

Video Content Modeling Techniques: An Overview

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

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 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 March 2003, the search engine offers access to over 3 billion web documents, serving more than 200 million searches per day with its Image search comprising more than 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 240 million images, videos and audios. Although search engines such as above mainly use keyword based textual search techniques on text annotations for visual information, 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 [18]. 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 Content Based Video Indexing and Retrieval (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 corresponding 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. This article primarily focuses on techniques and algorithms used for video content modeling dealing with uncompressed as well as compressed domain representation of data. We avoid the description of domain-specific content modeling schemes such as those tailored for sports, news, or medical videos.

We refer to Dimitrova et al [7] for an excellent review of the applications of these techniques that have matured into commercial systems. As our survey will reveal, 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. A classification of content modeling schemes based on gradient model of human visual perception and major contributions of research communities involved is outlined in Sidebar 1. In the next sections, we first describe low-level content modeling techniques most of which lie at level I (see Sidebar1), but some of them can be classified in level II. To cover the full spectrum of video content modeling techniques, we then review approaches that bridge the semantic gap between low-level (feature/object based) and high-level (object/concept based) representation of video content. Most of these techniques overlap level II and III, but some lie solely at level III.

Low-Level Content Modeling

Techniques at this level tend to model the apparent characteristics of “stuff” (as opposed to “things”) inside video clips. Video data is continuous and unstructured. To analyze and understand its contents, the video needs to be parsed into smaller chunks suitable for perceptual

Sidebar 1: Classification of Content Modeling Schemes

Studies into human visual perception 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 [8]. Though there exist two different schools of thought on how to model these perceptions, but the existence of primitive and semantic clues in visual information is well understood. 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 the following figure, we classify these abstractions into three categories based on the gradient model of human visual perception.

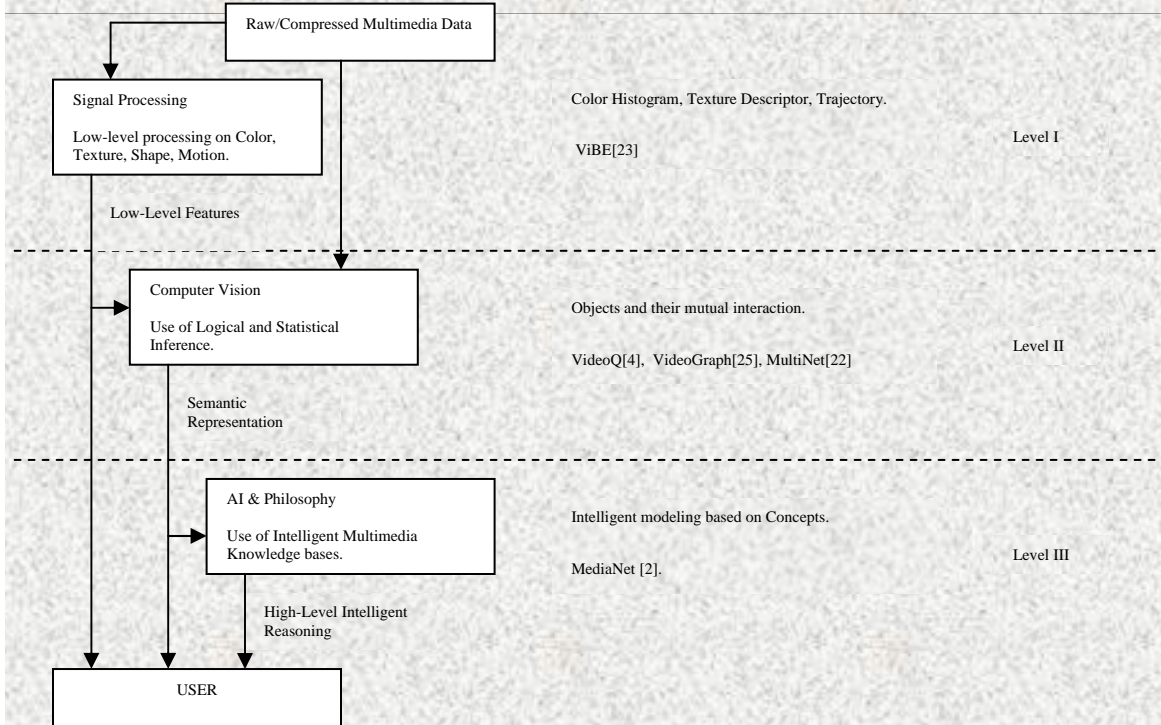


Figure: Classification of Content Modeling Techniques. Level I: Modeling of Raw Video Data. Level II: Representation of derived or logical features. Level III: Semantic level abstractions.

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”. 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”. 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”. A similar classification has been proposed by Hanjalic[13]. Shih-Fu Chang et al. [5] have provided a broad taxonomy of CBVIR systems based on functionality and mode of operation.

feature analysis or conceptual abstraction analysis. Most existing video database systems start off with temporal segmentation of video into a hierarchical model of frames, shots and scenes. The next logical step after temporal segmentation is compact representation and modeling of contents inside each shot using keyframes and objects, so as to facilitate the robust matching between any two shots at query time. In the following sections, we review techniques that rely on low-level feature

based modeling for temporal segmentation and shot representation.

Temporal Segmentation

Video data can be viewed as a hierarchical structure in which the smallest unit is a frame; a collection of frames that results 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 [19]. Transitions or boundaries between shots can be abrupt (Cut) or they can be gradual (Fade, Dissolve, Wipe). 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. For a review of major conventional shot boundary detection techniques, refer to [3] which also provides a comparison between five different techniques based on pixel difference from raw data, DCT coefficients difference and motion compensated difference.

A novel approach proposed in [17] 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. 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 [11] 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 [15] 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:

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

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

$$X_j^k = \left[(\bar{Y}_j^k - \theta_0)^T \Sigma^{-1} (\bar{Y}_j^k - \theta_0) \right]^{1/2}$$

In the expression of X_j^k , \bar{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.

Exploiting the object-based video coding paradigm of MPEG-4 (see Sidebar2), Berna et al [9] use motion as a cue for partitioning the VOs into temporal segments with uniform activity level. This motion information can be used in shot boundary detection in a video sequence as well as a feature in video objects retrieval. 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.

Shot 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

Sidebar 2: MPEG-4

MPEG-4 is the object based video coding standard meant for separate coding and manipulation of physical objects in a video clip [20], making the job of content-based indexing easier. 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. 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. The encoder sends together with the VOP, the 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-Coded and Not-Coded) are some of the major similarities in the two encoders. 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 their infancy.

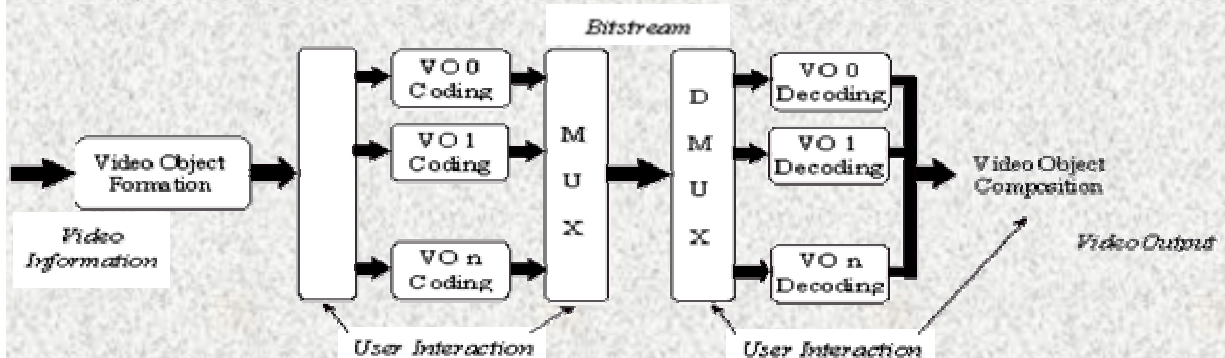


Figure: Block diagram of MPEG-4 codec.

frame from the shot, called keyframe. One approach is to use first frame of each shot as a keyframe. 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 [27] 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 [16]. They use two content descriptors, Dominant Color Histogram (DCH) and Spatial Structure Histogram (SSH), to measure content variation and to represent subshots. Chang

et al [28] 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. Delp et al [23] 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 [23] 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 [4] 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.

Exploiting the object-based video coding paradigm of MPEG-4 (see Sidebar2), Berna et al [10] 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. The process is made spatial-shift invariant by aligning the mass centers of current and candidate VOP before calculating distance between them.

Semantic Modeling

Video bitstream that contains audio stream and possibly closed caption text along with sequence of images contains a wealth of rich information about objects and events being depicted. Once the feature level summarization is done, semantic level description based on conceptual models, built on a knowledge base is needed. In the following sections, we review techniques that try to bridge the semantic gap and present a high-level picture obtained from video data.

Multimodal Probabilistic Frameworks

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. 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 [21] 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 j^{th} frame and assuming uniform priors on presence/absence of any concept in any region, the probability of occurrence of each concept in j^{th} frame is obtained from Bayes' rule as:

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

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 $F_i, i \in \{1 \dots N\}$ (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 $\chi = \{\bar{X}_1, \dots, \bar{X}_M\}$ the conditional probability of Multiject i being present in any region in the current frame is:

$$P(F_i = 1 | \chi) = \max_{j \in \{1, \dots, M\}} P(R_{ij} = 1 | \bar{X}_j)$$

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 [22]. 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.

Intelligence Based Systems

The next step towards future CBVIR systems will be marked by the full induction of intelligence into 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 [1]. Intelligence is referred to as the capabilities of 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 [2] 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. Hanjalic et al [12] have reported techniques for extraction and modeling of the *Affective* content from videos. The affective content of a video is viewed as the type/intensity of feeling/emotion mediated towards a viewer. Their computational method uses the so-called “dimensional approach to affect” concept underlined by psychophysiology studies. They obtain time curves that represent the two affect dimensions (*arousal* and *valence*) for a video from low-level video characteristics.

Video Data Models for Querying

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 model the rich information content of the video. SemVideo [24] 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 [25] 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 [14] 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

Sidebar 3: Open Issues in CBVIR Systems

In the following, we identify several open problems and research issues related to the field of content based video indexing and retrieval.

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 [26], 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. 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 fully automatic realization, but there has been a shifting trend towards more user inclusion in the process recently 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 – 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 benchmarking 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 – Any 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.

well as top-down object-oriented data abstraction concepts. Declair et al [6] 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.

Conclusions

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. Some of the future research directions and open issues are outlined in sidebar 3.

References

- [1] Benitez A.B., Smith J.R., "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.
- [2] Benitez A.B., Smith J.R., Chang S.F., "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.
- [3] Borecsky J.S., Rowe L.A., "Comparison of video shot boundary detection techniques", In Proceedings of SPIE, vol. 26670, pages 170-179, 1996.
- [4] Chang S.F., Chen W., Meng H.J., Sundaram H., Zhong D., "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.
- [5] Chang S.F., Smith J.R., Beigi M., Benitez A.B., "Visual Information Retrieval from Large Distributed On-line Repositories", Communications of the ACM, Vol. 40, No. 12, pp. 63-71, Dec 1997.
- [6] Declair C., Hacid M.S., Kouloumdjian J., "A Database Approach for Modeling and Querying Video Data", 15th International Conference on Data Engineering, Sydney, Australia, 1999.
- [7] Dimitrova N., Zhang H.J., Shahraray B., Sezan I., Huang T., Zakhori A., "Applications of Video-Content Analysis and Retrieval", IEEE Multimedia, Vol. 9, No. 4, 2002.
- [8] 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.
- [9] Erol B., Kossentini F., "Partitioning of video objects into temporal segments using local motion information", Proceedings of IEEE ICIP Conference, September 2000.
- [10] Erol B., Kossentini F., "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.
- [11] Hanjalic A., "Shot-Boundary Detection: Unraveled and Resolved?", IEEE Transactions on Circuits and Systems for Video Technology, Vol. 12, No. 2, February 2002.
- [12] Hanjalic A., Xu L.Q., "User-oriented Affective Video Analysis", IEEE Workshop on Content-based Access of Image and Video Libraries, in conjunction with the IEEE CVPR 2001 conference, December 2001, Kauai, Hawaii (USA).
- [13] Hanjalic A., "Video and image retrieval beyond the cognitive level: the needs and possibilities", in SPIE Proc. Storage and Retrieval for Media Databases 2001.
- [14] Khokhar A., Day Y.F., Ghafoor A., "A Framework for Semantic Modeling of Video Data for Content-Based Indexing and Retrieval", ACM Multimedia, 1999.
- [15] Lelescu D., Schonfeld D., "Real-time scene change detection on compressed multimedia bitstream based on statistical sequential analysis," IEEE International Conference on Multimedia and Expo, New York NY, 2000, pp. 1141-1144.
- [16] Lin T., Zhang H.J., Shi Q.Y., "Video Content Representation for Shot Retrieval and Scene Extraction", International Journal of Image & Graphics, Vol. 1, No. 3, July 2001.
- [17] Liu X.M., Chen T., "Shot Boundary Detection Using Temporal Statistics Modeling", IEEE Intl. Conf. On Acoustics, Speech and Signal Processing, ICASSP 2002., Orlando, FL, U.S., May 2002.
- [18] Mandal M.K., Idris F., Panchanathan S., "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.
- [19] Monaco J., "How to Read a Film: The Art, Technology, Language, History, and Theory

of Film and Media”, Oxford University Press, New York, NY, 1977.

- [20] MPEG-4, “Coding of Moving Pictures and Audio”, ISO/IEC JTC1/SC29/WG11 N3908 (January 2001/Pisa).
- [21] Naphade M.R., Kristjansson T., Frey B., Huang T.S., “ Probabilistic Multimedia Objects Multijets: A novel Approach to Indexing and Retrieval in Multimedia Systems”, Proc. IEEE International Conference on Image Processing, Volume 3, pages 536-540, Oct 1998, Chicago, IL.
- [22] Naphade M.R., Kozintsev I.V., Huang T.S., “A Factor Graph Framework for Semantic Video Indexing”, IEEE Transactions on Circuits and Systems for Video Technology, Vol. 12, No. 1, January 2002.
- [23] Taskiran C., Delp E.J., Bouman C.A., “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.
- [24] Tran D.A., Hua K.A., Vu K., “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.
- [25] Tran D.A., Hua K.A., Vu K., “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
- [26] White D., Jain R., “Similarity indexing: Algorithms and performance”, In Proc. SPIE Storage and Retrieval for Image and Video Databases, 1996.
- [27] Zhang H., Wu J., Zhong D., Smoliar S.W., “An integrated system for content-based video retrieval and browsing”, Pattern Recognition, Vol. 30, no.4, pp. 643-658, 1997.
- [28] Zhong D., Chang S.F., “Spatio-Temporal Video Search using the Object Based Video Representation”, IEEE International Conference on Image Processing, Oct. 26-29, 1997, Santa Barbara, CA.