MPEG-7 Content-Based Image Retrieval with Spatial and Temporal Information

Pei-Jeng KUO Terumasa AOKI and Hiroshi YASUDA

Yasuda-Aoki Laboratory, The University of Tokyo

E-mail: {peggykuo, aoki, yasuda}@mpeg.rcast.u-tokyo.ac.jp

Abstract: **The MPEG-7 standard was initiated with the intention to allow for efficient searching, indexing, filtering and accessing of multimedia content. While MPEG-7 provides a way to retrieve information of the image content, many of the MPEG-7 semantic descriptions still require manual annotations. As the rapid expansion of personal digital photographs in recent years requires a systematic indexing, categorizing and browsing interface to manage these enormous data, efficient metadata creation has become the bottleneck in building a MPEG-7 enabled multimedia database. We propose a MPEG-7 based multimedia content annotation scheme to generate MPEG-7 metadata with spatial and temporal information provided by modern GPS technology and a sophisticated location-based information database. The establishment of our proposed architecture would provide an interoperatable methodology for future multimedia content management applications at semantic level.**

1. INTRODUCTION

The rapid expansion of personal digital photographs would eventual make it impossible to retrieve in later days without a systematic indexing, categorizing, and browsing interface. The MPEG-7 standard, based on Extensible Markup Language (XML) schema, has therefore emerged with the intention of allowing for efficient searching, indexing, filtering, and accessing of audio-visual (AV) content. The birth of MPEG-7 aims to enable interoperability among devices and applications that manage the continuing growth of multimedia contents. [12] While MPEG-7 provides a way to get information about the AV data without the need of performing the actual decoding, many of the MPEG-7 semantic descriptions such as object, event, and location information still require a time-consuming manual annotation process. As a result, efficient metadata creation has become the bottleneck of building a MPEG-7 enabled multimedia database.

We propose a MPEG-7 based multimedia content annotation scheme to generate MPEG-7 metadata with spatial and temporal information provided by modern GPS technology and a sophisticated location-based information database. The establishment of our proposed architecture would provide an interoperatable methodology for future multimedia content management applications at semantic level. We also discuss the architecture and various application scenarios of the proposed system where the spatial and temporal information described with MPEG-7 metadata can bring users with a more customized image delivery service.

Section 2 summarizes our proposed approach with a brief comparison of previous related works on content-based image retrieval (CBIR) and provide a brief introduction of the MPEG-7 concept. Section 3 describes the general architecture of our proposed spatial and temporal based CBIR approach and the possible utilization models. Section 4 concludes this paper.

2. PROPOSED APPROACH WITH MPEG-7 SEMANTIC BASIC TOOLS

While immeasurable amount of multimedia information is accumulating in digital archives, from mobile terminals, on the web, in broadcast data stream and in personal and professional

database, the value of the information depends on how easily we can manage, find, retrieve, and filter it.

The birth of MPEG-7 aims to enable interoperability among devices and applications that manage the continuing growth of multimedia contents. [12] MPEG-7 descriptors are based on Extensible Markup Language (XML) and are designed to enable descriptions of both low level audio-visual (AV) features such as color, texture, motion and audio energy as well as high-level features of semantic objects, events and abstract concepts of digital multimedia contents. MPEG-7 provides a way to get information about the AV data without the need of performing the actual decoding. Many of the MPEG-7 descriptors are evolved from various signal and semantic level content based content retrieval researches. The following subsections summarize related signal and semantic level works on digital content management. The last subsection points out where our proposed schema differs from previous approaches and how we implement it with existing MPEG-7 tools.

2.1 Related Works on Signal-Based Content Retrieval

Most content based retrieval researches have focused on general-purpose multimedia database. Those retrieval systems roughly rely on the approach of extracting signal level features such as color histogram, color layout, and region-based signatures from the multimedia content. For example, the operation of Photobook [14] developed by MIT Media Lab is performed by comparing features associated with images, not the images themselves. The main features include color, texture, and shape, and parameter values of particular models are fitted to each image to evaluate their image similarity.

IBM QBIC project [15] is the first commercial content based image retrieval system, which allows for queries of large image databases based on visual image content properties such as the color percentages, color layout, and textures occurring in the images. Some systems use a combination of various algorithms and weight the retrieval results with sum matching metric or other merging schemes [11].

The Columbia VisualSEEK [18] and WebSEEK search engines support queries based on visual features such as color set or wavelet transform-based texture information. The spatial relationships between visual features are also considered and a user can then make queries based on color region sketches.

2.2 Related Works on Semantic-Based Content Retrieval

The Blobworld [17] developed at Berkeley presents a framework for image retrieval based on segmentalizing image into regions and querying images with those region properties. On top of the framework, statistical clustering models are introduced to associate annotated words with image regions for semantic image retrieval. It is an attractive alternative for semantic object recognition. However, different image features are suitable for the retrieval of images in different semantic types. To enable automatic semantic classification of images, the Stanford SIMPLicity [11] system proposed a semantics-sensitive approach for the general-purpose image database. It categorizes picture library into separate semantic types such as "graph", "photograph", "textured", "nontextured", "indoor", "outdoor", "city," or "landscape". A specific feature or a corresponding matching metric are then applied to image retrieval of each semantic classes in the system.

2.3 Related Works on Time and Event-Based Content Retrieval

In the Microsoft PhotoTOC[8] project, a time-based clustering algorithm is introduced to detect noticeable gaps between digital content creation times. This approach aims at digital personal photograph collections and helps users to automatically clustering their photos into albums for future browsing and retrieving. This approach is especially important for semi-automatic event detection and content clustering thanks to the time tag metadata attached to most current digital recording equipments.

2.4 Proposed Schema with MPEG-7 Tools

The MPEG-7 standard, formally named "Multimedia Content Description Interface", provides a rich set of standardized tools to describe multimedia content [10]. The MPEG-7 standard provides a metadata system, which describes the signal low-level AV content features such as color, texture, motion, audio energy as well as high-level features of semantic objects, event, content management related information and so forth. While most signal level descriptions can be extracted automatically, higher level attributes still require manual annotation afterwards [13].

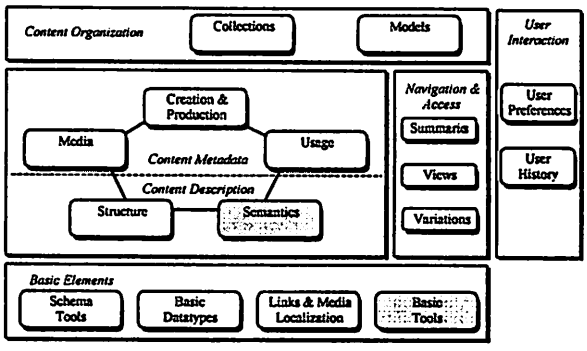

Figure 1 – MPEG-7 Multimedia Description Scheme (MDS) description tools overview

There are five parts of the MDS description tools. The basic elements on the bottom level of Figure 1 form the building blocks for higher level description tools. On the middle level in Figure 1, the content description tools describe the features of the multimedia content and the immutable metadata related to the multimedia content. The tools for navigation and access are shown as well at the middle level in Figure 1. Content organization tools shown at the top level describe collections and models of the multimedia content and the user interaction tools on the right part of Figure 1 contains user preferences as well as user history.

As marked on Figure 1, we propose a novel semantics description tool using the *TextAnnotation* datatype basic tool. The proposed semantics description tool can describe the "real-world" semantic features such as objects, events, and concepts that are related to or captured by the multimedia content. The *TextAnnotation* datatype of MPEG-7 MDS can contain multiple forms of an annotation including translations in multiple languages, or a combination of both structured and unstructured descriptions of the same annotation. We adopt the basic tool of *StructuredAnnotation* datatype, which is one of the several *TextAnnotation* datatype forms. The *StructuredAnnotation* datatype describes a structured textual annotation of multimedia contents in terms of who (people and animals), what object, what action, where (places), when (time), why, and how, which forms the main framework of our proposed description tool.

3. SPATIAL AND TEMPORAL BASED RETRIEVAL

3.1 Background

In addition to temporal information attached to digital images, spatial metadata annotations are made possible by the GPS technology. Although digital cameras with GPS modules have not become largely commercial available, mobile phones with camera modules and GPS functionality are emerging in Japan. This opens a potential utilization model of digital images produced by mobile devices.

Currently, most users store mobile images on memory sticks or online photo-album service sites. Occasionally, they share those images with friends or relatives via email or mobile-emails. In Japan, this is called "Sha-Mail" (email with photos).

With the more and more sophisticated camera modules and functionality available for mobile phones, CBIR services would be essential as a means of querying the ever-increasing image data in the long term. While MPEG-7 provides a way to get information about the AV data without the need of performing the actual decoding, many of the MPEG-7 semantic descriptions such as object, event, and location information still require manual annotations. As a result, efficient metadata creation has become the bottleneck of building a MPEG-7 enabled multimedia database.

3.2 Concept

We propose a novel methodology to facilitate semi-automatic MPEG-7 metadata generation for mobile image database, which makes use of the modern GPS technology, sophisticated location information database, map software as well as digital image devices (digital camera, camera equipped mobile phone, or PDAs). Figure. 2 shows a generic process of our concept. And Figure 3 shows a demonstrative illustration of our proposed semi-automatic annotation process.

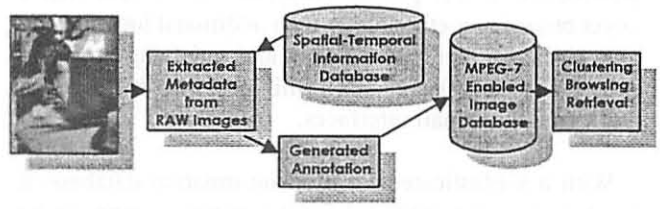

Figure 2 – Proposed Semi-Automatic Annotation Concept

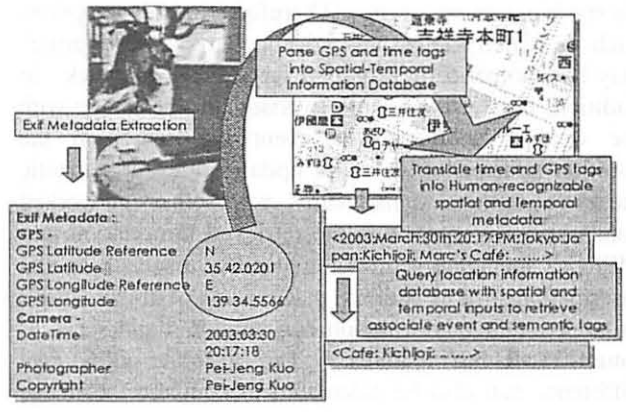

Figure 3 – Proposed Semi-Automatic Annotation Methodology

3.3 Spatial and Temporal Based Image Managment

The MPEG-7 Multimedia Description Schemes (MDS) consist of several structured parts. At the lower level, several basic tools of the MPEG-7 MDS basic elements are proposed [16]. Due to the length limitation of this paper, we only emphasize the StructuredAnnotation datatype of the TextAnnotation tool here. Detailed information of MPEG-7 MDS can be found in Refs. [6], [12] and [16]. The StructuredAnnotation datatype is designed to answer the questions "Who? What? Where? How? When?" which are the main parts of our proposed semi-automatically generated MPEG-7 metadata. With the generated answers of "Who? What? Where? How? When?", the next step is to index the image collections accordingly. In our proposed system, we integrate both visual and spatial features with temporal information. Therefore, clusters such as "Year 2003 Winter Scene of Inokashira Park at Tokyo" and "Year 2003 Spring Scene of Inokashira Park at Tokyo" can be indexed.

In Ref. [8], a browsing user interface with temportal order list of personal photograph collections is introduced. We do not emphasize browsing interface design with our proposed spatial and temporal information at this point. However, we believe that a novel browsing methodology with additional location and time clues can facilitate efficient and satisfying browsing experience for users and will perform better than traditional thumbnail interfaces.

With a sophisticated location information database, a number of useful attributes for the image content can be provided for our semi-automatic metadata creation check lists. For example, if a photograph was taken on April 5th at the Ueno Park in Tokyo, very likely it is related to the cherry blossoming event. Therefore, metadata options such as "Cherry Blossom", "Ueno Park", and "Spring" may be automatically provided for the user to check. In addition, it is also possible to associate the image with the weather condition, or event information if the location information database updates relevant semantic metadata options dynamically with other networked databases. By converting the retrieved information into MPEG-7 metadata semi-automatically, associate spatial information such as address, place, name of object, event, or even weather information can be stored and serve as future retrieving features. The absolute GPS data difference can also be calculated to compare the image similarity.

4. CONCLUSION AND FUTURE WORKS

We propose a MPEG-7 based multimedia content annotation scheme to generate MPEG-7 metadata with spatial and temporal information provided by modern GPS technology and a sophisticated location-based information database. The establishment of our proposed architecture would provide an interoperatable methodology for future multimedia content management applications at semantic level. We are currently building a prototype system with of the described methodology and the result and evaluation would be shown in our future publications.

REFERENCES

[1] N. Day, "Search and Browsing", *Introduction to MPEG-7 Multimedia Content Description Interface*, Ch20, John Wiley & Sons, Ltd, 2001.

[2] N. Day, S.Sekiguchi and M. Sasaki "Mobile Applications", *Introduction to MPEG-7 Multimedia Content Description Interface*, Ch21, John Wiley & Sons, Ltd, 2001.

[3]ISOIEC 15938-1, "Multimedia Content Description Interface – Part 1: Systems", 2001.

[4] Digital Library Project, U.C. Berkeley, http://elib.cs.berkeley.edu/.

[5] Digital Video and Multimedia Group, Columbia University, http://www.ctr.columbia.edu/dvmm/

[6] A. B. Benitez, H. Rising, C. Jörgensen, R. Leonardi, A. Bugatti, K. Hasida, R. Mehrotra, A. Murat Tekalp, A. Ekin, T. Walker, "Semantics of Multimedia in MPEG-7", *Proceedings of IEEE 2002 Conference on Image Processing (ICIP-2002)*, 2002.

[7] K. Rodden and K. Wood, "How do People Manage Their Digital Photographs?", *ACM Conference on Human Factors in Computing Systems (ACM CHI 2003)*, Apr 2003.

[8] J. C. Platt, M. Czerwinski and B. A. Field, "PhotoTOC: Automatic Clustering for Browsing Perosnal Photographs", *Microfort Research Technical Report*, Feb 2002.

[9] A. B. Benitez, and S. F. Change, "Perceptual Knowledge Construction from Annotated Image Collections", *Proceedings of the 2002 International Conference On Multimedia & Expo (ICME-2002)*, Aug 2002.

[10] ISO/IEC JTC1/ SC29/WG11 N4980, "MPEG-7 Overview", Jul 2001.

[11] J. Z. Wang, J. Li, and G. Wiederhold, "SIMPLIcity: Semantics-sensitive Integrated Matching for Picture LIbraries", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 9, pp. 947-963, 2001.

[12] ISO/IEC 15938-5:2001, "Multimedia Content Description Interface – Part 5 Multimedia Description Schemes," version 1.

[13] P. Salembier and J. Smith, "Overview of Multimedia Description Schemes and Schema Tools", *Introduction to MPEG-7 Multimedia Content Description Interface*, Ch6, John Wiley & Sons, Ltd, 2001.

[14] A. Pentland, R. Picard, and S. Sclaroff, "Photobook: Tools for Content-Based Manipulation of Image Databases", *SPIE Proceding*, Feb 1994.

[15] http://wwwqbic.almaden.ibm.com/

[16] ISO/IEC 1/SC 29/WG 11/N3964, "Multimedia Description Schemes XM", version 7.0, Mar 2001.

[17] C. Carson, M. Thomas, et al. "Blobworld: A System for Region-Based Image Indexing and Retrieval", *Proc. Visual Information Systems*, Jun 1999.

[18] J. R. Smith and S.-F. Chang, "VisualSEEk: a Fully Automated Content-Based Image Query System", *Proceedings, ACM Multimedia '96 Conference*, Nov 1996.