An Overview of Multimedia Content-Based Retrieval Strategies

Ankush Mittal
Deptt. of Electronics and Computer Engg.
Indian Institute of Technology, Roorkee,
India 247667
E-mail: ankumfec@iitr.ernet.in

Keywords: content based retrieval, syntactic indexing, semantic indexing, perceptual features, matching techniques, learning methods, mpeg-7

Received: October 14, 2005

Recently, information retrieval for text and multimedia content has become an important research area. Content based retrieval in multimedia is a challenging problem since multimedia data needs detailed interpretation from pixel values. In this paper, an overview of the content based retrieval is presented along with the different strategies in terms of syntactic and semantic indexing for retrieval. The matching techniques used and learning methods employed are also analyzed, and key directions for future research are also presented.

Povzetek: Opisane so strategije iskanja multimedijskih informacij.

1 Introduction

The last two decades have resulted in a substantial progress in the multimedia and storage technology that has led to building of a large repository of digital image, video, and audio data. There are a number of text-search engines on the web and incidentally, the sites hosting them are amongst the busiest sites. However, searching for a multimedia content is not as easy because the multimedia data, as opposed to text, needs many stages of pre-processing to yield indices relevant for querying. Since an image or a video sequence can be interpreted in numerous ways, there is no commonly agreed-upon vocabulary. Thus, the strategy of manually assigning a set of labels to a multimedia data, storing it and matching the stored label with a query will not be effective. Besides, the large volume of video data makes any assignment of text labels a massively labor intensive effort.

In recent years research has focused on the use of internal features of images and videos computed in an automated or semi-automated way [1], [2]. Automated analysis calculates statistics which can be approximately correlated to the content features. This is useful as it provides information without costly human interaction.

The common strategy for automatic indexing had been based on using syntactic features alone. However, due to its complexity of operation, there is a paradigm shift in the research of identifying semantic features [3]. User-friendly Content-Based Retrieval (CBR) systems operating at semantic level would identify motion-features as the key besides other features like color, objects etc., because motion (either of camera motion or shot editing) adds to the meaning of the content. The focus of present motion-based systems had been mainly in identifying the principal object and performing retrieval based on cues derived from such motion. With the objective of deriving semantic level indices, it becomes important to deal with the learning tools. The learning phase followed by the classification phase are two common envisioned steps in CBR systems. Rather than the user mapping the features with semantic categories, the task could be shifted to the system to perform learning (or training) with pre-classified samples and determine the patterns in an effective manner.

This paper is organized as follows. In section 2, analysis of level of abstraction of the content in CBR systems is presented. Syntactic indexing and semantic indexing are also discussed in this section. Section 3 discusses the motion feature as indexing cue with several examples. Section 4 elaborates on matching techniques in CBR systems while the learning methods in retrieval is discussed in section 5. The structure in multimedia content is discussed in section 6 followed by conclusion in section 7.

2 Level of abstraction of the content

Multimedia content can be modeled as a hierarchy of abstractions. At the lowest level are the raw pixels with unprocessed and coarse information such as color or brightness. The intermediate level consists of objects and their attributes, while the human level concepts involving the interpretation of the objects and perceptual emotion form the highest level.

Based on the above hierarchy, descriptive features in multimedia, furnished to the users of content-based technology, can be categorized as either syntactic features or semantic features [3]. A syntactic feature is a low-level characteristic of an image or a video such as an object

boundary or color histogram. A semantic feature, which is functionally at a higher level of hierarchy, represents an abstract feature such as the label grass assigned to a region of an image or a descriptor 'empathy of apprehension' for a video shot (a shot is a sequentially recorded set of frames representing a continuous action in time and space by a single camera [4],[5]). Succinctly, the retrieval process can be conceived of as the identification and matching of features in the user's requested pattern against features stored in the database. While extraction of the syntactic features is relatively undemanding, the semantic features are more appealing to the user as they are closer to the user's personal space. At higher level of user interaction the semantic features are more useful as compared to the syntactic features. For example, it is more common to have a query like "show next sports shot", "show interesting shots from a soccer match" as compared to the query "search for next zoom". One interesting point to note in the above example is that zoom-in may be one of the characteristics for an interesting shot in a soccer match but the user does not need to know it. Thus the user will not be required to construct his query in low level details in the former paradigm.

To make the distinction clearer, consider the CBR systems like QBIC [1], Virage [6], and JACOB [7], where image and video content are represented by a set of syntactic attributes like color, textures, shape, layout and global motion. The users are queried through this set of features alone. On the other hand, some examples of semantic attributes are: City vs. Landscape or Indoor vs. Outdoor (in Vailaya et al. [8]), action description of single object, interaction description of multiple objects and event recognition (in Kurokawa et al. [9]), categorization in the film genres like news cast, tennis, basketball etc. (in Mittal et al. [10]), and categorization in terms of violence or motion (Vasconcelos [11]).

2.1 Syntactic indexing

Some of the prominent CBR systems are IBM's QBIC [1], ViBE [12] at Purdue University, Visualseek [13] & VideoQ [14] at Columbia University, Photobook [15] & FourEyes [16] at M.I.T., Chabot [17] at University of California-Berkeley, MARS [18] at UIUC, Virage [6] at University of Michigan, Netra at University of California (Santa Barbara) and Jacob [7, 19] at Italy. These systems use syntactic features as the basis for matching and employ either Query-by-Example or Query-through-dialog box to interface with the user. Thus, they operate at a lower level of abstraction and therefore, the user needs to be highly versed in the details of the CBR system to take advantage of them.

Popular automatic image indexing systems (as CHABOT [17], VisualSEEK [13]) employ user composed queries which are provided through the dialog box. However this method is not convenient as the user needs to know the exact details of the attributes and their implementation as well as details of the search method. However, the operation of such systems is highly technical.

The only alternative to 'Query through dialog box' was thought to be 'Query by example' technique where the user is presented with a number of example images and he indicates the closest. The various features of the chosen image are evaluated and matched against the images in the database. The features which have been commonly used in previous work are color, shape, textures and spatial distribution. Using some distance metric, the distance between the feature vectors (i.e. vectors containing the set of features) for the example image and a database image is computed. A few images which have a distance less than a threshold are retrieved. The user browses through them and if he is not satisfied, could formulate a new query in terms of either one of the retrieved images or the old image.

There has been a parallel of 'Query by example' in the field of video indexing. The majority of work in video indexing has focused on the detection of key frames called representative frames or R-Frames [20], [21], [22]. The R-Frames are chosen based on some predefined criteria and the feature set is constructed using the R-Frames. The user is again provided as output the choice between various R-Frames of video clips which are close to the user query.

There are a number of defects with retrieving items with 'Query by example':

1. In contrast to a clearly defined text search, in image search, using 'query by example', the image can be annotated and interpreted in many ways. For example, a particular user may be interested in a waterfall, another may be interested in mountain and yet another in the sky, although all of them may be present in the same image.

2. It is reasonable for the user to wonder "why do these two images look similar?" or "what specific parts of these images are contributing to the similarity?"(see CANDID [23]). Thus the user is required to know the search structure and other details for efficiently searching the database.

3. Since there is no matching of exactly defined fields in query by example, it requires a larger similarity threshold as it usually involves many more comparisons than query via the dialog box. The number of images retrieved are so many that it makes the whole task tedious and sometimes meaningless.

We deem that there is a significant lacuna in addressing human level perception and cognitive capabilities of a common user as neither the 'Query by example' nor the 'Query through dialog box' attempt any higher-level analysis of the multimedia content. The syntactic features provided to the user may be adequate only if the goal is to find frames with similar distributions of color or texture or other low level characteristics. However the user often deals with and is more concerned with higher level objects. Rudimentary and unprocessed syntactic features inherently lack the power of descriptiveness required for the user to properly interact with and utilize CBR system. Some progress

is made recently with the work by Krishnapuram et al. [24]. They develop a fuzzy framework which can handle exemplar-based, graphical-sketch-based, as well as linguistic queries involving region labels, attributes, and spatial relations. The system uses Fuzzy Attributed Relational Graphs to represent images, where each node in the graph represents an image region and each edge represents a relation between two regions.

2.2 Semantic indexing

Researchers have recently been reviewing the appropriateness of these approaches based on syntactic features. There has been some effort in the direction of developing techniques which are based on analyzing the contents of images and videos on a higher level. A number of psychological studies and experiments emphasize the need for extracting the semantic information from images and video data. The two important researches in this direction are: a) Demonstrating that higher similarity-ratings are produced by perceptually-relevant semantic features as opposed to the features derived from color histograms on the images ([25]), and b) the performance and the efficiency of searching is generally greatly improved by using semantic cues ([26]) as compared to when low-level features are employed.

One can find a lot of work, developed lately, employing semantic technique. Shannon et al. [27] have analyzed and looked specifically at video-taped presentations in which the camera is focused on the speaker's slides projected by an overhead projector. By constraining the domain they are able to define a "vocabulary" of actions that people perform during a presentation. In the work done by Gong et al. [28], video content parsing is done by building a priori model of a video's structure based on domain knowledge. Out of the set of recorded shots, shots pertaining to news category are retrieved and the user can define his choice with respect to them. Sudhir et al. [29] have worked on automatic classification in 'Tennis'. Their approach is based on generation of an image model for the tennis court lines and players. Automatically extracted tennis court lines and the players' location information are analyzed in a high-level reasoning module and related to useful high-level tennis play events.

Ferman et al. [30] and Naphade et al. [31] have recently employed probabilistic framework to construct descriptors in terms of location, objects and events. Vasconcelos et al. [11] have integrated shot length along with global motion activity to characterize the video stream with properties such as violence, sex or profanity. An interesting insight that comes out from their work is that there exists a relationship between the degree of action and the structure of visual patterns that constitute a movie.

Hanjalic [32] has given a framework for adaptive extraction of highlights from a sport video based on excitement modeling. The system utilizes the expected variations in a user's excitement by observing the temporal behavior of selected audiovisual low-level features and the editing scheme of a video. Another work by Rasheed et al. [33] classifies movies into four broad categories: Comedies, Action, Dramas, or Horror films. Inspired by cinematic principles, four computable video features (average shot length, color variance, motion content and lighting key) are combined in a framework to provide a mapping to these four high-level semantic classes. Mean shift classification is used to discover the structure between the computed features and each film genre.

Recently, many researchers have worked in semantic image classification and natural image database organization into categories like Indoor vs. Outdoor ([34], [8] etc.), city vs. landscape ([35],[36] etc.), man-made vs. natural ([8],[37]), sunset vs. forest vs. mountain ([38] and so on.

3 Motion feature as indexing cue

Since it is often through motion that the content in a video is expressed and the attention of the viewers captivated, we review here some prominent work that has used motion features as indices for video classification.

Dimitrova et al. [39] have used object motion recovery for video classification and querying. From the low-level motion analysis, they build motion vectors. 'N-tuples' of motion vector constitute each trajectory. At the high-level motion analysis they associate an activity to a set of an object using domain knowledge rules. The visual query system allows the user to specify the path of a moving object like a player.

Courtney [40] detects moving object in the video sequence using motion segmentation module. By tracking the individual objects through the segmented data, a symbolic representation of the video is generated in the form of a directed graph describing the objects and their movements. This graph is then annotated using a rule-based classification scheme to identify the events of interest like appearance/disappearance, entrance/exit, and motion/rest of objects. He suggests the potential application of such a technique to surveillance video analysis.

Nam et al. [41] developed a scheme for video indexing based on the motion behavior of video objects. Moving objects are extracted by analyzing the layered images constructed from coarse data in 3-D wavelet decomposition. The moving objects are modeled as collections of interconnected rigid polygonal shapes and the motion signatures of these objects are computed and stored as potential query terms.

Recently developed VideoQ [14] brings up the idea of an animated sketch to formulate queries. In an animated sketch, motion and temporal duration are the key attributes assigned to each object in the sketch in addition to the usual attributes such as shape, color and texture. Using the visual palette, a scene is sketched out by drawing a collection of video objects. According to its theory, it is the spatio-temporal ordering and relationships of these objects that fully define a scene. However, since VideoQ only provides

for the temporal sketching of dominant object motion in 2-D space for querying, these queries are very technical. Note that imagining such sketches is not a straightforward task.

A key observation from most of the above studies is that high-level index formation has not been the main concern. These researchers were more interested in deriving low level descriptors such as the dominant direction, distribution of flow and trajectory of the object.

4 Matching techniques

In this section, matching techniques used in the popular CBR systems are examined. By matching technique, we mean the method of finding similarity between the two sets of multimedia data, which can either be images or videos. The parameters of such a technique, which we discuss and analyze herein, are:

1. Level of abstraction of features
2. Distance measures
3. Normalization of features, if supported, or else the method of relatively weighing the features.

In VisualSEEK [13] a query is specified by the colors, sizes and arbitrary spatial layouts of the color regions, which include both absolute and relative spatial locations. A query specified by the user is translated directly into pruning operations on intrinsic parameters. For example, given the single region query: to find the region that best matches Q =$\{ c_q, (x_q, y_q), area_q, (w_q, h_q)\}$, the query is processed by first computing the individual queries for color, location, size and spatial extent. Each of the color, size and location measures form different modules with each module utilizing a specific distance measure. The intersection of the region match lists is then computed to obtain a set of common images. Finally, the single region distance is given by the weighted sum of the color set($d^{set}_{q,t}$), location ($d^{s}_{q,t}$), area($d^{a}_{q,t}$) and spatial extent distances ($d^{m}_{q,t}$). The best match minimizes the total distance.

In JACOB [7], queries are based on color and texture measures. The user chooses a value between 0 and 1 to indicate the relative importance of a set of features over each other. Apart from this naive procedure no other technique for normalization is implemented. In QBIC (Query by Image Content) [1], the query is built on either color, texture, or shape of image objects and regions. QBIC computes each of the features by separate distance measures. The distance measure used for each feature is the weighted Euclidean measure where the weights reflect the importance of components of each feature. CHABOT [17] facilitates image search based on features like location, colors and concepts, examples of which are 'mostly red', 'sunset', 'yellow flowers' etc. Equal weightage is assigned in this system to all the features in retrieving the image.

A common strategy can be discerned in these different CBR systems: they employ only low level features with distance measures similar to Euclidean distance, with no method to automatically generate the weights of the features.

None of the indexing schemes discussed so far is capable of dealing with multimodal distribution. Another problem which may arise is that the probability distribution may not be Gaussian, even though it may be unimodal. The distance measures used by these systems inherently assume that with increasing distance from the mean vector, the probability decreases. Thus, some sort of Gaussian assumption is implicitly accepted. This is the case for the Bayesian Network employed in [30] which may turn out to be ineffective.

Identifying the meaningful set of features for a given domain is important yet unexplored. Many systems (like JACOB [7]) either resort to having the user specify the relative weights to the features or like CHABOT [17], they assign equal weightage to all the features in retrieving the image or video shot. By asking the user to specify the weight of various features, an injudicious assumption is made that the user is knowledgeable enough to ascertain these to a fine degree. To rely upon human experience is not a pragmatic approach when the aim is to build an integrated system with quite a few classes and many features. Different researchers (like Doulamis et al. [42], Peng et al. [43], and Sheikholeslami et al. [44]) have identified the importance of automatically identifying the relevance of the features. They have used different variations of neural network approaches in trying to achieve this task. A technique is required by which the relevant features for a class are automatically extracted and a higher relevance is assigned to them as compared to the other features. Moreover, the issue of dealing with diverse feature measures by normalization or otherwise has not been properly dealt with.

5 Learning methods in retrieval

Recently, strategies involving learning a supervised model are emerging in the field of CBR. When there are clearly identified categories, as well as, large domain-representative training data, learning can be effectively employed to construct a model of the domain. A model generally represents a strong spatial order within the individual images and/or a strong temporal order across a sequence. In this section, the learning strategy, the domains, as well as, the learning tools are discussed with reference to various research projects.

Minka et al. [45] use an interactive learning system (based on relevance feedback) based on a society of models. Instead of employing universal similarity measure or alternately manual selection of relevant features, this approach provides a learning algorithm for selecting and

combining groups of the data. The user generates both the positive and negative retrieval examples (relevance feedback). A greedy strategy is used to select a combination of existing groupings from the set of all possible groupings.

Yang and Kuo [46] propose a hierarchical procedure using a two-level classification based on K-Nearest neighbor classifier. The coarse classification uses low-level image features, while the fine level classification is based on semantic meanings. In the coarse classification, color and edge information analysis is used to summarize the image collections with image models. In fine classification, a supervised training algorithm based on multiple feature templates is adopted to refine the classification result of each coarse class.

Demsar et al. [47] have used decision-trees for classification of images based on user's feedback with positive and negative examples. Their work is in the domain of retrieving images with particular color combination (like sunset images, images containing human faces etc.).

Ratan et al. [48] have used a multiple-instance learning scheme to model ambiguity in the supervised learning examples in natural scenes. Each image can represent multiple concepts. To replace one of these ambiguities, each image is modeled as a bag of instances (sub-blocks in the image). A bag is labeled as a positive example of a concept, if there exist some instances representing the concept, which could be a car or a waterfall scene. If there does not exist any instance, the bag is labeled as a negative example. The concept is learned by using a small collection of positive and negative examples and this is used to retrieve images containing a similar concept from the database.

Torabba et al. [37] have used discriminant structural templates for organizing scene along various semantic axes. They classify the global scene representation of an image in the following axes: degree of naturalness (artificial/natural images) and degree of openness (panoramic views/closed environments). A supervised learning stage using linear discriminant analysis (LDA) is used to generate the decision boundaries along the various semantic axes. The classification is based on Gabor textures derived from the sub-blocks of the image.

Naphade et al. [31] use Markovian framework to build probabilistic multimedia objects called multijects, which are fused from low-level features from the multiple modalities. A probabilistic framework is used to encode the higher level relationship between the multijects, which enhances or reduces the probabilities of concurrent existence of various multijects. The fundamental components of their model are sites, objects and evenets and the model is evaluated to detect explosions and waterfalls in the movies.

In the previously discussed work by Fischer et al. [49], the classification module is comparable to a human expert who is asked about his/her evaluation of closeness of a particular feature. The estimates of different classification modules are combined into a final guess. Their strategy depends on the assistance from a human knowledge base to distinguish the style profiles of the features.

The researchers working in the semantic image classification have typically used color, texture, objects etc. as features for mapping to higher level concepts by learning through K-nearest neighbor (like Szummer et al. [34]), Rule-based systems (Gorkhani et al. [35]) , Linear discriminant analysis (Torralba et al. [37]), Vector quantization (Vailaya and Jain [38]), Decision trees (Forsyth et al. [36]) and Support vector machine (Sadlier et al. [50]). Recently, a framework for cluster-based retrieval of images by unsupervised learning is also proposed by Chen et al. [51]. Data mining techniques have also been employed to bridge the gap between semantic labels and low-level features. In particular, association rule mining has been used by Zhu et al. [52] for semantic indexing and event detection.

The ability to infer high-level understanding from a multimedia content has proven to be a difficult goal to achieve. The goal is to present supervised learning framework where the content of the semantic indices are properly modeled and learnt. Of course, not all semantic categories can be understood and extracted by present algorithms easily, for example, the category "John eating icecream". Such categories might require the presence of sophisticated scene understanding algorithms along with the understanding of spatio-temporal relationship between entities (like the behavior eating can be characterized as repeatedly putting something eatable in mouth).

However, there are still a large number of multimedia categories (especially in the domain of video) that demonstrate structure in their elements. This structure can be exploited to build models. The structure is present because content creation is not a random process, but rather, it obeys a series of well-established codes and conventions. These structures in many cases can be detected by paying more attention to the features directly encoding knowledge or manifesting psychological significance. Automatic techniques for properly mapping the feature space to the high-level descriptors are then required, otherwise, the design process for CBR system becomes highly complex with hundreds of features and a large number of categories.

6 Structure in multimedia content

Most multimedia data are viewed as part of a casual activity, for example, people customarily watch news over breakfast, watch movies while talking on the phone, and listen to radio while driving [53]. This requires only a share of the viewer's cognitive resources and therefore, the message is generally laid out in a way that minimizes the effort required to decode it. Furthermore, to achieve efficiency in content-production and due to the limited number of available resources, standard techniques are employed. While there are clear incentives for innovation, content production evolves by building on previously developed formulae that have sustained the testing of time and market [54, 55]. Thus, it can be naturally assumed that most of the video content exhibits a significant amount of structure in its ele-

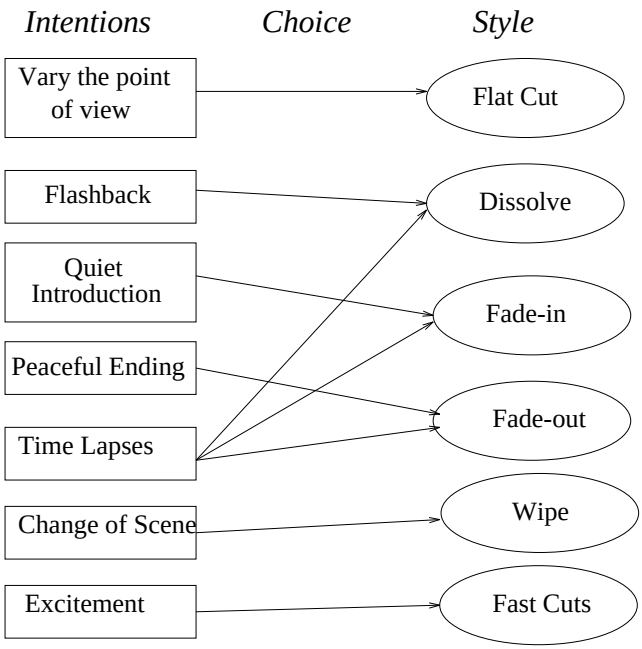

Figure 1: Conveying the meaning through styles

ment.

The structures are present as a result of the s table nature of the world and the ways in which the viewers perceive and interact with the world. For a perceiver to develop the inferential leverage necessary to disambiguate among several conflicting configurations of the world, the world must behave regularly [56, 57]. In other words, these structures embody the relationship of the observers with the world [58] or in this case, the virtual world presented by the video. Movie picture viewing or communication is possible due to constancies of these relationships. These patterns of interaction also make it possible to represent the events or movie theme. In this section, why and how the video classes are generally structured are considered from two angles: from the producer's end and from the nature of the content itself.

Some relevant works have been done founded around the observation that the media has structure. In the work by Fan et al. [3], the hierarchical structure of the semantics-sensitive video classifier is derived from the domain-dependent concept hierarchy of video contents in the database. Relevance analysis is used to shorten the semantic gap by selecting the discriminating visual features and suitable importance. The EM algorithm is used to determine the classification rule for each visual concept node.

6.1 The designer's end

The intention of video making is to represent an action or to evoke emotions using various storytelling methods. Figure 1 gives an analysis of the basic techniques of shot-transitions that are used to convey particular intentions. A similar study can be done for camera motion, lighting effects etc. (please refer to cinematographic literature [55, 59]). For example, panning for a long duration is used for 'an establishing shot', zooming-in is used for increasing the interest of the user and so on. Although these rules are mere guidelines and can be violated, their use tends to deepen the filmic reality. Consider, for example, that a director fails to use fast cutting at a scene of climax in a movie. This would reduce the thrill in the mind of the audience, although the entire set-up remains the same. Nack and Parkes [60] remark, while comparing movies with a theatrical performance, that the denotative material of the film becomes real through the audience's identifications and projections. This leads to a generic deduction that film styles like editing effect, movement of the camera, subjects in the frame, colors, variation of lighting effects etc. are meaningfully-directed and intentional.

6.2 The nature of content

The structure of a multimedia class like sports, commercials, news etc. also stems from the pattern inherent in the material that is portrayed. These patterns then become the characteristics of the class and distinguish it against others. To illustrate the point with some common examples: car-race video has unusual zoom-in and zoom-out, basketball has left-panning and right-panning that last for certain maximum duration (say 20 seconds), the color of tennis sequence is mainly restricted to that of the four types of court according to the international standards, the motion activity in interesting shots in sports is higher than its surrounding shots and so on. As examples of works demonstrating structure in domain other than movies, it was shown (Mittal et al. [61]) that the classes form separable cliques in the feature spaces (with feature representation improved by making it fine-grained) and reasonable classification accuracy is achieved. Another work by Eickeler et al. [62] exploits the special structure of news in 'begin shot', 'newscaster shot', 'interview', 'weather forecast' etc. and builds a video model of news. The feature vectors are modeled and classified using HMMs in the domain of broadcast news.

6.3 Discussion and Future of CBR systems

The analysis presented in the paper has two implications. First is that since there is some structure in the film-making process, there is a possibility of deriving some conclusions about the intentions or meaning conveyed through a shot. Of course, as Figure 1 shows there is ambiguity in making such conclusions, for example, dissolve can be either due to 'flashback' or due to 'time lapse'. However, by inclusion of several cues, especially context, much clearer distinction is possible. To take the same example, in a moving window of seven shots, if the number of dissolves is two, the dissolves belong most probably to 'flashback'; however, if it is more than two, the dissolves probably denote a 'time lapse'. The second implication is that the cinematic

theories of psychology and techniques used by cameramen and directors in making a film clearly expound the need to have features which have psychological relationship with humans. Naturally, the integration of higher-level features would increase the classification accuracy of video classes belonging to non-movie domains like news, soccer etc.

The process of information representation remains incomplete without the features which are at a perceptual level. Perceptual-level features also reveal fundamental structure about the content of the video data. For example, the presence of zoom-in, followed by relative camera stability, have been shown to be a good indicator of interesting shots in the home videos. These structures are such a fundamental characterization of the multimedia classes that even including a large number of classes in the CBR system does not cause problems in distinguishing them from one another. In other words, such primitives are applicable in general environments.

To realize the need for CBR system, the systematic development of the new member of the MPEG family, called "Multimedia Content Description Interface" (in short 'MPEG-7') is currently pursued. MPEG-7 will extend the limited capabilities of proprietary solutions in identifying content that exist today, notably by including more data types. In other words, it will specify a standard set of descriptors that can be used to describe various types of multimedia information. MPEG-7 will also standardize ways to define other descriptors as well as structures (Description Schemes) for the descriptors and their relationships. This description (i.e. the combination of descriptors and description schemes) will be associated with the content itself, to allow fast and efficient searching for material of a user's interest.

7 Conclusions

While the above works in the semantic domain disclosed the potentiality of description in semantic terms, a systematic exploration of construction of high-level indexes is lacking. The literature survey presented before evinces the fact that most systems operate only at syntactic level and provide low-level descriptors such as color, shape, and textures. Some attempted work at semantic level (for example, [27], [28]) confined themselves to data modeling in specific domains. Other works at semantic level (for example, [11], [49]) exclusively tried to derive semantic properties from low-level properties. This paradigm of deriving semantic indices needs to be explored further. However, none of the work has considered exploring features close to the human perception.

The need to have features which have psychological relationship with human is clearly expounded by cinematic theories of psychology and techniques used by cameramen and directors in making a film. Nack and Parkes [60] remark, while comparing movies with a theatrical performance, that the denotative material of the film becomes real through the audience's identifications and projections. One aspect of film reality is, therefore, the imagination of the audience. Consider our experience of camera movement as it appears on the screen prior to our conscious reflection about it. The experience is a relatively 'invisible' one - particularly if we are used to viewing narrative rather than experimental films. We become aware of camera movement as our movement and perceive the camera as an invisible but present subject. This has a lot of implication in creating semantic indices as camera movement in a video is meaningfully-directed and intentional. For example, a pan which is described as a particular rotation of the camera on its vertical axis from a stationary point, may be used to establish the contiguity of screen space, and leads the viewer to understand and feel from this expression the 'sweep' and 'scope' of a monument valley landscape and the stagecoach crossing it ([63]).

In summary, there is a great need to extract semantic indices for making the CBR system serviceable to the user. Though extracting all such indices might not be possible, there is a great scope for furnishing the semantic indices witha certain well-established structure.

References

[1] M. Flickner et al. Query By Image and video Content : the QBIC system. *IEEE Computer*, pages 23–32, September 1995.

[2] H.J. Zhang, C.Y. Low, S.W. Smoliar, and J.H. Wu. Video Parsing Retrieval and Browsing : an Integrated and Content Based Solution. In *Proc. of Multimedia '95, San Francisco, CA USA*, pages 15–24, 1995.

[3] J. Fan, A. K. Elmagarmid, X. Zhu, W. G. Aref, and L. Wu. Classview: Hierarchical video shot classification, indexing, and accessing. *IEEE Transactions on Multimedia, vol. 6*, pages 70–86, 2004.

[4] Y. A. Aslandogan and C. T. Yu. Techniques and systems for image and video retrieval. *IEEE transactions on Knowledge and data engineering, Vol. 11 No. 1*, pages 56–63, 1999.

[5] A. K. Jain, A. Vailaya, and X. Wei. Query by video clip. *Multimedia systems*, pages 369–384, 1999.

[6] J. Bach et. al. The virage search engine: An open framework for image search engine. In *SPIE conference on Storage and retrieval of Image and video databases*, pages 76–87, 1996.

[7] M. L. Cascia and E. Ardizzone. JACOB : Just a content-based query system for video databases. In *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 1216–1219, May 1996.

[8] A. Vailaya and A. K. Jain. Incremental learning for bayesian classification of images. *International Conference on Image Processing*, 2:585–589, 1999.

[9] M. Kurokawa, T. Echigo, A. Tomita, J. Maeda, H. Miyamori, and S. Iisaku. Representation and retrieval of video scene by using object actions and their spatio-temporal relationships. *International Conference on Image Processing*, 2:86–90, 1999.

[10] A. Mittal and L.-F. Cheong. Addressing the problems of bayesian network classification of video using high-dimensional features. *IEEE Transactions on knowledge and data engineering*, pages 230–244, 2004, vol. 16.

[11] N. Vasconcelos and A. Lipman. Towards semantically meaningful feature spaces for the characterization of video content. *Proc. of Int. Conf. on Image Processing*, pages 25–28, 1997.

[12] J.-Y. Chen, C. Taskiran, E. J. Delp, and C. A. Bouman. Vibe: A new paradigm for video database browsing and search. *Workshop on Content-based access of image and video libraries*, pages 96–100, 1998.

[13] J.R. Smith and S. F. Chang. VisualSEEK: a fully automated content-based image query system. *ACM Multimedia*, pages 87–98, November 1996.

[14] S. F. Chang, W. Chen, H. J. Meng, H. Sundaram, and D. Zhong. VideoQ: An automated content based video search system using visual cues. *ACM Multimedia*, pages 313–324, 1997.

[15] A. Pentland, R. Picard, and S. Sclaroff. Photobook: Tools for content-based manipulation of image databases. In *SPIE conference on Storage and retrieval of Image and video databases*, pages 34–47, 1994.

[16] R. W. Picard and T. P. Minka. Vision texture for annotation. *Multimedia systems*, 3:3–14, 1995.

[17] V. E. Ogle and M. Stonebraker. CHABOT: Retrieval from a relational database of images. *IEEE Computer*, pages 40–48, September 1995.

[18] S. Mehrotra, Y. Rui, M. Ortega, and T. S. Huang. Supporting content-based queries over images in mars. *Internation conference on multimedia computing and systems*, pages 632–633, 1997.

[19] E. Ardizzonei and M. L. Cascia. Automatic video database indexing and retrieval. In *Multimedia Tools and Applications, Kluwer Academic Publishers, Boston MA*, 1996.

[20] K. Otsuji and Y. Tonomura. Projection-detecting filter for video cut detection. *Multimedia Systems, Vol. 1*, pages 205–210, 1994.

[21] R. Zabih, J.Miller, and K. Mai. Video browsing using edges and motion. *IEEE Int. Conf. on Computer Vision and Pattern Recognition*, pages 439–446, 1996.

[22] H. Zhang, A. Kankanhalli, and S. W. Smoliar. Automatic partitioning of full motion video. *Multimedia Systems, Vol.1*, pages 10–28, 1993.

[23] P. M. Kelly, T. M. Cannon, and D. R. Hush. Query by image example : the CANDID approach. *Proc. of the SPIE, Storage and Retrieval for Image and Video Databases III, Vol. 2420*, pages 238–248, 1995.

[24] R. Krishnapuram, S. Medasani, S.-H. Jung, Y. Choi, and R. Balasubramaniam. Content-based image retrieval based on a fuzzy approach. *IEEE Trans. Knowl. Data Eng., vol. 16*, page 2004, 1185-1199.

[25] B. E. Rogowitz, T. frese, J. R. Smith, C. A. Bouman, and E. Kalin. Perceptual image similarity experiments. *SPIE Conference on Human and Electronic Imaging*, pages 576–590, 1998.

[26] T. V. Papathomas, T. E. Conway, I. J. Cox, J. Ghosn, M. L. Mitter, T. P. Minka, and P. N. Yianilos. Psychophysical studies of the performance of an image database retrieval system. *SPIE Conference on Human and Electronic Imaging*, pages 591–602, 1998.

[27] X. J. Shannon, M. J. Black, S. Minneman, and D. Kimber. Analysis of gesture and action in technical talks for video indexing. *IEEE Int. Conf. on Computer Vision and Pattern Recognition*, pages 595–601, 1997.

[28] H.J. Zhang, Y. Gong, S.W. Smoliar, and S. Y. Tan. Automatic parsing of news video. In *Proc. Of Int. Conf. On Multimedia Computing and Systems, Boston, Massachusetts, USA*, pages 45–54, May 1994.

[29] G. Sudhir, C. M. Lee, and A. K. Jain. Automatic classification of tennis video for high-level content-based retrieval. In *IEEE Workshop on Content-based Access of Image and Video Databases*, 1998.

[30] A. M. Ferman and A. M. Tekalp. Probabilistic analysis and extraction of video content. *Proc. Of ICIP*, pages 91–95, vol. 2, 1999.

[31] M. R. Naphade, T. Kristjansson, B. Frey, and T. S. Huang. Probabilistic multimedia objects (multijects): A novel approach to video indexing and retrieval in multimedia systems. *Proc. of ICIP*, pages 536–40, 1998.

[32] A. Hanjalic. Adaptive extraction of highlights from a sport video based on excitement modeling. *IEEE Transactions on Multimedia, vol. 7*, page 2005, 1114-1122.

[33] Z. Rasheed, Y. Sheikh, and M. Shah. On the use of computable features for film classification. *IEEE Trans. Circuits Syst. Video Techn., vol. 15*, page 2005, 52-64.

[34] M. Szummer and R. W. Picard. Indoor-outdoor image classification. *Int. Workshop on Content-based access of image and video databases*, 1998.

[35] M. M. Gorkhani and R. W. Picard. Texture orientation for sorting photos. *International conference on Pattern recognition*, pages 459–464, 1994.

[36] D. Forsyth, J. Malik, M. Fleck, H. Greenspan, T. Leung, S. Belongie, C. Carson, and C. Bregler. Finding pictures of objects in large collections of images. *Int. Workshop on Object Recognition for Computer Vision*, pages 335–360, 1996.

[37] A. B. Torralba and A. Oliva. Semantic organization of scenes using discriminant structural templates. *International Conference on computer vision*, pages 1253–1258, vol. 2, 1999.

[38] A. Vailaya and A. K. Jain. Reject option for vq-based bayesian classification. *International Conference on pattern recognition*, pages 48–51, 2000.

[39] N. Dimitrova and F. Golshani. Motion recovery for video content classification. *ACM transactions on Information Systems*, pages 408–439, Vol. 13, No. 4, 1995.

[40] J. D. Courtney. Automatic video indexing via object motion analysis. *Pattern Recognition*, pages 607–625, Vol. 30, No. 4, 1997.

[41] J. Nam and A. H. Tewfik. Motion based video object indexing using multiresolution analysis. *Proc. of SPIE: Storage and Retreival for Image and Video database*, pages 688–695, 97.

[42] N. D. Doulamis, A. D. Doulamis, and S. D. Kollias. A neural network approach to interactive content-based retrieval of video databases. *International Conference on Image Processing*, pages 116–120, 1999, vol. 2.

[43] W. S. Peng and N. DeClaris. Heuristic similarity measure characterization for content-based image retrieval. *IEEE conference on systems, man and cybernetics*, pages 7–12, 1997.

[44] G. Sheikholeslami, S. Chatterjee, and A. Zhang. Neuromerge: An approach for merging heterogeneous features in content-based image retrieval systems. *IEEE workshop on multimedia database management systems*, pages 106–113, 1998.

[45] T. P. Minka and R. W. Picard. Interactive learning using a 'society of models'. *Pattern recognition*, 30:565, 1997.

[46] Z. Yang and C. C. J. Kuo. A semantic classification and composite indexing approach to robust image retrieval. *International Conference on Image Processing*, pages 134 – 138, 1999, vol. 1.

[47] J. Demsar and F. Solina. Using machine learning for content-based image retrieving. *International Conference on Pattern Recognition*, pages 138–142, 1996, vol. 3.

[48] A. L. Ratan, O. Maron, W. E. L. Grimson, and T. Lozano-Perez. A framework for learning query concepts in image classification. *Proc. IEEE Computer vision and pattern recognition (CVPR)*, pages 423–429, 1999.

[49] S. Fischer, R. Lienhart, and W. Effelsberg. Automatic Recognition of Film Genres. In *ACM Multimedia 95 - Electronic Proceedings, San Francisco, California*, pages 295–304, November 1995.

[50] D. Sadlier and N. O'Connor. Event detection in field sports video using audio-visual features and a support vector machine. *IEEE Transactions on Circuits Systems and Video Technology*, pages 1225–1233, 2005.

[51] Y. Chen, J. Z. Wang, and R. Krovetz. Clue: cluster-based retrieval of images by unsupervised learning. *IEEE Transactions on Image Processing, vol.14*, pages 1187–1201, 2005.

[52] X. Zhu, X. Wu, A. K. Elmagarmid, Z. Feng, and L. Wu. Video data mining: Semantic indexing and event detection from the association perspective. *IEEE Trans. Knowl. Data Eng, vol. 15*, page 2005, 665-677.

[53] W. R. Neuman. *The future of the mass audience*. NY: Cambridge University press, 1991.

[54] N. Vasconcelos and A. Lipman. A bayesian framework for semantic content characterization. *Proc. of IEEE Conference on Computer vision and pattern recognition*, pages 566–561, 1998.

[55] B. Salt. *Film Style and Technology:History and Analysis*. Starwood, London. 2nd. Edition. 1992.

[56] D. Lowe. *Perceptual organization and visual recognition*. Kluwer Academic publishers, Boston, 1985.

[57] A. Jepson, W. Richards, and D. Knill. Modal structure and reliable inference. In *Perception as Bayesian Inference*, Cambridge University press, 1996.

[58] C. Fermuller and Y. Aloimonos. Vision and action. *Image and vision computing*, pages 725–44, vol. 13, no. 10, 1995.

[59] G. F. Kawin. *How Movies work*. Macmillan Publishing, New York. 1987.

[60] F. Nack and A. Parkes. The application of video semantics and theme representation in automated video editing. *Multimedia tools and applications*, pages 57–83, 1997.

[61] A. Mittal and L.-F. Cheong. Framework for synthesizing semantic-level indices. *Journal of Multimedia tools and application*, pages 135–158, 2003.

[62] S. Eickeler and S. Muller. Content-based video indexing of tv broadcast news using hidden markov models. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 2997–3000, 1999.

[63] V. Sobchack. Toward inhabited spce: The semiotic structure of camera movement in the cinema. *Semotica*, pages 317–335, 1982.