2004－AVM－44　(10)
2004／3／5

Personal Tourist Photograph Library: A MPEG-7 Ontology-Based MPEG-7 Image Indexing, Retrieving and Managing System

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 current trend of image retrieval is to incorporate the image visual features used in Content Based Image Retrieval (CBIR) and semantics annotations used in Metadata Based Image Retrieval to enhance retrieval performance. Because of the pervasive of consumer imaging devices, building personal digital photograph libraries became an increasingly interested domain. Personal digital photograph collections have specific characteristics compare to general purpose image databases. Hence, annotation architecture specially designed for that plays an important role in building an interoperatable data repository for future indexing, browsing and retrieving purposes. We propose a MPEG-7 based multimedia content description architecture, Dozen Dimensional Digital Content (DDDC), which annotates multimedia data with twelve main attributes regarding its semantic representation. In addition, we also proposed a machine-understandable "Spatial and Temporal Based Ontology" representation for the above DDDC semantics description to enable semi-automatic annotation process.**

Keywords:

Ontology, Digital Image Database, MPEG-7, Spatial-Temporal, image retrieval, metadata, Semantic Web, semi-automatic annotation

1. INTRODUCTION

The consumer behavior of taking photos has been changing rapidly since the introduction of digital cameras. During the past two to three years, while the performance of consumer digital cameras is approaching to traditional compact film cameras, the price of that has dropped rapidly. Currently, affordable digital camera and large capacity memory enables people to take many digital photographs whenever and wherever they want with very low cost. As a result, general users tend to produce a larger amount of digital photographs compare to the time when they used traditional film cameras.

In Japan, almost all the mobile phones on the market are equipped with camera modules at this time. This means, most people are bringing at least one camera with them all the time regardless of their personal preferences. The trend of using digital cameras, cameras in cellular telephones, and other pervasive devices, along with the increasing use of high-speed internet connections allow people to accumulate personal digital photographs faster than ever. As a result, finding suitable photographs for a particular purpose is increasingly problematic even for normal users.

While an increasing amount of people are building their online photo albums with the aid of off the shelf digital album tools as well as web album hosting sites, an effective and semantic way of retrieving context relevant images from the large repository of personal digital archives has yet appeared.

Two approaches have been studied in the research community:

1. Content-Based Image Retrieval (CBIR): CBIR research has been on-going for sometime. [14,15,17,18] Most of the Content-based approaches compare images based on their visual features such as color histogram, color layout, texture or shape. However, the retrieval precision has yet to be satisfactory.

2. Metadata-Based Image Retrieval: In Metadata-Based Image Retrieval, external metadata annotations such as keywords or free text descriptions are used when dealing with conceptually higher levels of content. [24]

In this paper, we focus on metadata-based image retrieval with an emphasis on management of personal photograph collections including novel indexing, clustering and retrieving with our proposed architecture.

Typically, individuals can publish their digital photographs online with a few key words annotated. Some users might choose some of their best shots among their digital repository and annotate those photos with semantic descriptions regarding to the context of those images.

However, it is very laborious, if not impossible, for consumers to annotation thousands of photographs they can easily capture with their digital devices. Therefore, mostly consumer digital photographs are either poorly annotated with a few keywords or are just stay with their numerical file names came along with the camera without any semantic annotations. [7]

Some commercial image providers, such as GettyImages [30] and National Geographic Image Collections [31], have invested time, and human resources to manually annotate keywords as well as relevant metadata descriptions for individual images among their collections. Current image search engines, such as Goggle image search [32], uses its text search engine to search based on an image's filename and surrounding text. However, preliminary metadata annotations are still the prerequisite for above mechanism to work.

Those approaches are not practically feasible for normal consumers as well as amateur photographers due to the limitation of time and human resources needed. However, without efficient retrieving and browsing algorithm, the ever increasing personal photograph collections would just like another photograph shoebox stored in the closet.

There are still some problems which hamper the development of "semantic" level image retrieval given the availability of carefully annotated external metadata [21, 22, 24, 26, 27]:

1. There is lack of common annotation architecture for personal digital image library. Several existing metadata initiatives [29] such as TV-Anytime (telco broadcasting), NewsML (multi-media news) , Dublin Core (simple resource discovery), CIDOC CRM (cultural heritage documentation) , INDECS (intellectual right management) , FGDC (Geographic Data) and GEM (Educational Material) have been constructed and used for various digital library purposes. Some of them have been adopted as the metadata scheme for special purpose image database such as museum collections. However, those were developed for different purposes and weren't well suited for extensive context-oriented annotation for personal digital photograph library.

2. Annotations require domain knowledge. Different annotator might use a different terms to annotate the same concept. In addition, the users who do not have specific domain knowledge might not be able to input the right keywords or natural language query for semantic image retrieval.

We try to tackle the above two problems with the following steps:

1. Construct common annotation architecture for building personal digital photograph libraries –We proposed The "Dozen Dimensional Ditigal Content (DDDC)" architecture extended from MPEG-7 Multimedia Description Scheme.

2. Construct a machine-understandable "Spatial and Temporal Based Ontology" representation for the above DDDC semantic description to enable semi-automatic annotation process.

We have proposed a semantic description tool of multimedia content [20] constructed with the *StructuredAnnotation* Basic Tool of MPEG-7 Multimedia Description Schemes (MDS). The proposed content description tool annotates multimedia data with twelve main attributes regarding its semantic representation. The twelve attributes include answers of who, what, when, where, why and how (5W1H) the digital content was produced as well as the respective direction, distance and duration (3D) information. We define digital multimedia contents including image, video and music embedded with the proposed semantic attributes as Dozen Dimensional Digital Content (DDDC).

In Section 2, we will describe the general concept and MPEG basic description tools we adopt to form the proposed MPEG-7 Multimedia Description Schemes (MDS) semantics description tools and the architecture of our proposed DDDC scheme. Section 3 provides a detailed explanation of our concept of building up the Spatial and Temporal Based Ontology with an example of personal tourist photograph library. Section 4 explains the proposed system architecture and summaries the annotation mechanism to conclude this paper.

Figure 1 Sample Image with free text annotation "2001_07_14_People and Eiffel Tower in Paris on the National Day of France"

2. DOZEN DIMENSIONAL DIGITAL CONTENT (DDDC) ARCHITECTURE

Extended from the *StructuredAnnotation* Basic Tool of MPEG-7 Multimedia Description Schemes (MDS), we propose a semantic description tool of multimedia content. The proposed content description tool annotates multimedia data with twelve main attributes regarding its semantic representation. The twelve attributes include answers of who, what, when, where, why and how (5W1H) the digital content was produced as well as the respective direction, distance and duration (3D) information. We define digital multimedia contents including image, video and music embedded with the proposed semantic attributes as Dozen Dimensional Digital Content (DDDC).

2.1 Annotate Multimedia Content with TextAnnotation Datatype

How should we annotate multimedia content using *TextAnnotation* datatype? Figure 1 is a sample image which was taken near The Eiffel Tower in Paris, France at 20:17 on the National Day of France at year 2001, which is the 14th of July. This image was annotated with free text "2001_07_14_People and Eiffel Tower in Paris on the National Day of France". Temporal and spatial information as well as the condition how this image was taken can be either manually inputted or retrieved from the original metadata provided by the recording equipment such as a GPS-equipped digital camera if available.

Figure 2 shows an example of the MPEG-7 *TextAnnotation* datatype. Two forms of *TextAnnotation* datatype are available, that are *FreeTextAnnotation* and *StructuredAnnotation.* The *TextAnnotation* part depicts the

```
<TextAnnotation id="Ann1">
    <FreeTextAnnotation xml:lang="en">
    2001_07_04_People and Eiffel Tower in Paris on the
    National Day of France
    </FreeTextAnnotation>
    <StructuredAnnotation>
        <Who>
            <ControlledTerm>
                <Name xml:lang="en">People
                </Name>
            </ControlledTerm>
        </Who>
        <What>
            <Name xml:lang="en">
                Eiffel Tower
            </Name>
        </What>
        <Where>
            <Name xml:lang="en">
                Paris; France
            </Name>
        </Where>
        <When>
            <Name xml:lang="en">
                2001-07-14T20:17+01:00;
                National Day of France
            </Name>
        </When>
    </StructuredAnnotation>
</TextAnnotation>
```

Figure 2 TextAnnotation example of Figure 4 with FreeTextAnnotation and StructuredAnnotation

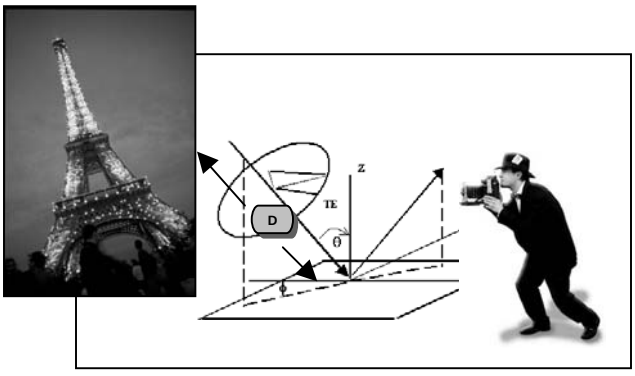

Figure 3 Concept of direction and distance information

original annotation and the *StructuredAnnotation* part annotates Figure 1 with specific tags of <who>, <what>,<where> and <when> attributes.

2.2 The Twelve Attributes

The above example shows a simple annotation for a digital image with the *TextAnnotation* datatype. More specifically, we propose a methodology to annotate multimedia content such as video, audio and images with twelve main attributes. The twelve attributes we proposed extend the StructuredAnnotation datatype semantics specified in [12] and include answers of who, what, when, where, why and how (5W1H) the digital content was produced as well as the respective direction, distance and duration (3D) information. Due to limited space, only brief explanation on the twelve attributes is given as following, and detailed explanation and example codes can be found in [20]:

Who

The *who* attribute describes animate objects or beings such as "people" and "animals" or "person groups" using Person Description Scheme (Person DS) or free text.

What

The w*hat* attribute describes inanimate object using either free text or a term from the classification scheme such as "The Eiffel Tower".

When

The *when* attribute describes the time point while the specific scene within the digital content happened. Free text or term from a classification schema such as "The National Day of France" can also be associated given specific relationship of the place of interests along with the time point information.

Where: Longitude

Where:Longitude attribute describes the spatial information of the digital content. Here we adopt the GeographicPoint Semantics specified in [12] and hence three attributes longitude, latitude and altitude are required to annotate the location where a specific digital content was taken.

Where: Latitude

The *where: latitude* attribute describes the latitude in degrees. Negative value represents southern latitude.

Where: Altitude

The *where: Altitude* attribute describes the altitude in meters. The reference altitude, indicated by zero, of the measurement is set to the sea level as default. Free text or term from a classification schema such as "Paris, France" can also be associated given specific relationship of the place of interests along with the GPS information. While lacking of GPS information, manual annotations might be needed.

Why

The *why* attribute describes the purpose that specific digital content such as audio, video or image was recorded.

How

The *how* attribute describes the device condition information while the specific digital content such as audio, video or image was recorded. This attribute can be described with free text or a combination other classification schemes. The information of how the specific digital content was recorded can be retrieved from the raw multimedia file available with most current digital recording devices.

Direction: Theta (θ)

The *direction: Theta (θ)* and d*irection: Phi (Φ)* annotations describe relative direction between the recording device and the recorded object. While the *where* attributes describe the GPS information recorded by respective recording devices, it can only specify the location of the device itself but not the object which was recorded in the digital file. The difference between the recording device and the object positions might be neglectable for image content such as person's portrait or street images. However, the real position of the object becomes ambiguous if the recorded object is a mountain far away from the camera or a star on the sky. In those cases, direction information between the recording device and the recorded object becomes important and can not be neglected. The direction vector from the photographer to the object is determined by two polar angles (theta – θ and phi - Φ) as shown in Figure 3.

Direction: Phi (Φ)

As explained above, the second polar angle d*irection: Phi (Φ)* attribute is required for the *direction* annotation.

Distance

To specify the real location of recorded object, the *distance* information between the recording device and recorded object

is also required. The distance between the photographer and the object is determined by the attribute of *distance* d (m) and it can be calculated based on the focal length information provided by most advanced digital recording devices. Concept if the *distance* attribute is also illustrated in Figure 6.

Duration

For multimedia content, especially audio and video, another attribute, *duration*, is also important when describing its semantic presentation. For audio and video files, the duration information can be retrieved from the starting and ending time tags and for image files, the shutter speed can serve as the duration attribute.

3. SPATIAL AND TEMPORAL BASED ONTOLOGY

In [21], several difficulties have been pointed out in terms of the annotation process. First, different annotator might use a different terms to annotate the same concept. Second, the users who do not have specific domain knowledge might not be able to input the right keywords or natural language query for semantic image retrieval. And third, the manual annotation of a large amount of personal digital photograph collections, if not impossible, is a laborious task.

In [22], the idea of Ontology-Based Photo Annotation was described. An ontology is a formal, explicit specification of a domain. Typically, an ontology consists of *concepts*, concept *properties*, and *relationships* between concepts. [26] Ontology concepts are represented by terms, which can hep the user in formulating the information needed, the query, and the answers [24]. While images in a content repository are annotated according to specific domain ontology, the same conceptualization can also offer to the users to facilitate focused image retrieval using the right terminology.

Figure 4 illustrate our proposed Spatial and Temporal Ontology. Our investigation on the experiment image database shows that there is a strong association between the image context and its respective spatial and temporal clues. Our observation on couple online personal photograph albums also shows that consumers tend to take more photographs on famous tourist stops. Based on this assumption, we propose to build locational specifc Domain Ontology for popular tourist stops such as the city of Paris, Tokyo and New York based on their respective spatial and temporal attributes. Illustrations in Figure 7 and 8 present example of spatial and temporal based hierarchical ontology for the city of Paris.

In building Spatial Ontology, we firstly separate Paris into several popular tourist districts such as “The Latin Quarter”, “The Eiffel Tower Quarter”, “Champs-Elysées” and “St-Germain des Prés”. Under each district, we again separate it into sub-districts or point of interests such as “Café de Flore”, “The Eiffel Tower” and “Café les Deux Magots”. Each node of the sub layer inherits the properties of their upper layers; therefore, when we annotate a photograph with “Café de Flore” metadata, upper layer properties of “St-Germain des Prés” and “Paris”, “France” would also be included.

The construction of Temporal Ontology requires more domain knowledge of the specific location. For example, the seasonal events periodically happen in the area, or special event occurs on specific date. As suggested in [27], there is no single correct class hierarchy for any given domain. And the ontology should not contain all the possible information about the domain but only specific enough for what you need in the application. We suggest building up the location specific Temporal Ontology according to the photographer’s personal interest and experience. In addition, we can also construct that with the aid of third party databases such as

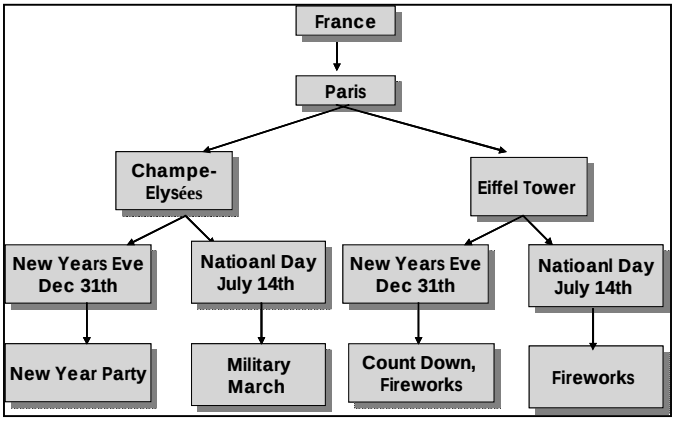

Figure 4 Concept of Proposed Temporal; Ontology

travel information portals or existing geographic metadata initiatives. In Figure 4, we demonstrate event tags come from our Temporal Ontology such as “New Years Party”, “Military March”, “Fireworks” and “Count Down”, which are associated with different image groups that were taken at the location of “Champs-Elysées” and “The Eiffel Tower” at special time such as “New Year’s Eve” or “National Day”.

4. CONCLUSION AND FUTURE WORKS

Interoperability standard such as MPEG-7 has been initiated in order to facilitate the collection of multimedia and related metadata. MPEG-7 standard emerged with the intention of allowing for efficient searching, indexing, filtering, and accessing of audio-visual (AV) content. However, the various attributes available for MPEG-7 can be as chaotic to the developers, not mention to the users.

The DDDC architecture we proposed annotates multimedia data with twelve main attributes regarding its semantic representation. In addition, we also proposed a machine-understandable "Spatial and Temporal Based Ontology" representation for the above DDDC semantics description to enable semi-automatic annotation process. As personal digital photograph libraries have specific characteristics and are particularly Spatial and Temporal associated, we envision various multimedia content management applications at semantic level can be developed based on the proposal described in this paper.

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.

[19] P. J Kuo, T. Aoki and H Yasuda, "Semi-Automatic MPEG-7 Metadata Generation of Mobile Images with Spatial and Temporal Information in Content-Based Image Retrieval", *Proceddings, 2003 International Conference on Software, Telecommunications and Computer Networks, Oct 2003.*

[20] P. J Kuo, T. Aoki and H Yasuda, "MPEG-7 Based Dozen Dimensional Digial Content Architecture for Semantic Image Retrieval Services", *Accepted Paper, 2004 IEEE International Conference on e-Technology, e-Commerce and e-Service (EEE-04).*

[21] V. W. Soo et al., "Automated Semantic Annotation and Retrieval Based on Sharable Ontology and Case-based Learning Techniques", *Proceedings, JCDL'2003, May 2003.*

[22] A. T. Schreiber et al., "Ontology-Based Photo Annotation", *IEEE Intelligent Systems*, May 2001.

[23] A. Graham et al., "Time as Essence for Photo Browsing Through Personal Digital Libraries", *Proceedings,JCDL'02*, Jul 2002.

[24] E. Hyvönen, A. Styman and S. Saarela, "Ontology-Based Image Retrieval", *Towards the semantic web and web services, Proceedings of XML Finland 2002 Conference*, 2002..

[25] A. Stent and A. Loui, "Using Event Segmentation to Improve Indexing of Consumer Photographs", *Proceedings, ACM SIGIR'01*, Sep 2001.

[26] A. Jaimes and J. R. Smith, "Semi-Automatic, Data Driven Construction of Multimedia Ontologies", *Proceedings, IEEE ICME 2003*, July 2003.

[27] N. F. Noy and D. L. McGuinness, "Ontology Development 101: A Guide to Creating Your First Ontology", *SMI Technical Report*, 2001.

[28] J. Hunter, "Enhancing the Semantic Interoperability of Multimedia Through a Core Ontology", *IEEE Transactions on Circuits and Systems for Video Technology*,Jan 2003.

[29] J. Hunter, "Adding Multimedia to the Semantic Web-Building an MPEG-7 Ontology", *SWWS, Stanford*, Jul 2001.

[30] http://www.gettyimages.com/

[31] https://www.ngsimages.com/