[12] Christos Faloutsos, Ron Barber, Myron Flickner, Jim Hafner, Wayne Niblack, Dragutin Petkovic and Will Equitz, “Efficient and Effective Querying by Image Content”, Journal of Intelligent Information Systems, 3, 3/4, July 1994, pp. 231-262.

[13] A. P. Pentland, R. W. Picard, and S. Sclaroff, “Photobook: Content-based Manipulation of Image Databases”, Int. Journal of Computer Vision, Vol. 18, No. 3, pp. 233--254, 1996.

[14] J. R. Smith and S.-F. Chang, “VisualSEEk: a Fully Automated Content-Based Image Query System”, Proceedings, ACM Multimedia '96 Conference, Boston, MA, November 1996.

[15] S. W. Smoliar and H. Zhang, “Content-Based Video Indexing and Retrieval", IEEE Multimedia, Vol. 1, No. 2, Summer 1994, pp. 62-72.

Video indexing and Retrieval Systems

[16] A. Nagasaka and Y. Tanaka, “Automatic video indexing and full-video search for object appearances”, in Visual Database Systems II, 1992.

[17] H.Zhang, J.Wu, D.Zhong, and S.W.Smoliar, “An integrated system for content-based video retrieval and browsing”, Pattern Recognition, Vol. 30, no.4, pp. 643-658, 1997.

[18] T.Lin, H.J.Zhang, and Q.-Y.Shi,“Video Content Representation for Shot Retrieval and Scene Extraction”, International Journal of Image & Graphics, Vol. 1, No. 3, July 2001.

[19] J-Y Chen, C. Taskiran, A. Albiol, E. J. Delp and C. A. Bouman, “ViBE: A Compressed Video Database Structured for Active Browsing and Search”, submitted to IEEE Transactions on Multimedia 2001.

[20] C. Taskiran, C.A. Bouman, and E. J. Delp, “The ViBE Video Database System: An Update and Further Studies”, Proceedings of the SPIE/IS&T Conference on Storage and Retrieval for Media Databases 2000, January 29-28, 2000, San Jose, California, pp. 199-207.

[21] Shih-Fu Chang, William Chen, Horace J, Meng, Hari Sundaram, Di Zhong, “A Fully Automated Content-Based Video Search Engine Supporting Spatiotemporal Queries”, IEEE Transactions on Circuits and Systems for Video Technology, Vol. 8, No. 5, September 1998.

[22] Di Zhong, Shih-Fu Chang, “Spatio-Temporal Video Search using the Object Based Video Representation”, IEEE International Conference on Image Processing, Oct. 26-29, 1997, Santa Barbara, CA.

[23] M. R. Naphade, R. Mehrotra, A. M. Fermant, J. Warnick, T. S. Huang, A. M. Tekalp, “A High Performance Shot Boundary Detection Algorithm using multiple cues” , Proc. I.E.E.E. International Conference on Image Processing, Volume 2, pages 884-887, Oct 1998, Chicago, IL.

[24] H. J. Zhang, J. Y. A. Wang, and Y. Altunbasak, “Content-based video compression and browsing: a unified solution”, Proc. IEEE Int. Conf. Image Proc., vol. 1, pp. 13-16, Santa Barbara, CA, 26-29 Oct. 1997.

[25] V. Kobla, and D. Doermann, “Indexing and Retrieval of MPEG Compressed Video”, Journal of Electronic Imaging, Vol. 7(2), pp. 294-307, April, 1998.

[26] J.S. Borecsky, L.A. Rowe, “Comparison of video shot boundary detection techniques”, In Proceedings of SPIE, vol. 26670, pages 170-179, 1996.

[27] JungHwan Oh, Kien A. Hua, Ning Liang, “A Content-based Scene Change Detection and Classification Technique using Background Tracking”, Proc. of IS&T/SPIE conference on Multimedia Computing and Networking 2000. pp. 254-265 Jan. 24 - 28, 2000, San Jose, CA.

[28] JungHwan Oh and Kien A. Hua, “An Efficient Technique for Summarizing Videos using Visual Contents”, Proc. IEEE International Conference on Multimedia and Expo. July 30 - August 2, 2000. pp. 1167-1170. Hilton New York & Towers, New York, NY, USA.

[29] X. M. Liu and T. Chen, “Shot Boundary Detection Using Temporal Statistics Modeling”, ICASSP 2002. , Orlando, FL, U.S.A., May 2002.

[30] S V Porter, M Mirmehdi, B T Thomas, “Video Cut Detection using Frequency Domain Correlation”, In Proceedings of the 15th International Conference on Pattern Recognition, pages 413--416. IEEE Computer Society, September 2000.

[31] Alan Hanjalic, “Shot-Boundary Detection: Unraveled and Resolved?”, IEEE Transactions on Circuits and Systems for Video Technology, Vol. 12, No. 2, February 2002.

[32] D. Lelescu and D. Schonfeld, “Real-time scene change detection on compressed multimedia bitstream based on statistical sequential analysis,” Proceedings of the IEEE International Conference on Multimedia and Expo, pp. 1141-1144, New York, New York, 2000.

MPEG-4 Based Video Indexing and Retrieval

[33] “Coding of Moving Pictures and Audio”, ISO/IEC JTC1/SC29/WG11 N3908 (January 2001/Pisa)

[34] Mohammad Ghanbari, “Video Coding: an introduction to standard codecs”, The Institution of Electrical Engineers, London, United Kingdom, 1999.

[35] B. Erol and F. Kossentini “Automatic Key Video Object Plane Selection Using the Shape Information in the MPEG-4 Compressed Domain”, IEEE Transactions on Multimedia, vol. 2, no 2, pp.129-138, June 2000.

[36] B. Erol and F. Kossentini, “Retrieval of video objects using compressed domain shape features”, Proceedings of IEEE ICECS Conference, December 2000.

[37] B. Erol and F. Kossentini, “Partitioning of video objects into temporal segments using local motion information”, Proceedings of IEEE ICIP Conference, September 2000.

[38] B. Erol and F. Kossentini, “Similarity Matching of Arbitrarily Shaped Video by Still Shape Features and Shape Deformations”, Proceedings of IEEE ICIP Conference, October 2001.

[39] B. Erol and F. Kossentini, “Color Content Matching of MPEG-4 Video Objects”, Proceedings of IEEE Pacific-Rim Conference on Multimedia, 2001.

[40] Thomas Sikora, “The MPEG-4 Video Standard Verification Model”, IEEE Transactions on Circuits and Systems for Video Technology, Vol. 7, No. 1, February 1997.

[41] Noel Brady, “MPEG-4 Standardized Methods for the Compression of Arbitrarily Shaped Video Objects”, IEEE Transactions on Circuits and Systems for Video Technology, Vol. 9, No. 8, December 1999.

[42] Thomas Meier, King N. Nagan, “Automatic Segmentation of Moving Objects for Video Object Plane Generation”, IEEE Transactions on Circuits and Systems for Video Technology, Vol. 8, No. 5, September 1998.

Semantic Description and Video Database Systems

[43] Ahmed K. Elmagarmid, Haitao Jiang, Abdelsalam A. Helal, Anupam Joshi, Magdy Ahmed, “Video Database Systems: Issues, Products, and Applications”, Kluwer Academic publishers, 1997.

[44] A. Yoshitaka, S. Kishida, M. Hirakawa, and T. Ichikawa, “Knowledge-Assisted Content-Based Retrieval for Multimedia Database”, IEEE Multimedia, Vol. 1, No. 4, Winter 1994, pp. 12-21.

[45]N. Dimitrova and F. Golshani, “*Px* for Semantic Video Database Retrieval”, Proceeding of the ACM Multimedia'94, San Francisco, CA, pp. 219-226.

[46] M.R.Naphade, Igor V.Kozintsev, Thomas S. Huang, “A Factor Graph Framework for Semantic Video Indexing”, IEEE Transactions on Circuits and Systems for Video Technology, Vol. 12, No. 1, January 2002.

[47] M.R.Naphade, Igor Kozintsev, Thomas S. Huang, Kannan Ramchandran, “A Factor Graph Framework for Semantic Indexing and Retrieval in Video”, IEEE workshop on Content-Based Access of Image & Video Libraries, H.H.Island, South Carolina, June 12, 2000.

[48] M.R.Naphade, T.Kristjansson, B.Frey, T.S.Huang, “ Probabilistic Multimedia Objects Multijects: A novel Approach to Indexing and Retrieval in Multimedia Systems”, Proc. I.E.E.E. International Conference on Image Processing, Volume 3, pages 536-540, Oct 1998, Chicago, IL

[49] M.R.Naphade, Roy Wang, Thomas S. Huang, “Multimodal pattern matching for audio-visual query and retrieval”, Proc. SPIE, Storage and Retrieval for Media databases, Voulme 4315, pages 188-195, Jan 2001, San Jose, CA.

[50] M.R.Naphade, “Semantic Video Indexing using a probabilistic framework”, International Conference on Pattern Recognition, Barcelona, Spain, 3-8 September 2000

[51] M.R.Naphade, Thomas S. Huang, “A Probabilistic Framework for Semantic Indexing and Retrieval in Video”, IEEE International Conference on Multimedia and Expo, New York, 31 July-2 August 2000

[52] M.R.Naphade, “Probabilistic Semantic Video Indexing”, to be presented at Neural Information Processing Systems 2000, Denver, Colorado

[53] M.R.Naphade, “Semantic Filtering of Video Content”, SPIE, Storage and Retrieval for Media Databases, Jan 2001, San Jose.

[54] Duc A. Tran, Kien A. Hua, and Khanh Vu. “Semantic Reasoning based Video Database Systems”, Proc. of the 11th Int'l Conf. on Database and Expert Systems Applications, pp. 41-50, September 4-8, 2000, London, England.

[55] Duc A. Tran, Kien A. Hua, and Khanh Vu. “VideoGraph: A Graphical Object-based Model for Representing and Querying Video Data”, In the proc. of ACM Int'l Conference on Conceptual Modeling (ER 2000), October 9-12, Salt Lake city, USA

[56] Cyril Decleir, Mohand-Said Hacid, Jacques Kouloumdjian, “A Database Approach for Modeling and Querying Video Data”, 15th International Conference on Data Engineering, Sydney, Australia, 1999.

[57] Ashfaq Khokhar, Young Francis Day, Arif Ghafoor, “A Framework for Semantic Modeling of Video Data for Content-Based Indexing and Retrieval”, ACM Multimedia, 1999.

[58] A. B. Benitez, J. R. Smith, and S.-F. Chang, “MediaNet: A Multimedia Information Network for Knowledge Representation”, Proceedings of the SPIE 2000 Conference on Internet Multimedia Management Systems (IS&T/SPIE-2000), Vol. 4210, Boston, MA, Nov 6-8, 2000

Miscellaneous

[59] Kai-Chieh Liang, C.C.Jay Kuo, “WaveGuide: A Joint Wavelet-Based Image Representation and Description System”, IEEE Transactions on Image Processing, Vol. 8, No. 11, November 1999.

[60] Ashfaq Khokhar, E. Albuz ,E. Kocalar, “Quantized CIELab* Space and Encoded Spatial Structure for Scalable Indexing of Large Color Image Archives”, Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP).

[61] Ashfaq Khokhar, E. Albuz ,E. Kocalar, “Vector-wavelet based Scalable Indexing and Retrieval System for Large Color Image Archives” , Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing ICASSP, 1999.

[62] S. Krishnamachari and M. Abdel-Mottaleb, “Hierarchical clustering algorithm for fast image retrieval”, in IS&T/SPIE Conference on Storage and Retrieval for Image and Video Databases VII, 1999.

[63] Alex P. Pentland “From pixels to predicates: Recent advances in computational and robotic vision”, Ablex Publishing Corporation, 1986.

[64] David Marr, “Vision”, San Fransisco: W.H.Freeman.

[65] Deborah Swanberg, Chiao-Fe Shu, Ramesh Jain. “Architecture of multimedia information systems for content-based retrieval”, In Audio-Video workshop, San Diego, CA, Nov. 1992.

[66] James Monaco, “How to Read a Film: The Art, Technology, Language, History, and Theory of Film and Media”, Oxford University Press, New York, NY, 1977.

[67] D. White and R. Jain. “Similarity indexing: Algorithms and performance”, In Proc. SPIE Storage and Retrieval for Image and Video Databases, 1996.

[68] E. Oomoto, and K. Tanaka, “OVID: Design and Implementation of a Video-Object Database System”, IEEE Transactions on Knowledge and Data Engineering, Vol. 5,No. 4, August 1993, pp. 629-643.