A Survey on Video Data Storage for Content Based Retrieval

-Krishna Mahesh , Akella
(akella2@illinois.edu)

In accordance with course requirements of
CS 412 Introduction to Data Mining,
Fall 2011

University of Illinois ,Urbana Champaign
IL,USA -61820

A Survey on Storage and Content Based Retrieval of Multimedia data

Abstract :
A query for a video clip can be significantly different than a query for a record in a relational database . Video data includes semantics which do not have proper structure to be represented. Video data retrieved also has a temporal dependency and is a composition of audio and text in some cases.
The survey attempts to cover some of storage models and retrieval techniques used for managing multimedia data. Although multimedia includes images,audio and video ,it primarily focuses on video data. The proposed methodologies over years are discussed . Certain important characteristics of video data are examined and their applications are studied. It also covers some of the content based retrieval techniques possible with different data modeling schemas are also discussed.

Introduction:
Advances in technology have produced a deluge of multimedia data over the Internet. Due to the numerous sources which generate the such data like cameras, software,etc it can be in various formats.

Multimedia data typically entails digital images, audio, video, animation and graphics together with text data. In this survey our concentration would be on video data .

Multimedia data is different from regular data in significant ways . Especially with video and audio data there is new temporal dimension attached .

Unlike the traditional relational data bases used to store text data ,a multimedia database (MMDB) has to provide a framework for storing, processing, retrieving, transmitting and presenting a variety of media data types in a wide variety of formats.

It is therefore important to model the video data meticulously to make retrieval more intuitive and efficient.

Contents of Multimedia Database (MMDB)
An MMDB needs to store different types of information about the multimedia data. They are:

- Media data - This is the actual data representing images, audio, video that are captured, digitized, processed, compressed and stored.
- Media format data - This contains information pertaining to the format of the media data after it goes through the acquisition, processing, and encoding phases. For instance, this consists of information such as the sampling rate, resolution, frame rate, encoding scheme etc.
- Media keyword data - This contains the keyword descriptions, usually relating to the generation of the media data. For example, for a video, this might include the date, time, and place of recording , the person who recorded, the scene that is recorded, etc This is also called as content descriptive data.

- Media feature data - This contains the features derived from the media data. A feature characterizes the media contents. For example, this could contain information about the distribution of colors, the kinds of textures and the different shapes present in an image. This is also referred to as content dependent data.

  The last three types are called meta data as they describe several aspects of the media data. The media keyword data and media feature data are used as indexes for searching purpose. The media format data is used to present the retrieved information.

As can be clearly seen from above, to represent and store multimedia data correctly additional meta data is required. The queries use the meta data to retrieve the data.

Differences in Relational and multimedia queries :

With relational data the queries are direct and retrieval involves direct comparisons and table joins . Concisely, the queries in a relational database are of objective nature.

In contrast, the queries for the multimedia data are quite subjective in nature. Two users can look at the same video with different perspectives and issue different queries.

Besides ,a query can comprise of any of the above mentioned meta data . It can be the title of video, duration of the video or the content of the video.

Therefore , it is important that the meta data should be sophisticated enough to handle even content-based queries. In the following sections we shall look at some of the popular approaches at modeling video data to support retrieval with content-based queries efficiently.

Widely used Structure of Video data :
A stream of video data is stored as contiguous groups of frames called segments. Each segment comprises of collections of frames. Usually video comparisons compare each frame to find the similarity between two video streams.

The efficiency of a model depends on the method used to decompose a given video stream . It is the amount of information stored along with the decomposed fragments that helps in intuitive retrieval of video data.

Various models have been suggested by different researchers, based on relational, hierarchical and object oriented concepts. The two major modeling schemas are:

Annotation Dependent Models:
These models heavily depend on hierarchical or relational databases. In these type of models, some textual data (meta data) about the contents of video are embedded into the video data. Usually an annotator is required to write the captions, but in some models text can be generated dynamically from the audio.

Object Oriented Models:
Object Oriented Database Management Systems are considered to be good for defining multimedia models, as they can entertain both spatial as well as temporal features of any media. Another property of a media object is that it can define its own data rate, abstraction and other attributes. Again an AV (audio video) object can be accessed concurrently.

In this survey we concentrate on certain object oriented models.

- Decomposition of data into Visual Objects and their indexing.
- Decomposition of data into Themes and Thematic Indexing .

Mathematical Models :

These are object oriented models based on Bayesian networks and coordinate systems. The survey concentrates on one model Bounded Coordinate System. This is a very efficient model in terms of retrieval time and is appropriate for real time applications.

The following 3 methods segment video streams/data with aid of visual information. The shot boundaries which separate the segments are based on color analysis, visual objects ,etc. Identification of shot boundaries intuitively forms the basis of content based retrieval . We will look at some methods which use audio and text based information to segment the video data and for shot boundary detection.

Decomposition into Visual Objects [1]:

Hierarchical decomposition of Video data :

It is based on the idea that video can be considered as a structured document, with various segments from sequence or storie , to low level audiovisual objects.

In [1] they have developed a spatial segmentation method for analysis of still images into visual objects. Resulting visual objects are mainly semantic ones. Detection of some types of visual objects then contributes to video segmentation into sequences and stories.

The hierarchical decomposition of video into several levels of details (Figure 1) is well known among most research .

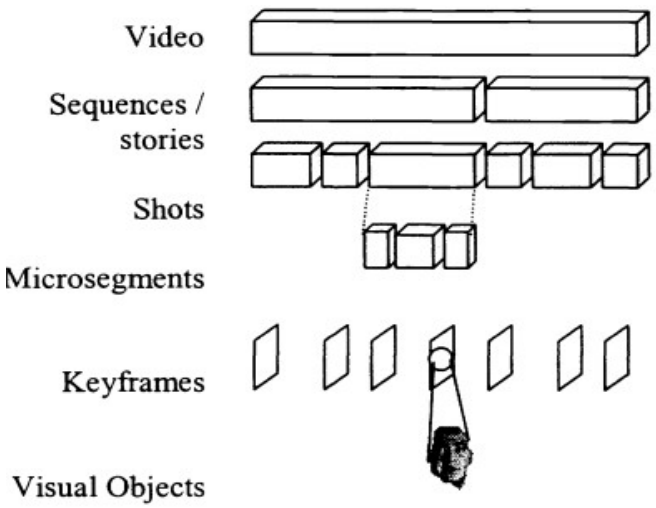


Fig. 1 : hierarchical decomposition of video

The main objective is to split the documents into elementary temporal segments . The video decomposition starts with detection of shot boundaries by using histogram-based method with

adaptive threshold selection. Then a key frame/image will be identified for generation of visual objects.

Identification of the best key frame is a challenge because the key frame should represent the sequence faithfully. Every key frame is considered to be a collection of visual objects. Extraction of visual objects from the key frame is performed by application of Motion detection ,facial recognition , clusters of color patterns ,etc.

Every visual object has the following features : colors, texture, size, position, shape and possibly motion . As a result , two key frames can be compared by comparing their visual objects . Also to retrieve the most relevant sequence or shot,the visual objects can be used.

Indexing on visual objects :
Indexing is performed on the visual objects obtained from the various key frames of a video data.
To enable indexing of visual objects , the attributes of the visual objects should be assigned values. Most of this values assignment happens manually . The visual objects are associated with visual concepts. So the users should classify the visual objects as belonging to different visual concepts. For example , from the sample data the user would classify a visual concept called “face” . This is depicted in figure 2.

Fig 2 : user classified visual concept “face”

However, for a given user, concepts proposed by other people can be more or less relevant. Therefore, in the similarity retrieval process, each prototype concept can be weighted according to its relevance. Weights are automatically calculated from user feedback, based on classical information retrieval methods .

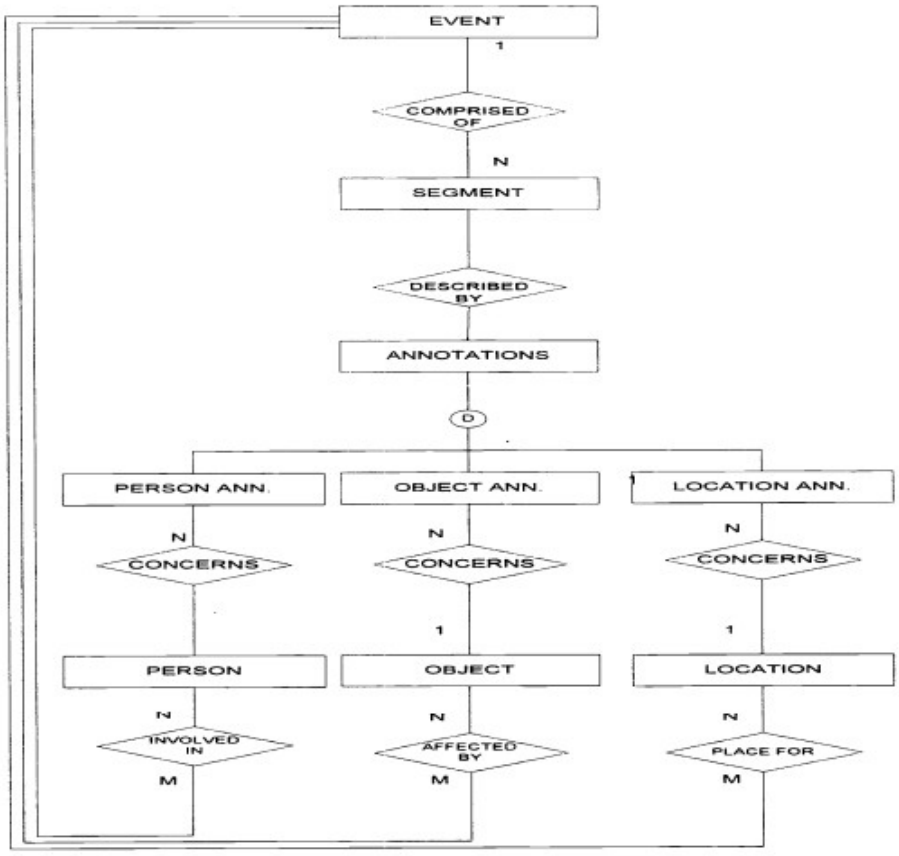
When a query is received , the most relevant weighted concept will be returned as a result.

Decomposition into themes and Thematic indexing [2] :

The model used here is called tvDBMS Model.
The object oriented approached is used in tVDBMS, so that each element is identified by a unique object identity .
This model also uses a hierarchical structure but differs in some ways.

In this model , the structure of a video document is represented by a hierarchy of structural components. Each structural component identifies a frame sequence that consists of the frames that belong to the component. The entire video document or a single frame can also become a structural component, but a whole document is too coarse as a level of abstraction and a single frame is rarely a unit of interest.
They use the abstractions such as scenes or events to reference video information. By using these abstractions it is easier to comprehend contents of video . Therefore more emphasis in tVDBMS is given to metadata which are categorized as Event, Location, Person, Object_of_Interest.

The following is the component structure adopted in the tvDBMS model.

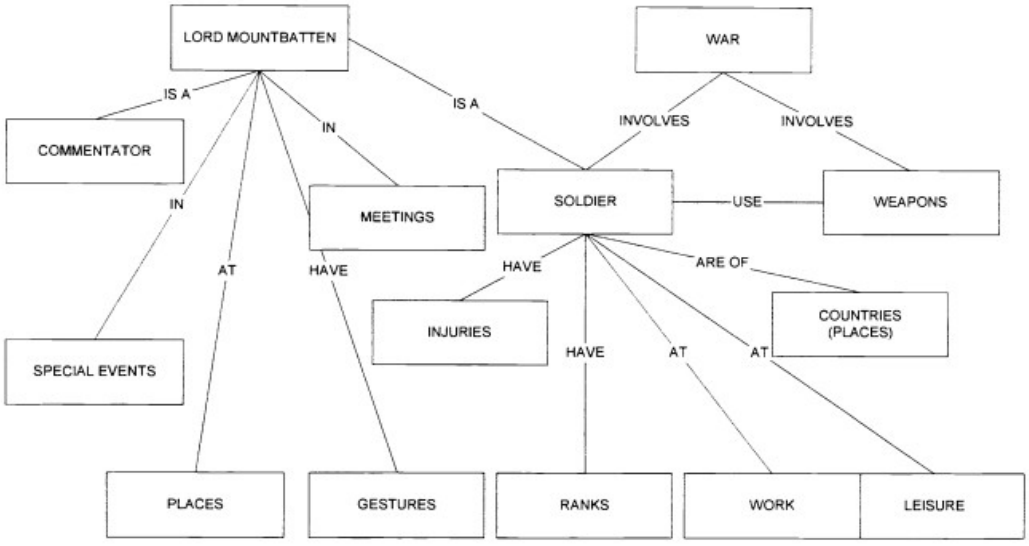


In thematic indexing ,themes of a video are identified.

A theme is defined as a topic of discourse or discussion . While watching a movie, one can find that many themes are interwoven within and around a story.

For example , In the documentary of Lord Mountbatten, it was realized that two main themes substantially form the whole story of Lord Mountbatten. One theme is of Lord Mountbatten himself, his life, his career as a commander in the armed forces and as a dignitary in Burma and India. The second theme is about the war between the Allied Forces and the Japanese. Since a war is fought between nations, soldiers are involved, weapons are used and injuries and fatalities occur. Here the two themes combine as Lord Mountbatten himself is a soldier and a soldier fights a war.

An object model for the above themes is represented in following figure.

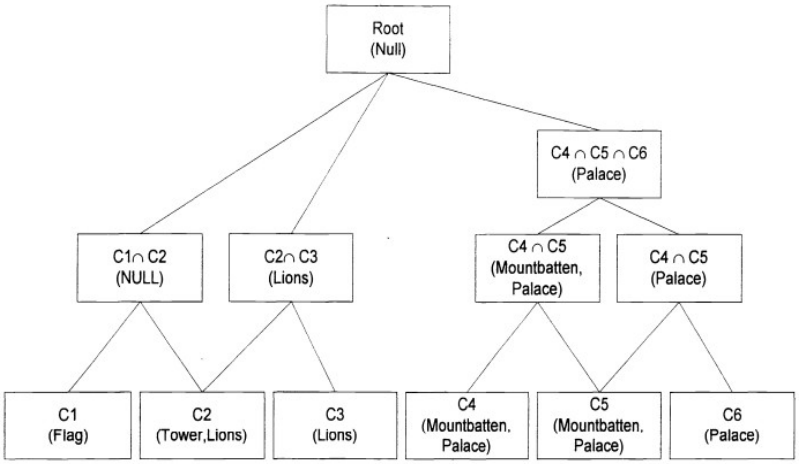


As can be seen the two themes are quite interrelated.

Now the objects are identified from the video clips and are annotated according to the above mentioned model.

Objects in the video clips:
A video is segmented into shots. Each shot is annotated with objects it contains ,the events it portrays and the people present ,etc. These video objects are then sorted in a common video tree model.

For example, in the documentary mentioned above the important objects could be Lord Mountbatten , the palace of Burma ,etc. We identify those video clips where these objects appear and annotate them with the respective object names. Since there is a possibility that multiple objects of importance can exist in the same video clip,we keep track of the connections and relations too . After careful analysis and identification of different objects ,a video tree like the following might result.



In a video shot there can be many objects that are more important then the other objects. For example, the person being shown in the shot is more important than the background objects. So more metadata should be provided for important objects with more links than for unimportant objects .

In the above figure , each node represents the clip and the video object associated with it . The intersection of C_i and C_j denotes that the object occurred in both the clips .

Once we have the tree ready then the video clips are annotated with the relationships mentioned in object model before. Indexes are then created for the video objects so identified. There can be multiple indexes for both important and unimportant objects to facilitate queries on both.

The metadata is generated from audio components as well as video components. Textual information about an object is also inserted. Each visual object has its own metadata, which creates a many-many relationship with the segment metadata. Upon association of metadata to the video clips ,retrieval is straight forward.

Based on the video segmentation concepts discussed so far ,one of the examples where this kind of segmentation can be applied is to the news videos . The frames are relatively easy to identify as they are just alternations of news reader and the news footage. By identifying the primary objects in the news footage and annotating with the voice modulations ,we can easily form indexes which makes the appropriate news retrievable.

The approaches discussed till were produced in the late 90s and are widely in use in most of the systems . There are a quite a few altogether different modeling approaches proposed later. One such approach is to use Bounded Coordinate System to model the video data.

Representation of Video Frames with Bounded Coordinate System (BCS) [3] :

This approach seizes an important characteristic of video clips that a clip often has a central theme that shows a “moment of significance, humor, oddity, or prodigy performance.” A video clip is likely to have the dominating visual content and corresponding content changing trends to express the moment.

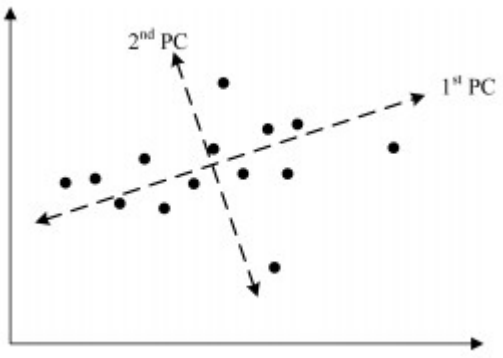
Inspired by this, a novel video clip representation model called the Bounded Coordinate System (BCS) is proposed. It is the first single video representative that captured the dominating content and content changing trends by exploiting frame content distribution. It summarizes a video clip by a coordinate system, where each of its coordinate axes is identified by principal component analysis (PCA) and bounded by data projections along the axis. It describes a video clip by a system origin and some bounded coordinate axes. BCS is a single video representative with a small size (only linear in the dimensionality of feature space).

Every video has certain moments of significance. Given that each moment lasts for considerable amount of time there would be limited changes in the frames . A data point is collected for each such frame and is plotted on the coordinate system whose axes represent the features of clip .The moment of significance is represented by a BCS .

Every video is regarded as a sequence of frames, each of which is typically represented by a high-dimensional feature vector. Each frame is characterized with certain features taken as the principal components with varying weights. A data point represents the the weights of different features in the coordinate system where the feature vectors form the coordinate axes. The number of such points depends on the number of frames which in turn depends on the video length and frame rate.

Similarity can be measured between two BCS representations by translation and rotation of axes . The similarity measure of BCS integrates two distances: the distance between two origins by translation, and the distance between each pair of bounded axes by rotation and scaling. The first distance indicates the global difference between two sets of frames representing the video clips, while the second distance indicates the difference of all the corresponding bounded axes which reflect the content-changing trends and ranges.

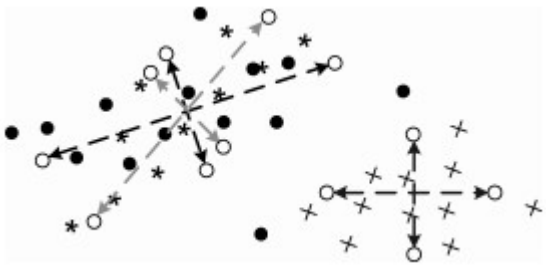
The Bounded Coordinate System (BCS) captures the dominating content and content-changing trends of a video clip by extending principal component analysis (PCA).



A two-dimensional example of PCA is shown in figure above, where the first PC indicates the direction that exhibits a larger variance, and the second PC orthogonal to the first one indicates the direction with a relatively smaller variance.

To denote the video clips with BCS, the axes of the coordinate system represent each of the strongest changing trends. A data point corresponding each frame of the clip is plotted on the coordinate axes with reference to the changing trends.

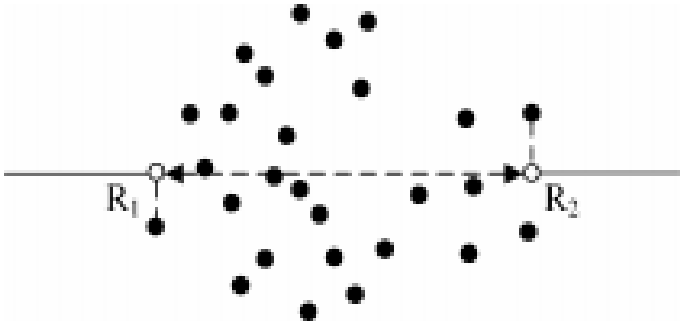
For example ,



The above figure shows two more sample video clips and their corresponding BCSs, where the * video clip (signified with *) has a very similar origin to that of the • video clip (signified with •) but different in orientation, and the × video clip (signified with ×) has a different origin and orientation from those of the * and • video clips.

Indexing of video clips based on BCS:
The B+ Tree is extended to index the BCS. They calculate a quantity called Bidirectional Transformation of a given point. Every point is transformed using this calculation to have two key values which are then stored in a B+ tree kind of structure.

Bidirectional Transformation (BDT):
Given a dataset, the one-dimensional transformation for a point P can be achieved by a mapping function D(P, R) which computes the distance between the point and the selected reference point R, where D is the distance function used. The derived one-dimensional distance values for all the points are then sorted and indexed by a B+tree. Given a query Q and a search radius r, an efficient range search [D(Q, R) − r, D(Q, R) + r] in B+ tree is performed. All the points whose distances are in the range are eligible for actual distance computations.



The intuition of BDT is that two furthermost projections (i.e., two optimal reference points R1 and R2 as shown in Figure above) separated by the first bounded principal component are far away from each other. Points that are close to one optimal reference point will be far from the other. Given a query point Q and a search radius r,
the probability for P of satisfying both

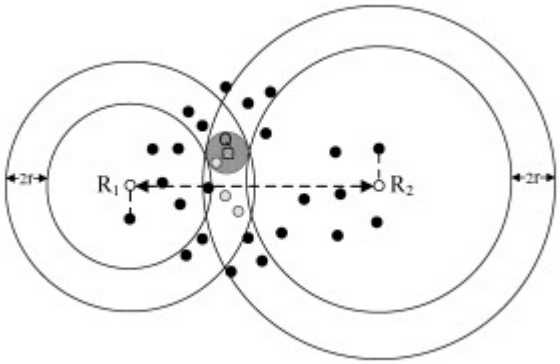
$D(Q, R1) - r \leq D(P, R1) \leq D(Q, R1) + r$ and
$D(Q, R2) - r \leq D(P, R2) \leq D(Q, R2) + r$

simultaneously is much less than that of satisfying either

$D(Q, R1) - r \leq D(P, R1) \leq D(Q, R1) + r$ or
$D(Q, R2) - r \leq D(P, R2) \leq D(Q, R2) + r$

Using BDT to retrieve the relevant video clips :
Given a query point P, the Bidistance Transformation is introduced to generate two distance values (i.e., two indexing keys) by using two furthermost projections on the first principal component as reference points. Formally, given a point P, its indexing keys are a pair of distance values computed as follows:

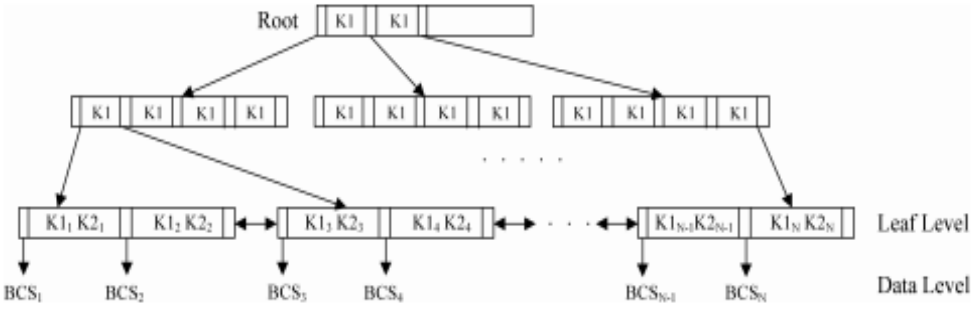
$K1(P) = D(P, R1)$
$K2(P) = D(P, R2)$.



Therefore, each point is associated with two indexing keys. Given a query Q and a search radius r, a point P can be safely pruned if it does not satisfy both
$D(Q, R1) - r \leq D(P, R1) \leq D(Q, R1) + r$ and
$D(Q, R2) - r \leq D(P, R2) \leq D(Q, R2) + r$.

The data shown in grey color are the ones returned .

After the generation of keys the tree structure looks like ,



Given a query Q, its two indexing keys are first computed. Assume K 1s are indexed. A range search on K 1s is then performed in the extended B+ -tree. At the leaf node level, the points whose K 1s are in the range of [D(Q, R1) − r, D(Q, R1) + r] are then checked as to whether their K 2s are in the range of [D(Q, R2) − r, D(Q, R2) + r]. Only the points falling into both ranges are accessed for actual distance computations.

This summarizes the video segmentation based on visual cues . There was one another approach proposed to segment video data based on the audio and text . The following part of the survey will give a brief overview about this approach.

The Study of Documentary Segmentation through Audio and Text Understanding [4] :

In this approach, the segmentation of documentary video data through audio and text understanding is investigated. To segment a continuous documentary video stream into subtopic segments, music markers and domain-independent video speech text segmentation are explored.

The primary focus is on a video documentary data as it can be considered as a long document with multiple subtopics. Defining a topic is not a simple task . In reality a topic can be considered as a stretch of discourse about something specific . So the proposal revolves around the idea that , between two contiguous pieces of discourse which are intuitively considered to be two different topics , there should be a point where there is a shift from one topic to the other . This approach concentrates on detecting this topic change with the help of background music.
In documentaries , the filmmakers intentionally use music to structure the content and communicate the semantic. There is a temporally repetitive pattern of interleaving speech and music in videos. In general, videos start with a music segment that introduces the context and end with a music segment that summarizes the video content, predicts the future and sometimes presents copyright information. In the middle, speech and music segments of variable lengths alternate. To represent this pattern we can say ,

Documentary :== Intro-music{{speech}{music}}nEnd-music

This structure suggests that interleaving music segments could be used for subtopic detection. After the above formulation , there is a 3 step procedure for subtopic detection , namely

- feature calculation
- speech/music classification
- subtopic segmentation

To be able to use the background music for semantic video segmentation , we need to differentiate the music from the speech . For this purpose 3 clip-level audio feature groups are selected. They are ZCR based,volume based , spectral flux.
By collecting the feature values some sort of clustering is adopted to assign the segments some labels . Those clips with same labels are expected to belong to the same subtopic .Finally, for each music segment in the significant group, we define its music marker as the midpoint of that temporal interval. These music markers segment a continuous video stream into subtopic segments .

Also TextTiling is used to detect subtopic shifts . TextTiling is domain independent text segmentation method. It is a technique for dividing long texts into multi-paragraph units that represent subtopics. Domain- independent text segmentation is derived from the theory of lexical cohesion , which states that text blocks with similar vocabulary are likely to be part of a coherent topic. Thus, finding topic boundaries could be achieved by detecting topic transitions that are indicated by vocabulary change.

Text-based segmentation result is then combined with that of audio segmentation.

Retrieval:
To claim a hit, the boundary of a detected text segment have to match what is manually determined and/or the corresponding music marker has to be within the temporal interval of the nearest music segment that is determined manually. A hit from both speech text segment and audio segment is counted as a appropriate clip for retrieval .

This technique can also be applied to the news example, where we can annotate the news footage by the background audio and speech recognition techniques .

Conclusion :

A multimedia database stores images, videos and other multimedia files. It should also include several indexing techniques to help in content-based retrieval. It should handle a variety of data compression and storage formats.

Video and audio are inherently temporal in nature. For example, the frames of a video need to be presented at the rate of at least 30 frames/sec. for the eye to perceive continuity in the video.

Multimedia data has significant spatial objects associated with it .The rich interpretations of spatial objects constitute multiple semantic meaning leading to multiple representation schemes that digital libraries need to possess as their human users do.

If an information retrieval function only supports a character string query of names, titles, and identification fields of spatial objects, these fields are definitely not content-based features.

A portion of interest in a video can be queried by using either

1) a few sample video frames as an example,

2) a clip of the corresponding audio track or

3) a textual description using keywords.

The retrieval of records of persons having certain facial features from a database of facial images requires both content-based queries and similarity-based retrievals. This requires indexes that are content dependent, in addition to key-word indexes .

Video documents are better handled using an object-oriented database model. This enables us to attribute significant features of the video and their values . With the help of the object oriented approach , indexes can be constructed to support content based queries.

There was significant amount of research in this are over the last two decades. Some of the important contributions are ,

I**ndexing of video data by Segmentation Model:**

Video broken down to scenes,scenes to shots,shots to frames and frames are represented with visual objects. Every visual object has certain attributes. Two visual objects are compared based on the similarity of these attribute values .Indexes are constructed on the visual objects.

Theme based hierarchical model :

Video is believed to contains two or more themes. Spatially the video clips are still broken down by segmentation model. However , some primary objects of interest like a person or a background are identified and relations between the primary objects are recorded based on an video tree model and relationship symbols. Indexes are then constructed on the primary objects. Given a query ,looking at the video tree model constructed ,all the related video clips are returned.

Representation of video frames as Bounded Coordinate System :

Every video has certain moments on significance. Given the moments last for considerable amount of time there would be limited changes in the frames . A data point is collected for each such frame and a moment of significance is represented by a BCS . We regard the video as a sequence of frames, each of which is typically represented by a high-dimensional feature vector, such as a color feature. The number of frames depends on the video length and frame rate . Each frame is characterized with certain features taken as the principal components with varying weights. A data point represents the the weights of different features in the coordinate system where the feature vectors form the coordinate axes.

Similarity can be measured between two BCS by translation and rotation of axes .

Efficiency of different models and their applications:

Given a large collection of video clips, effective management is needed to search for user interests. To do so, two essential issues have to be addressed: obtaining a compact video representation with an effective similarity measure, and organizing the compact representations with an indexing method for efficient search.

Most of the old methods needed human involvement to provide the meta data to the video clips. After the segmentation model was proposed the task is automated based on shot boundary detection. Lot of speech recognition techniques used the audio track of the video clips to annotate them and store meta data. There are several others which take user feedback to classify the video clips according to the importance of objects that have been identified.

Some other methods capture the human expressions while the video is being watched through a camera . Using these expressions the attributes like happiness,sadness ,etc are assigned to clips. In some other applications , the GPS information is incorporated within the video data while encoding it with events and objects. Then the indexes are built on the location parameters. This helps in retrieval of of video data based on the location shown in the video.

The systems like Bounded coordinate system can be very effective real time applications because of the compression of video data into mathematical structures which are comparatively easier to compare and maintain.

References:

1.Indexing and retrieval of multimedia objects at different levels of granularity Faudemay, Pascal (LIP6, University of Pierre et Marie Curie, 4 place Jussieu, 75255 Paris Cedex 05, France); Durand, Gwenaël; Seyrat, Claude; Tondre, Nicolas **Source:** Proceedings of SPIE - The International Society for Optical Engineering, v 3527, p 112-121, 1998

2.Thematic video indexing to support video database retrieval and query processing Khoja, Shakeel A. (Univ of Southampton, Southampton, United Kingdom); Hall, Wendy **Source:** Proceedings of SPIE - The International Society for Optical Engineering, v 3846, p 371-380, 1999

3.Bounded coordinate system indexing for real-time video clip search (University of Queensland,Australia)Huang, Z., Shen, H.T., Shao, J., Zhou, X., (Peking University,China) Cui, B. 2009 ACM Transactions on Information Systems 27 (3), art. no. 17

4.**The study of documentary segmentation through audio and text understanding** Dong, Aijuan (Department of Computer Science, Hood College, Frederick, MD 21701); Wang, Baoying **Source:** ICALIP 2008 - 2008 International Conference on Audio, Language and Image Processing, Proceedings, p 460-465, 2008, ICALIP 2008 - 2008 International Conference on Audio, Language and Image Processing, Proceedings

5.**The locatable video: Acquisition, segmentation, retrieval**
Wu, Yong (College of Geographical Science, Fujian Normal University, Fuzhou, China); Liu, Xuejun;Lin, Guangfa Source: ICSDM 2011 - Proceedings 2011 IEEE International Conference on Spatial Data Mining and Geographical Knowledge Services, p 229-233, 2011, ICSDM 2011 – Proceedings 2011 IEEE International Conference on Spatial Data Mining and Geographical Knowledge Services

Other references:
Knowledge-Assisted Semantic Video Object Detection.
Dasiopoulou, Stamatia ; Mezaris, Vasileios ;Kompatsiaris, Ioannis ;Papastathis, Vasileios-Kyriakos ; Strintzis, Michael G. IEEE Transactions on Circuits & Systems for Video Technology; Oct2005, Vol. 15 Issue 10, p1210-1224, 15p

Using automatic facial expression classification for contents indexing based on the emotional component
Kowalik, Uwe (Research Center of Advanced Science and Technology, University of Tokyo, Bldg.#3, 4-6-1 Komaba, Meguro-ku Tokyo, 153-8904, Japan); Aoki, Terumasa; Yasuda, Hiroshi **Source:** Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), v 4096 LNCS, p 519-528, 2006, Embedded and Ubiquitous Computing - International Conference, EUC 2006, Proceedings

I**nteractive classification and indexing of still and motion pictures in VideoRoadMap**
Park, Youngchoon (Department of Computer Science, Arizona State University, Tempe, AZ 85287-5406, United States); Golshani, Forouzan; Panchanathan, Sethuraman; Candan, K. Selçuk **Source:** Proceedings of SPIE - The International Society for Optical Engineering, v 3527, p 122-133, 1998