DIGITAL IMAGE/VIDEO LIBRARY AND MPEG-7: STANDARDIZATION AND RESEARCH ISSUES

Yong Rui and Thomas S. Huang

Dept. of ECE & Beckman Institute
University of Illinois at Urbana-Champaign
Urbana, IL 61801
{yrui,huang}@ifp.uiuc.edu

Shih-Fu Chang

Dept. of EE & New Media Technology Center
Columbia University
New York, NY 10027
sfchang@ee.columbia.edu

ABSTRACT

Much research activity and interest has emerged recently in two closely related areas: Digital Image/Video Library (DIVL) and MPEG-7. In this paper, we review the critical research issues in DIVL from a signal processing viewpoint, the objectives and scope of MPEG-7, and the relationships between these two.

1. INTRODUCTION

We are seeing the proliferation of Digital Image/Video Library (DIVL) in the multimedia information era. A huge amount of visual information is being produced at a rapid speed. Examples include those in both the military and civilian domains. Yet, technology for accessing, searching, and retrieving the visual information is still limited and inadequate.

The traditional text-based approach to accessing DIVL faces two major problems. One is the vast amount of labor required in the manual annotation process. The other, which is more essential, concerns the rich content in the images/videos and the subjectivity of human perception. Perceptual subjectivity and annotation impreciseness may cause unrecoverable mismatches in the retrieval process.

To overcome these problems, content-based retrieval was proposed in the early 90's. Instead of using manual, textual annotations, images/videos are indexed by their inherent visual features, such as color, texture, shape etc. Since then, many techniques in this research area have been developed and many image/video retrieval systems, both research and commercial, have been built. The active research effort has been reflected in many special issues of leading journals dedicated to this topic [2, 6].

Work of the first author was supported in part by a CSE Fellowship, College of Eng., UIUC; work of the second author was supported in part by ARL Cooperative Agreement No. DAAL01-96-2-0003; and work of the third author was supported in part by the National Science Foundation under a CAREER award (IRI-9501266), a STIMULATE award (IRI-96-19124), and industrial sponsors of Columbia's ADVENT project.

Much progress has been made and lessons have been learned. While techniques continue to advance, a standardized description of the multimedia content is urgently needed to solve the inter-operability problem over large-scale distributed DIVLs [9]. With a broader theme, MPEG has started a new work called MPEG-7, i.e. *Multimedia Content Description Interface*. Its objective is to specify a standard set of descriptors that can be used to describe various types of multimedia information [4].

In view of the importance of and synergy between these two areas, we present high-level discussion about the relationship between DIVL and MPEG-7, and their common research issues in this paper. Our objective is to motivate further discussion in the signal processing community to discover new technical barriers and identify promising directions.

The rest of the paper is organized as follows. Section 2 discusses the major unsolved issues in DIVL. Section 3 introduces MPEG-7 based on our preliminary knowledge. In section 4, we present our views towards relationships between DIVL and MPEG-7.

2. OPEN RESEARCH ISSUES IN DIVL

In the past few years, many advances have been made in various aspects of DIVL, including visual feature extraction, cataloging and organization, multi-dimensional indexing, testbed development, etc. [15]. However, there are still many open research issues that need to be solved before current DIVL systems can be of practical use.

2.1. Human in the Loop

Because of the perceptual subjectivity towards multimedia content, human is an indispensable part of the DIVL retrieval process. A user searches for information with different expectations depending on the context of the task, the nature of the image collection, and sometimes his/her own personal knowledge. Early DIVL systems emphasize fully automated techniques and try to find the best features and matching methods for image retrieval. However, this goal is only at-

tainable in constrained domains, such as medical or remote sensing. For generic systems, more recent research emphasizes interactive systems with human in the loop. Representative works include the FourEyes system [13] using learning through user interaction, Netra [12] incorporating supervised learning for texture analysis, WebSEEk [18] for dynamic feature vector recomputation, and MARS [16] using a relevance feedback framework for content-based retrieval.

2.2. High-level Concepts and Low-level Features

Except for specific domains, general users prefer to use high-level concepts in accessing information, including images and video. However, results of fully automatic image/video analysis usually remain as low-level features. There is a significant gap between the high-level semantic concepts and the low-level features. New research is called for in this area. Several dimensions provide promises, including incorporation of knowledge from user through learning and incorporation of semantic models from specific domains. The former includes supervised or unsupervised learning [13, 12]. The latter includes work in [11] for image classification using domain specific models. One promising alternative involves integration of multimedia and multiple modalities. Work in [19] uses image features and associated captions for automatic indexing of images of people.

2.3. High Dimensional Indexing

DIVL requires high-dimensional indexes for a large number of features. Typical features like color histogram, shape, and motion all produce high-dimensional feature vectors. Today, only a few systems have explored efficient multi-dimensional indexing techniques. However, as the size of DIVL increase rapidly, retrieval performance (e.g., computational complexity, retrieval accuracy and recall) has become more difficult to track. Existing indexing methods from the traditional database area have proven to be inadequate.

2.4. Integration of Disciplines and Media

Development of powerful DIVL systems requires interdisciplinary effort. Obviously, Image Processing/Computer Vision and Database Management play essential roles. Techniques integrating these two disciplines have demonstrated great benefits [14]. In addition, research from Information Retrieval [17] provide much intuition for discovering new issues in the visual domain, such as the initial approaches for the evaluation metrics and benchmark procedures. Another important discipline is user interface design. As we place more emphasis on human interaction in visual query, new efforts are needed to investigate effective methods for user to specify query criteria, navigate through the visual space, and provide feedback to the system. The goal is to overcome the dichotomy between the limited 2D user interface on today's computers and the rich content contained in the visual information.

Another observation is that integration of multimedia and multi-modalities provide great potential for improved indexing and classification of images in general domains. Research in [18] has shown promising results in using both textual and visual features in automatic indexing of images. More sophisticated techniques for cross-mapping image classification between the high level using textual cues and the low level using the visual cues will very likely bear fruit.

2.5. Metadata and Heterogeneity

Unlike the text documents, images or videos do not share consistent formats, indexes, or meta-data [9]. For example, dozens of formats are used for representing images and videos on the Web; many different techniques are used for implementing the indexing features in DIVL systems. In addition, there is no standard for inter-operability between different DIVL systems.

The issues of heterogeneity must be solved in order to improve inter-operability. Some standardization activities are underway. For example, Dublin Core has been augmented to extend the meta-data schemes from text documents to images [20]. A standard taxonomy for graphic materials has been proposed by the Library of Congress [1]. MPEG-7, aiming at the development of a standard for the description of multimedia content. This will be discussed in the following section.

3. MPEG-7

3.1. Scope of MPEG-7

Observing the increasing availability of digital audiovisual content, MPEG recently started a new work item, MPEG-7, formally called *Multimedia Content Description Interface*. The objective is to provide an interoperable solution to extend the capabilities of today's proprietary solutions in identifying multimedia content. That is, MPEG-7 will specify a standard set of descriptors that can be used to describe various types of multimedia information [4].

The following terminology is essential in MPEG-7 [7]:

- *Data*: is the audiovisual information that will be described using MPEG-7, regardless of the storage, coding, display, transmission, medium, or technology.
- *Description scheme (DS)*: defines a structure for the descriptors and their relationships.
- *Descriptors*: are the description of the data according to the structure defined in the DS.
- *Description*: is the entity describing the data and consisting of DS and descriptors.
- *Coded description*: is a compressed description allowing easy indexing, efficient storage and transmission.

A JPEG image or a MPEG-2 compressed video are examples of *data*. Color feature of an image is a *description scheme* and an image color histogram is a *descriptor* [5].

MPEG-7 will standardize the following: Description scheme, Descriptors, Coding methods for compact representation of Descriptions. Although not part of the standard, tools for creating and interpreting DS and Descriptors are also needed for testing and implementation purposes.

Figure 1 shows a high-level block diagram of a possible MPEG-7 processing chain [4], which includes *feature extraction*, *description*, and *search engine*. Even though both *feature extraction* and *search engine* are important to MPEG-7, they are not part of the standard. The main reason for this is that their standardization is not required to allow inter-operability, while leaving space for industry competition. This basically continues the spirit of MPEG-1 and MPEG-2, in which only the decoding methods are standardized. Another reason is to make use of the expected improvements in these two research fields.

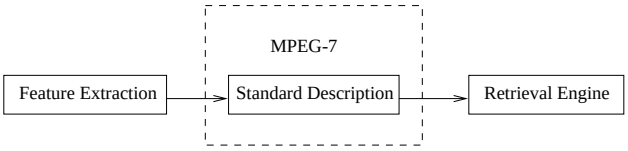

Figure 1: A possible MPEG-7 processing chain

3.2. MPEG-7's target applications

Example applications for MPEG-7 are quite broad and include [3]: visual retrieval systems (e.g., video databases, teleshopping, medical, and remote sensing applications), auditory retrieval systems (e.g., Karaoke and music sales and historical speech database), beyond-search applications (e.g., agent driven media selection and filtering, and intelligent multimedia presentation), and other applications such as Education or Surveillance.

4. RELATIONSHIP BETWEEN DIVL AND MPEG-7

As discussed in section 3.1, MPEG-7 only defines the *standard description* of multimedia content (the second block in Figure 1) while DIVL is more concerned with the *feature extraction* block and *retrieval engine* block. One may view that DIVL considers all issues in the entire system scope, while MPEG-7 focuses on the interoperability of internal representation of content descriptions. One may also view that MPEG-7 starts with more focus on audiovisual content, while DIVL involves many interactions with other disciplines such as document information retrieval. But as both fields are still in the early stage (particularly MPEG-7) and consist of much new input from new participants, the evolving relationships will certainly keep changing, but will also remain closely related.

There is great synergy between DIVL and MPEG-7. For example, even though tools for creating and interpreting DS and Descriptors are not part of the current scope of MPEG-7, they are indispensable for development of a successful *standard description*. An example in the opposite direction is that a well defined MPEG-7 standard will greatly benefit the inter-operability of DIVL retrieval systems, as we have discussed in Section 2.5.

4.1. Synergistic Data Model in DIVL and MPEG-7

Another aspect in considering the relationship between DIVL and MPEG-7 is to look at the hierarchical data model for images/video. In [16], an image object model was presented as follows,

$$O = O(D, F, R) \tag{1}$$

- D is the raw image data, e.g. a JPEG image.
- $F = \{f_i\}$ is a set of low-level visual features associated with the image object, such as color, texture, and shape.
- $R = \{r_{ij}\}$ is a set of representations for a given feature f_i , e.g. both color histogram and color moments are representations for the color feature.

In [8], an object-oriented video model was proposed to facilitate the content-based object-oriented video indexing and query, see Figure 2. A video clip may contain multiple salient video objects (e.g., foreground and background), and each video object may be decomposed into consistent video regions (e.g., uniform color regions) or classified into different semantic categories (e.g., people and automobiles). These hierarchical frameworks are synergistic with the video DS model being considered in MPEG-7 [7], as depicted in Figure 3. The hierarchy consists of video frame, shot, sequence, and higher level topic and knowledge about story.

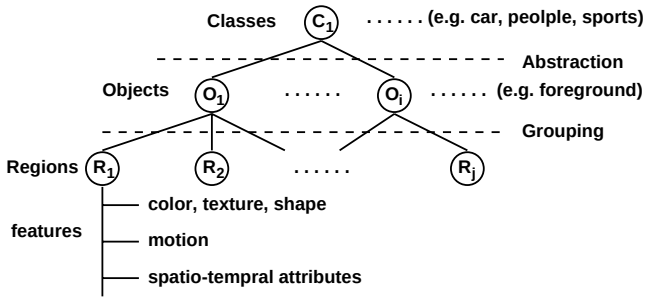

Figure 2: An Object-Oriented Video Model in VideoQ [8]

Using the context set by the above three data models, we can further understand the relationships between DIVL and MPEG-7. Research is both two areas

is called for to investigate efficient and effective features, matching criteria, and coding representations of features. Although we have said that the recent focus of DIVL has shifted from finding the optimal features to finding the best interactive mechanisms for visual query, pursuit of good features models and measures will still be an essential work for both DIVL and MPEG-7. In addition, efforts within MPEG-7 should be undertaken to study the best levels and set of content that should be standardized. We envision a flexible multi-level description scheme, including levels of region, object, frame, shot, scene, and story. Within each level, we need to determine what attributes should be indexed and which set should be indexed first. We expect to see much input from contributions provided by application developers and users in this area.

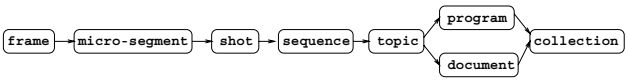

Figure 3: A Hierarchical Video DS in MPEG-7

4.2. Evaluation criteria

A general agreeable set of evaluation criteria and benchmarking procedures are essential for facilitating advances in any technical field, particularly for the fields that are emerging. Objective and subjective metrics focusing on the signal quality have been used in image/video compression. Precision and recall have been used in document-type information retrieval. However, for DIVL and MPEG-7, partly due to the human factor involved, there has not yet been a satisfactory solution. Some preliminary properties for good evaluation criteria have been discussed in MPEG-7 [5]: Accommodation of existing DSs, Expression efficiency, Effectiveness, Distinctiveness, Processing efficiency (amenability for fast processing), Storage-space efficiency, Scalability, and Flexibility. Contributions have been made in this area in the MPEG-7 community also [10].

Due to its actual, sometimes aggressive working schedules, we envision that MPEG-7, with enthusiastic involvement from industry and users, should be able to contribute to and achieve an effective working solution that is acceptable in practical applications. Similar contributions have been seen in MPEG-2 and MPEG-4, whose evaluation criteria has significant impact on later research in related areas.

5. REFERENCES

[1] The library of congress. *Thesaurus for Graphic Materials I.* http://lcweb.loc.gov/rr/print/tgm1.

[2] Special issue on content-based image retrieval systems. *IEEE Computer Magzine*, 28(9), 1995. Guest Editors: Venkat N. Gudivada and Jijay V. Raghavan.

[3] MPEG-7 applications document. *ISO/IEC JTC1/SC29/WG11 N1922, MPEG97*, Oct 1997.

[4] MPEG-7: Context and objectives (v.5). *ISO/IEC JTC1/SC29/WG11 N1920, MPEG97*, Oct 1997.

[5] MPEG-7 proposal package description (PPD) - v1.0. *ISO/IEC JTC1/SC29/WG11 N1923, MPEG97*, Oct 1997.

[6] Special issue on visual information management. *Communications of ACM*, Dec 1997. Guest Editor:Ramesh Jain.

[7] Third draft of MPEG-7 requirements. *ISO/IEC JTC1/SC29/WG11 N1921, MPEG97*, Oct 1997.

[8] S.-F. Chang, W. Chen, H.J. Meng, H. Sundaram, and D. Zhong. VideoQ - an automatic content-based video search system using visual cues. In *Proc ACM Multimedia 97*, 1997.

[9] S.-F. Chang, J. R. Smith, M. Beigi, and A. Benitez. Visual information retrieval from large distributed online repositories. *Communications of ACM, Special Issue on Visual Information Retrieval*, pages 12–20, Dec 1997.

[10] Pascal Faudemay. Benchmarking issues in the MPEG-7 process. *ISO/IEC JTC1/SC29/WG11 M2622, MPEG97*, Oct 1997.

[11] D. Forsyth and M. Fleck. Body plans. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 1997.

[12] B. S. Manjunath and W. Y. Ma. Image indexing using a texture dictionary. In *Proceedings of SPIE conference on Image Storage and Archiving System*, volume 2606.

[13] T. P. Minka and R. W. Picard. An image database browser that learns from user interaction. Technical report, MIT Media Laboratory and Modeling Group, 1996.

[14] Michael Ortega, Yong Rui, Kaushik Chakrabarti, Sharad Mehrotra, and Thomas S. Huang. Supporting similarity queries in MARS. In *Proc. of ACM Conf. on Multimedia*, 1997.

[15] Yong Rui, Thomas S. Huang, and Shih-Fu Chang. Image retrieval: Past, present, and future. *submitted to Journal of Visual Communication and Image Represenation.*

[16] Yong Rui, Thomas S. Huang, and Sharad Mehrotra. Relevance feedback techniques in interactive content-based image retrieval. In *Proc. of IS&T SPIE Storage and Retrieval of Images/Video Databases VI, EI'98*, 1998.

[17] Gerard Salton and Michael J. McGill. *Introduction to Modern Information Retrieval.* McGraw-Hill Book Company, 1983.

[18] J. R. Smith and S.-F. Chang. Visually searching the web for content. *IEEE Multimedia Magazine*, 4(3):12–20, Summer 1997. also Columbia U. CU/CTR Technical Report 459-96-25.

[19] R. K. Srihari. Automatic indexing and content-based retrieval of captioned images. *IEEE Computer Magzine*, 28(9), 1995.

[20] S. Weibel and E. Miller. Image description on the internet: A summary of the cni/oclc image metadata on the internet workshop, september 24 - 25, 1996, dublin, ohio. *D-Lib Magazine*, 1997.