MPEG-7 Enhanced Ubi-Multimedia Access

-- Convergence of User Experience and Technology

Yong Rui
Microsoft Research
One Microsoft Way
Redmond, WA 98052, USA
yongrui@microsoft.com

ABSTRACT
Unlike early versions of MPEG, MPEG-7 is shifting from a data-compression standard to a multimedia content description standard. As a result, it will have much broader implications for the human-computer interaction research. There are countless ways that the MPEG-7 meta-data description framework can enhance users' experiences in accessing multimedia content. In this paper, we will report the algorithms that we developed for automatic generation of MPEG-7's description schemes (DSs): a video Table of Contents DS to enhance active video browsing, a Summary DS to enable direct use of a meta-data producer's annotation, and a Still Image DS to allow interactive content-based image retrieval. Experimental results and user studies demonstrated the usefulness of MPEG-7's meta-data framework and effectiveness of our proposed automatic DS generation techniques.

1. INTRODUCTION

With advances in computer technologies and the advent of the World-Wide Web, there has been an explosion in the amount and complexity of information being generated, stored, transmitted, analyzed, and accessed. Even though the information format and framework have changed drastically during the past decade, the way we access it has remained almost the same. We still watch a video clip passively, regardless of whether it is from a TV channel or a local disk. We still need to watch an on-demand baseball game video linearly, even when there are only 10 minutes of exciting segments in the 2-hour game. We still have to go through hundreds of art pictures one by one if we want to study the vase shapes in an African art collection.

More and more people are convinced that we need a new human-media interaction paradigm to improve users' experiences in accessing media content. The success of this paradigm will heavily depend on the collaboration and cooperation of multiple research communities, including human-computer interaction (HCI) and signal processing (SP). Fortunately, during recent years each community has seen the importance of the other. The HCI community not only studies the human-computer interfaces, cognitive models, their usability and functionality, but also looks into the required SP techniques to support these models/frameworks [3]. On the other hand, the SP community has moved beyond the algorithms for audio-visual feature extraction, compression, and distribution to meta-data description and user experience. We see more and more interdisciplinary research exploring both users' experiences and signal processing techniques [5,14]. For example, MPEG started MPEG-7 to address users' experiences in accessing multimedia content [8,9,10,11].

To explore the potential of HCI-SP synergy, we have developed algorithms and systems that look into how MPEG-7 could be used in future media accessing applications. The remainder of this paper is organized as follows. We first review the HCI community's effort on media access and then describe how MPEG moves from a data-compression standard to a meta-data description standard. We will discuss MPEG-7's goal, important terminology, and potential applications. We then describe in detail the techniques that we have developed under the MPEG-7 framework to enhance users' experiences in *active* video browsing, *direct* access to a meta-data producer's annotation, and *interactive* retrieval of images based on their visual content. We report experimental results and user studies that we have carried out. We conclude the paper with discussions and observations on future directions in ubiquitous media access.

2. RESEARCH ON ACCESSING MULTIMEDIA

HCI's effort in facilitating users' experiences in multimedia access can be traced back to the early 90s. In 1992, Mills et. al. proposed a hierarchical magnifier tool for video navigation [7]. This technique allows a user to recursively magnify the temporal resolution of a video source while preserving the levels of magnification in a spatial hierarchy. Degen et. al. proposed a tool for audio browsing, where a user can linearly and non-linearly access the audio content by using timelines, markers, time compression, and visual representation of the audio [4].

In 1993, Tonomura et. al. proposed to represent a video's temporal and spatial characteristics as icons for efficient accessing [15]. Ueda et. al. addressed the video representation issue from a different angle [16]. Their system offers descriptions of shot cut detection, motion estimation for both background and objects, and automatic linking for identical objects.

In 1998, we see another peak in studying users' experiences in accessing multimedia content. Mackay and Beaudouin-Lafon developed the DIVA system [6], which provides a smooth transition between spatial and temporal view of the data by mapping the source and presentation streams into a two-dimensional space. CMU's Informedia team reported their work on how to use multimedia information sources (audio, video, text, etc.) to generate video skims [3]. They reported two studies that measured the effects of different video skimming techniques on comprehension, navigation, and user satisfaction.

As discussed above, many creative ideas and concepts (e.g. structured video, non-linear access of audio-visual content, etc.) have been developed by HCI researchers during the past 10 years. While the research focus has been on interface design and user evaluation, the required techniques to support such an interface have not been fully explored. On the other hand, the SP community is addressing the multimedia content accessing issue from a different perspective. Many video processing techniques (e.g. shot boundary detection, key frame extraction, skim generation from prosodic audio feature analysis, etc.) have been developed. MPEG, the SP community's best known standard, also moved from low-level data compression to high-level meta-data description to enhance users' experiences in accessing multimedia content.

3. THE EVOLUTION OF MPEG

MPEG, standing for Moving Picture Experts Group, is the nickname of a working group inside the International Standard Organization (ISO/IEC) that drafts standards for coding audio-visual information in a digital compressed format [1,8,9,10,11]. It was formed in 1988 and has created 5 stages of MPEGs (see Table 1).

Today, with quick advances in data transmission and storage, the value of data (information) depends less on how much compression it achieves and more on how easily it can be found, retrieved, and accessed [10]. This shift of value makes it necessary to develop audio-visual information descriptions that go beyond the simple waveform or sample-based, frame-based (such as MPEG-1 and MPEG-2) or even object-based (such as MPEG-4) representations [9,10]. MPEG-21 [1] aims at media access across a variety of networks and devices. It is the newest member of the MPEG family, and is still evolving. MPEG-7 is formally called the "Multimedia Content Description Interface", and is the focus of this paper.

MPEG-7 standardizes a set of Descriptors (Ds) that can be used to describe various types of multimedia content. MPEG-7 will also standardize a set of Description Schemes (DSs) to specify the structure of the Ds and their relationship. These sets of Ds and DSs make it possible for content providers, meta-data annotators, and content consumers to speak the same multimedia description language. While a baseball game video is located in city A, a user can watch the game selectively in city B by using a MPEG-7 Summary DS created by a meta-data producer in city C.

Table 1. Evolution of MPEG

	Standard	Bit Rate	Techniques	Approval date
MPEG-1	Storage media	1.5Mbps	Frame-based compression	Nov. 1992
MPEG-2	Digital TV	6.0Mbps	Frame-based compression	Nov. 1994
MPEG-4	Multimedia application	Low bit rate	Object-based compression	Oct. 1998
MPEG-7	Multimedia content description	N/A	Not data, but meta-data "compression"	July, 2001
MPEG-21	Multimedia delivery & consump.	N/A	Accessibility across devices	2003

The following example illustrates some of the MPEG-7 terminology. A video clip is a piece of MPEG-7 data. It can have a color D and a motion D associated with it to describe a frame's color information and how an object is moving with time. It can also have a Summary DS that highlights the exciting portions of the video clip. These Ds and DSs are the meta-data of the data (the clip). They follow a certain MPEG-7 syntax to ensure interoperability and they can be stored separately from the data to enhance distributed operations. Here are some of the important DSs that will be studied in this paper:

- Video Table of Contents (ToC) DS: This DS is targeting at *active* video browsing by using a hierarchical tree-structured ToC.
- Summary DS: This DS describes a summary of audiovisual content. In the same way an abstract allows a reader to grasp the gist of a book, this DS conveys the essential information about the video content and provides immediate access to the necessary summary data [8].
- Still Image DS: This DS contains various visual Ds (color, texture, layout, etc.) and their relationships. It is the basis for *interactive* content-based image retrieval.

4. SCENARIOS, TECHNIQUES & EXPERIMENTS

A major limitation of the current media accessing experience is the lack of interaction and collaboration. With the MPEG-7 meta-data DSs, a user's ability to interact with the media and collaborate with other users can be greatly improved. To encourage competition, MPEG-7 will only standardize a set of DSs, leaving how to generate the DSs open for different vendors. The same was true for previous MPEGs. For example, MPEG-2, which is the supporting technique for DVD, only standardized the

syntax of compression so that a decoder can decode the incoming data, which left the encoder open for competition.

In this section, we will describe three scenarios to motivate the use of MPEG-7 DSs and explore MPEG-7's "encoder" techniques – automatic DS generation. The techniques explored here include a Video ToC DS to enhance *active* video browsing, a Summary DS to enable *direct* use of a meta-data producer's annotation, and a Still Image DS to allow *interactive* content-based image retrieval.

4.1. Enhanced Experience by Video ToC DS

In many cases, we need to browse through a video clip and quickly find a particular segment that is of interest. Imagine a TV commercial designer who needs to find a video segment depicting a romantic scene of two people chatting in front of a house. He remembers he saw such a scene in movie A, but cannot remember where exactly that scene is in the clip. If there is a ToC DS associated with the clip, he will be able to pinpoint that scene almost instantaneously, instead of doing the tedious "fast forward" and "rewind" for hours.

We will describe below an algorithm to automatically generate such a ToC DS for a movie video. Instead of using the chapter-section ToC structure as we have in a book, we use a more natural scene-group-shot structure for a video clip [13]:

- Shot boundary detection and key frame extraction: Use image processing and unsupervised learning algorithms to detect the shot boundaries and extraction key frames.
- Spatial-temporal feature extraction: Extract spatial features (e.g. color and texture) from key frames and extract temporal features (e.g. motion) from shots. These joint spatial-temporal features provide the basis for later grouping and clustering.
- Time-adaptive grouping: The similarity of shots can be modeled by localization in both space and time. "Local in space" means similar in visual features, and "local in time" means the shots are in vicinity in time axis. Our time-adaptive grouping takes both factors into account and merges similar shots into groups.
- Rule-based clustering for ToC construction: Based on the study of film-directing rules, we analyze the group patterns obtained in previous step to cluster semantically related groups into scenes.

The input to our algorithm is a raw video clip and the output of the algorithm is a tree structured ToC. This concise representation of video content provides a convenient tool for users to *actively* and *non-linearly* access the video content.

Experiments

Table 2. Toc DS results

Movie Name	Frames	Shots	Groups	Detected Scenes	False Positive	False Negative
Movie1	21717	133	16	5	0	0
Movie2	27951	186	25	7	0	1
Movie3	14293	86	12	6	1	1
Movie4	35817	195	28	10	1	2

Extensive experiments using real-world video clips have been carried out. To verify the robustness of the developed algorithm, various movie types are tested. Specifically, the test set includes Movie1 (romantic-slow), Movie2 (romantic-fast), Movie3 (music), and Movie4 (comedy). Each video clip is about 10-20 minutes long, and the total length of the test data is about 100,000 frames. The experimental results are summarized in Table 2, where "detected scenes" denotes the number of scenes detected by the algorithm; "false negatives" indicates the number of scenes missed by the algorithm; and "false positives" indicates the number of scenes detected by the algorithm but are not considered as scenes by humans.

This algorithm is among the first in the field to explore automatic ToC generation for movie videos [13]. The results are very encouraging in that they are close to a human's performance. By using the time-adaptive grouping, this algorithm overcomes the "windowing" effect encountered by most other existing techniques [13]. Another important advantage is its on-line processing ability. That is, the algorithm does not need to wait the entire data to arrive before it can process [13].

The active browsing process is illustrated as follows, taking the TV commercial designer's scenario as an example. In Figure 1, five scenes are automatically created from the 21717-frame video clip (Movie1). By looking at the representative frames, the commercial designer can quickly refresh his mind of where the romantic chatting scene is. To verify, he can also expand the video ToC into more detailed levels, such as groups and shots by clicking

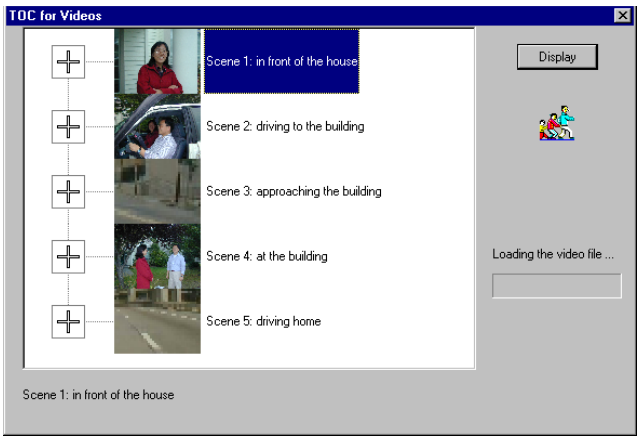

Figure 1. A condensed ToC

on the corresponding nodes (see Figure 2).

After confirming this is indeed the scene that he is looking for, he can *instantaneously* seek to the portion of interest by clicking on the "Display" button. A task requiring hours of work with a VCR, now can be done in minutes with a MPEG-7 ToC DS. The above *active* browsing process, enabled by MPEG-7 ToC DS, greatly facilitates users' access to video. It not only provides users with a non-linear access to video content (in contrast to conventional linear *fast-forward* and *rewind*), but also gives them a global context of where the scene is with respect to the whole story.

4.2. Enhanced Experience by Summary DS

Meta-data annotator is a new concept that has emerged in the past few years. The old "*content producer – content consumer*" paradigm is quickly being replaced by "*content producer – meta-data annotator – content consumer*" paradigm. A distinctive feature of MPEG-7 is its separation of meta-data from the data itself. This separation greatly enables a meta-data annotator to add its Ds or DSs to the MPEG-7 data. We have developed an algorithm for a meta-data annotator to automatically generate a Summary DS for a baseball game video. The game summary comprises the segments with high excitement.

A baseball fan comes home late because she had to trace a bug in her program. However, she needs to know the highlights of the Seattle Mariners games in order to chat with friends at the baseball chat club the next morning. It is already late at night and the recorded baseball game is two hours long. She is exhausted and wishes to look at a ten-minute summary instead of the whole game.

Equipped with MPEG-7's Summary DS, she can accomplish this. Her computer plays only the portion of the baseball game specified by the Summary DS from her favorite meta-data annotator. She successfully finishes the whole game in less than ten minutes without losing any major exciting segments. We next describe the algorithm that can automatically create such a Summary DS.

Features

In video delivery of a baseball game, the audio track provides a lot of useful information – the commentator's emotion is reflected in the way he or she speaks; the background audio from the audience is also a good indicator of the excitement of the game. Based on the above observations, we decided to use prosodic audio features as the basis for baseball game summary DS. Three types of audio features are used:

- *Pitch*: The waveform of voiced human speech is a quasi-periodic signal. The period in the signal is called the pitch of the speech. Independent of the waveform shape, this period can be shortened or enlarged as a result of the speaker's emotion and excitement level.

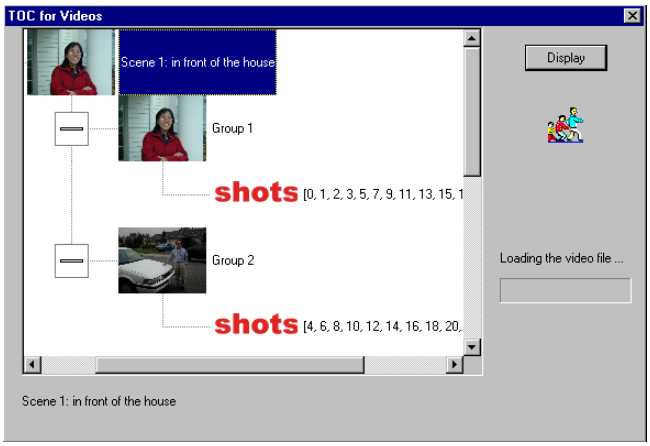

Figure 2. An expanded ToC

- *Intensity*: Audio intensity can be characterized by two features -- energy and magnitude. Energy (E) is defined as summation of the squared amplitude of the wave samples, while magnitude (M) is defined as summation of the absolute value of the wave samples. Because the squaring operation may over-emphasize a particular sample, M is preferred to E in characterizing signal intensity.
- *MFCC*: MFCC stands for Mel-frequency cepstrum coefficients. Cepstrum analysis is used extensively in signal processing. It is defined as the inverse Fourier transform of the log spectrum. Since the perception of human ears is not linear in frequency scale, Mel frequency scale, a log scale, is used to simulate this effect.

Model training and data classification

We propose to use a supervised classification algorithm. The first phase is model training. Both positive (exciting segments) and negative (non-exciting segments) examples are provided as training samples. This is a two-class training problem. The training samples are fit into two multi-dimensional *Gaussian* distributions. Let A and B be the two classes. Let (μ_A, σ_A) and (μ_B, σ_B) be the distribution parameters to be estimated for classes A and B. We obtain the distribution parameters by using Maximum Likelihood Estimation (MLE):

$$\mu = \frac{1}{n}\sum_{k=1}^{n} x_k \ , \ \sigma^2 = \frac{1}{n}\sum_{k=1}^{n}(x_k - \mu)$$

where x represents the training samples and n is the number of training samples.

The second phase is classification. Once the models are trained, they are capable of classifying the incoming video data into exciting segments and non-exciting segments.

The optimal *Bayesian* classification is used in our algorithm, where a class is selected if its *a posteriori* probability is the highest:

$$\arg\max_{i=A,B} P(i)p(x|(\mu_i,\sigma_i))$$

Experiments

We tested the proposed algorithm in both case-dependent classification and case-independent classification. In the formal case, the training samples and testing samples are from the same video clip, without overlapping. In the latter case, the training samples and testing samples are from different video clips. The results are summarized in Table 3.

Table 3. Automatic Summary DS

Type	Original Length	Summary Length	Detected Segments	False Positive	False Negative
CD	7108	282	14	0	0
CI	3980	96	5	0	1

CD stands for case-dependent test and CI for case-independent test. The video data lengths are in seconds. As can be seen from the table, the results for both CD and CI are with high accuracy. The high accuracy in CI means we can train the model off-line based on previous videos and then classify new videos *without* retraining the model. We can further see that the baseball game can be summarized in a very dense way, e.g., reduced by two orders of magnitude. This proposed algorithm provides a powerful tool for a meta-data annotator to generate MPEG-7 Summary DS more efficiently and with lower cost.

4.3. Enhanced Experience by Still Image DS

In the traditional image retrieval applications, the images are retrieved based on their associated text (e.g., title or author) [12]. But there are many applications that need to look at the image visual content (color pattern, texture style, or object arrangement) *directly* and retrieve images based on the visual features. This is especially true for art study.

To support this functionality, MPEG-7 has drafted a Still Image DS, which covers the visual features mentioned above. Not only will MPEG-7 standardize the visual Ds for color, texture, and shape, it can further associate *weights* to each of the Ds to capture a particular user's preference – some users may be interested in looking at the color usage of the paintings while others may be interested in studying various shapes of ancient vases. Based on MPEG-7 Still Image DS, we have developed a relevance feedback image retrieval system that actively integrates users' knowledge into the retrieval process and greatly increases the retrieval performance [14].

Algorithm

The Still Image DS consists of multiple visual Ds, i.e., DS = [D_1, D_2, …, D_i, …,D_I], where I is the number of Ds in this DS. For example, the Image DS can have a color histogram D, a wavelet texture D, and a Fourier descriptor D. Each D is normally a vector, consisting of multiple components (Cs), i.e., $D_i = [C_{i1}, C_{i2}, \ldots, C_{in}, \ldots, C_{iN}]$, where N is the number of components in D_i. For example, in the color histogram D, there are 32 components. To reflect different importance of each component to a D and each D to a DS, various weights can be associated to Cs and Ds. These two sets of weights are denoted as W_i and W_{in}.

Unlike most of today's image retrieval systems, our system takes full advantage of the weights (W_i and W_{in}) in the DS. During retrieval, users can provide positive and negative examples to guide the research. The retrieval system dynamically updates the weights based on users' relevance feedback to better estimate users' true retrieval purpose.

To update W_i, we have developed an estimation algorithm based on rank-difference [14]. Weights will be automatically increased for those Ds that capture the user's retrieval need. To update W_{in}, we have developed a standard deviation based algorithm [13]. Let's consider how to estimate C_{in}'s weight in D_i. For all the images that are marked with *relevant* by the user, stack their D_i's to form a M by N matrix, where M is the number of relevant images. Each column of the matrix is then a length-M sequence. Intuitively, if all the relevant images have consistent values for the component C_{in}, then this component captures the user's retrieval need. On the other hand, if the values for a component vary significantly among the relevant images, then this component does not capture the user's retrieval need. Based on this observation, the inverse of the standard deviation of the C_{in} sequence is a good estimation of the weight W_{in} for C_{in}.

Experiments

An African art image collection consisting of 286 images was used in the experiments. Once the Image DS and its Ds are fixed, the set of weights (W_i and W_{in}) uniquely specify user's retrieval need. We have tested how the proposed algorithm can drive the weights from initial equal weights

Table 4. Convergence ratio for image retrieval

Condition	0 fdbk	1 fdbk	2 fdbk	3 fdbk
1	40.8	89.6	97.4	98.5
2	38.9	95.7	98.8	99.0
3	39.0	77.7	74.2	77.3
4	34.8	91.9	94.6	94.1
5	62.9	85.7	87.1	87.1
6	64.1	87.6	94.4	95.3
Average	46.8	88.0	91.1	91.9

to the true weights in users' mind. We use *convergence ratio* (CR) to measure the retrieval performance. CR is defined as the number of retrieved relevant images with respect to the total number of relevant images. Table 4 summarizes the retrieval performance change of six sets of weights. The numbers in the table are CRs in percentage. After 3 rounds of relevance feedback, CR is increased to about 90%. Another desired feature of the algorithm is that the largest CR increase occurs in the first round of feedback. Figure 3 illustrates an art image retrieval result after a round of relevance feedback.

Subjects from both universities and industry research labs were invited to evaluate our system. The system was put at the UIUC engineering library for students to play with. Students from library and information sciences and art institute conducted questionnaire-based evaluation. Students liked this new image retrieval tool, as an alternative to the traditional text-based search engines. Art students especially liked the system's ability to support visual characteristic of the images, which provides them with more efficient tools for their study. Furthermore, the concept of *relevance feedback* has been rated very high by all the users, as it allows them to use their knowledge to guide the search.

CONCLUDING REMARKS

In this paper, we described how MPEG-7's meta-data can be used in accessing multimedia content and how automatic techniques can be developed to generate MPEG-7 DSs. The experimental results and preliminary user studies have demonstrated the usefulness of MPEG-7 and the potential of our proposed automatic techniques.

The three scenarios presented here are just our first attempt to address how to make multimedia content more accessible under the MPEG-7 framework. MPEG-7 poses many challenges as well as exciting opportunities for both the HCI and the SP communities. Users' experiences in accessing multimedia content will greatly benefit from interdisciplinary research. For future research, we are looking at MPEG-21 to ensure ubi-media interoperability across multiple networks and devices.

ACKNOWLEDGMENTS

The image collection was used with permission from the Fowler Museum of Cultural History at UCLA.

REFERENCES

1. Burnett, I., Van de Walle, R., Hill, K., Bormans, J. and Pereira, F., MPEG-21: goals and achievements, IEEE Multimedia, Oct.-Dec. 2003, pp.60-70
2. Cheesbrough, E., The ins and outs of MPEG, http://www.rcc.ryerson.ca/rta/brd038/papers/1996/mpeg1.html
3. Chistel, M. G., Smith, M., Taylor, C. R., and Winkler, D. B., Evolving video skims into useful multimedia abstractions, in *Proceedings of CHI'98 (Los Angeles, CA, 1998)*, 171-178.

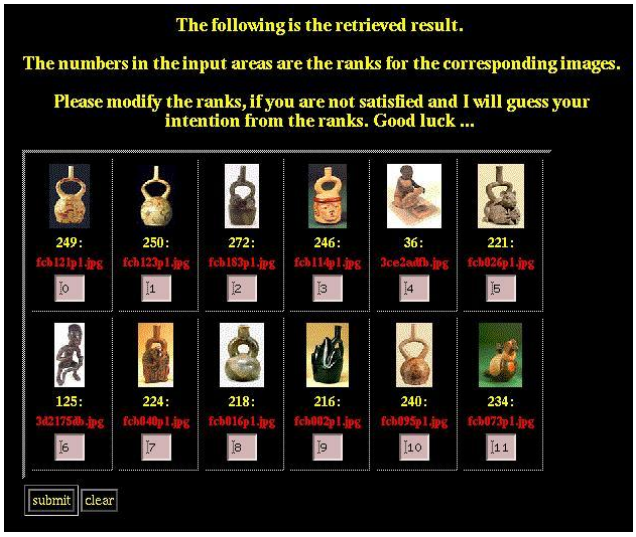

Figure 3. After relevance feedback

4. Degen, L., Mander, R., and Salomon, G., Working with audio: integrating personal tape recorders and desktop computers, in *Proceedings of CHI'92*, 1992, 413-418.
5. Faloutsos, C., Flickner, M., Niblack, W., Petkovic, D., Equitz, W., and R. Barber, Efficient and Effective Querying By Image Content, *IBM Research Report*, Aug., 1993
6. Mackay, W. E., and Beaudouin-Lafon, M., DIVA: Exploratory data analysis with multimedia streams, in *Proceedings of CHI'98 (Los Angeles, CA, 1998*), 416-423.
7. Mills, M., Cohen, J., and Wong, Y. Y., A magnifier tool for video data, in *Proceedings of CHI'92*, 1992, 93-98.
8. MPEG N2844, MPEG-7 Description Schemes (V0.5), ISO/IEC JTC1/SC29WG11, July, 1999, Vancouver, Canada
9. MPEG N2860, MPEG-7 Applications Document V.9, ISO/IEC JTC1/SC29WG11, July, 1999, Vancouver, Canada
10. MPEG N2861, MPEG-7: Context, Objectives band Technical Roadmap V.12, ISO/IEC JTC1/SC2WG11, July, 1999,
11. Rui, Y., Huang, T. S. and Chang, S.-F., Digital Image/Video Library and MPEG-7: standardization and Research Issues, In *Proceedings of IEEE ICASSP*, May, 1998, Seattle, WA
12. Rui, Y., Huang, T. S, and Chang, S.-F., Image Retrieval: Current Techniques, Promising Directions, and Open Issues, *Int. J. Vis. Commun. Image Rep*., 1999, Vol. 10, 1-23.
13. Rui, Y., Huang, T. S., Mehrotra, S., Constructing Table-of-Content for Videos, *Journal of ACM Multimedia Syst*, Sept, 1999.
14. Rui, Y., Huang, T. S., Ortega, M., and Mehrotra, S., Relevance Feedback: A Power Tool in Interactive Content-Based Image Retrieval, *IEEE Transaction on Circuits and Systems for Video Technology*, Vol.8, No. 5, 1998, 644-655.
15. Tonomura, Y., Akutsu, A., Otsuju, K., and Sadakata, T., VideoMAP and VideoSpaceIcon: tools for anatomizing video content, in *Proceedings of CHI'93*, 1993, ACM Press.
16. Ueda, H., Miyatake, T., Sumino, S., and Nagasaka, A., Automatic structure visualization for video editing, in *Proceedings of CHI'93*, 1993, ACM Press, 137-141