Dozen Dimensional Digital Content with MPEG-7

Pei-Jeng KUO　Terumasa AOKI　and　Hiroshi YASUDA

Yasuda-Aoki Laboratory, The University of Tokyo

Research Center for Advanced Science and Technology
4-6-1 Komaba, Meguroku, Tokyo, 153-8904 JAPAN

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

Abstract. We propose a MPEG-7 based multimedia content description tool which 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). The establishment of DDDC would provide an interoperatable methodology for multimedia content management applications at semantic level. We also envision novel algorithms associate with DDDC attributes for semantic based content retrieval in the future.

1 INTRODUCTION

In this paper, we propose a semantic description tool of multimedia content 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). Various multimedia content management applications at semantic level can be envisioned based on the description tool proposed here.

Section 2 summarizes the general concept and MPEG basic description tools we adopt to form the proposed MPEG-7 Multimedia Description Schemes (MDS) semantics description tools. Section 3 describes the architecture of our proposed DDDC scheme. Sample MPEG-7 markup codes would be provided as reference. Section 4 addresses novel utilization scenarios with the proposed mechanism, and Section 5 concludes this paper.

2 MPEG-7 SEMANTIC BASIC 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].

In MPEG-7 standard, the followings are specified:

- *Description Schemes (DS)* specify the structure and semantics of their components, which may be Description Schemes, Descriptor, or datatypes.

- *Descriptors (Ds)* describe multimedia features, attributes, or group of attributes.
- *Datatype* are the basic reusable datatypes used by DS and Ds.
- *System tools* support delivery of descriptions, multiplexing of descriptions with multimedia content, synchronization, and so forth. [12]

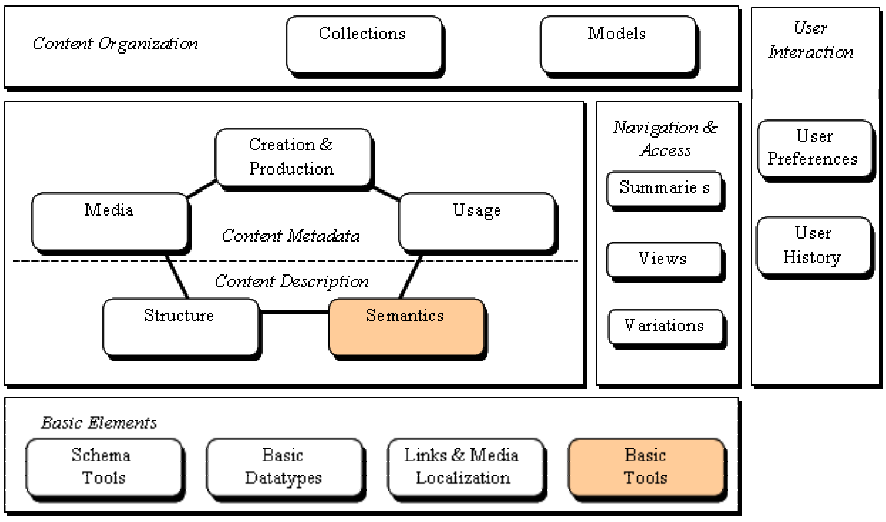

Figure 1 – MPEG-7 Multimedia Description Scheme (MDS) description tools overview

Here in this proposal, we focus on the Multimedia Description Schemes (MDS) part of the MPEG-7 standard. Figure 1 is an overview diagram of the MDS description tools including Description Schemes, Descriptors, and datatypes. 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 DOZEN DIMENSIONAL DIGITAL CONTENT (DDDC)

We propose a semantic description tool of multimedia content constructed with an extended 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). Semantics of each attribute with reference sample codes are described in the following subsections.

3.1 Annotate Multimedia Content with TextAnnotation Datatype

How should we annotate multimedia content using TextAnnotation datatype? Figure 2 is a sample image which was taken at Marc's Café in Kichijoji in Tokyo at 20:17 on the 30th of March 2003. This image was annotated with free text "Peggy Kuo is Reading a Book in Front of her VIAO Computer at Marc's Cafe in Kichijoji." Temporal and Spatial information as well as the condition how this image was taken can be retrieved from the original metadata provided by the recording equipment such as a GPS-equipped digital camera.

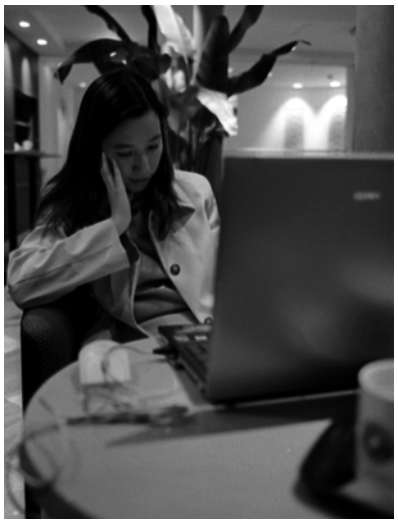

Figure 2 – Sample Image with free text annotation "Peggy Kuo is Reading a Book in Front of her VIAO Computer at Marc's Cafe in Kichijoji."

Figure 3 shows an example of the TextAnnotation datatype, it annotates Figure 2 which depicts the event of "Peggy Kuo is Reading a Book in Front of her VIAO Computer at Marc's Cafe in Kichijoji." In this example, two forms of expressions are given in describing the same event, where FreeTextAnnotation is simply an English description of what happened in the scene without any structuring and StructuredAnnotation identifies the who and whatObject as well as whatAction in a more structured annotation.

3.2 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. Explanation and sample metadata of Figure 2 annotated with each attribute are given in the following sections.

```
<TextAnnotation id="Ann1">
 <FreeTextAnnotation xml:lang="en">
          Peggy Kuo is Reading a Book in Front of her VIAO Computer at Marc's Cafe in Kichijoji.
 </FreeTextAnnotation>
 <StructuredAnnotation>
       <Who>
              <ControlledTerm>
                     <Name xml:lang="en">Peggy Kuo</Name>
              </ControlledTerm>
       </Who>
       <WhatObject>
              <Name xml:lang="en">
                     A book
              </Name>
       </WhatObject>
       <WhatObject>
              <Name xml:lang="en">
                     VIAO Computer
              </Name>
       </WhatObject
       <WhatAction>
              <Name xml:lang="en">
                     Peggy Kuo is Reading a Book in Front of her VIAO Computer at Marc's Cafe in
                     Kichijoji.
              </Name>
       </WhatAction>
  </StructuredAnnotation>
</TextAnnotation>
```

Figure 3- TextAnnotation example of Figure 2 with FreeTextAnnotation and StructuredAnnotation

```
<Mpeg7>
  <Description xsi:type="CharacterType">
    <CharacterInformation>
    <Character>
      <Agent xsi:type="PersonType">
        <Name xml:lang="en">
              <GivenName>Peggy</GivenName>
              <FamilyName>Kuo</FamilyName>
         </Name>
        <Name xml:lang="ja">
              <GivenName>ペギー</GivenName>
              <FamilyName>カク</FamilyName>
         </Name>
        <Affiliation>
              <Organization>
                     <Name>The University of Tokyo</Name>
              </Organization>
         </Affiliation>
        <ElectronicAddress>
              <Email>peggykuo@mpeg.rcast.u-tokyo.ac.jp</Email>
        </ElectronicAddress>
      </Agent>
    </Character>
    </CharacterInformation>
  </Description>
</Mpeg7>
```

Figure 4 - Example annotation of Figure 2 with the "Who" attribute

3.2.1 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. Figure 4 shows an example annotation of the "who" attribute. The animate object of "Peggy Kuo" in Figure 2 is annotated with both its English and Japanese names. Besides, the affiliation and email address of the person appeared in the scene are also annotated. In the PersonType semantics specified in MPEG-7 MDS, other attributes such as Person Groups and Address can also be described.

3.2.2 What

Figure 5 depicts the "What" attribute of the sample image with two semantics: WhatObject, which describes inanimate object using either free text or a term from the classification scheme such as a person or animal's name. WhatAction, which describes the actions occurred in the scene with either free text or a classification scheme term. Two objects, "A book" and "A VIAO Computer", and the action of "Peggy Kuo is Reading a Book in Front of her VIAO Computer at Marc's Cafe in Kichijoji." are depicted in the annotation example.

```
 <TextAnnotation id="Ann1">
<FreeTextAnnotation xml:lang="en">
            Peggy Kuo is Reading a Book in Front of her VIAO Computer at Marc's Cafe in Kichijoji.
</FreeTextAnnotation>
<StructuredAnnotation>
        <Who>
           <ControlledTerm>
             <Name xml:lang="en">Peggy Kuo</Name>
           </ControlledTerm>
        </Who>
        <WhatObject>
             <Name xml:lang="en">A book</Name>
        </WhatObject>
        <WhatObject>
             <Name xml:lang="en">VIAO Computer
             </Name>
        </WhatObject
        <WhatAction>
            <Name xml:lang="en">
                  Peggy Kuo is Reading a Book in Front of her VIAO Computer at Marc's Cafe in Kichijoji.
            </Name>
        </WhatAction>
</StructuredAnnotation>
 </TextAnnotation>
```

Figure 5 - Example annotation of Figure 2 with the "What" attribute

3.2.3 When

```
<Time>
       <TimePoint>2003-03-30T20:17+09:00</TimePoint>
</Time>
```

Figure 6 - Example annotation of Figure 2 with the "When" attribute

When attribute describes the time point while the specific scene within the digital content happened. We adopt the TimePointType semantics specified in [12] to describe the time. Since most current digital recording devices provide time code attached with the recorded multimedia content, we propose to use the retrieved time code from original recorded file as our media time. For digital content reproduced from audio, video or image contents recorded prior to the availability of digital time code, manual input of the time would be required. Figure 6 is a sample piece of metadata of the "when" attribute annotation.

3.2.4 Where: Longitude

Where 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. We envision most digital recording devices in the near future will equip with Global Positioning System (GPS) functionality and we define the retrieved GPS information from those devices as our device location and could serve as the input of where attribute specified here. For digital content reproduced from audio, video or image contents recorded prior to the availability of digital GPS information, manual annotation of the GPS data would be required. The Where: longitude attribute describes the longitude in degrees. Negative values represent western longitudes.

```
<GeographicPosition>
        <Point longitude="135.75" latitude="35.5" altitude="50"/>
</GeographicPosition>
```

Figure 7 - Example annotation of Figure 2 with the "Where" attribute

3.2.5 Where: Latitude

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

3.2.6 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. If the geodetic datum system used to measure the altitude is specified elsewhere, the datum point is determined by the system specification. Negative values indicate altitudes below the reference altitude. [12] Figure 7 is an example of the "Where" annotation using GPS information.

3.2.7 Why

The Why attribute describes the purpose that specific digital content such as audio, video or image was recorded. Figure 8 is an example of the "Why" attribute.

```
<Why>
    <Name xml:lang="en">
            A Person's Portrait
    </Name>
</Why>
```

Figure 8 - Example annotation of Figure 2 with the "Why" attribute

3.2.8 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. For example, most current digital cameras record various information such as focal length and ISO speed inside the stored multimedia files. That information can then be retrieved and serves as the "How" attribute when recording status is a concern for the potential digital data management system. Figure 9 is an example of the "How" attribute.

```
<How>
        <ISOSpeed>400</ISOSpeed>

</How>
```

Figure 9 - Example annotation of Figure 2 with the "How" attribute

3.2.9 Direction: Theta (θ)

The direction annotation describes 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 with 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. Figure 10 illustrates the concept of direction information.

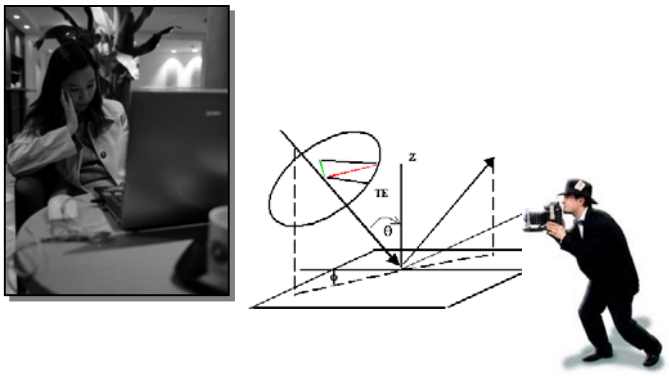

Figure 10 –Concept of direction information

A photographer who seated opposite to the object took the example photo of Figure 2. The direction vector from the photographer to the object is determined by two polar angles (theta – θ and phi - Φ) as in Figure 10.

3.2.10 Direction: Phi (Φ)

As explained above, two polar angle attributes are required for the "direction:" annotation. Figure 11 is an example of the "direction" attributes.

```
<Direction>
<Point theta=" 30 " phi="45.5" />
</Direction>>
```

Figure 11 - Example annotation of Figure 2 with the "Direction" attributes

3.2.11 Distance

To specify the real location of recorded object, the distance information between the recording device and recorded object is also required. A simple concept of distance attribute is shown in Figure 12. The example photo shown on Figure 12 was taken by a photographer who seated opposite to the object. 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. Figure 13 is an example of how to annotate the "distance" attribute.

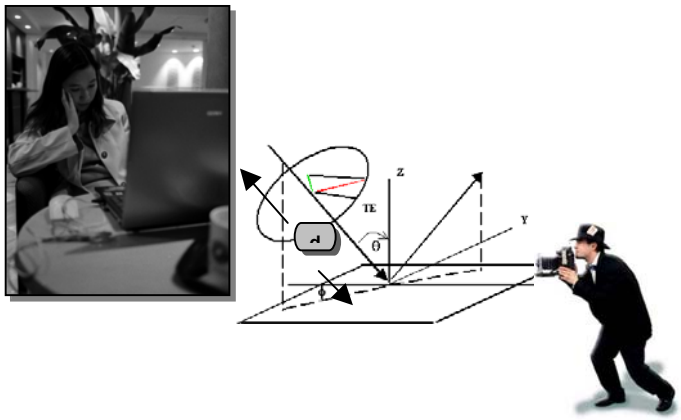

Figure 12 –Distance Attribute

```
<Distance>
        <Point distance=" 1.5 " />
</Distance>
```

Figure 13 - Example annotation of Figure 2 with the "Distance" attribute

3.2.12 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. Figure 14 is an example of how to annotate the "duration" attribute.

```
<Duration>
        <ShutterSpeed="125 " />
</Duration>
```

Figure 14 - Example annotation of Figure 2 with the "Duration" attribute

4 NOVEL UTILIZATION SCENARIOS

4.1 Indexing

The extended StructuredAnnotation datatype we proposed is designed to answer the questions of "Who? What? Where? How? When? How? (5W1H)" and " Direction, Distance, Duration (3D)" information of specific multimedia content. With the generated answers of proposed 5W1H3D attributes, multimedia content collections could be indexed accordingly. Figure 15 shows several indexing examples of visual, spatial, as well as spatial and temporal based clusterings of digital images made possible by our proposed DDDC metadata. We can cluster image collections into "Cat (with cat hair texture)", "Foliage (with a large portion of red color)" or "Sunset (with sunset color pattern)" with visual features. With Spatial features, photographs taken at different geographic locations can then be separated into clusters such as "Heian Jingu; Kyoto; Japan" and "Eiffel Tower; Paris; France". 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.

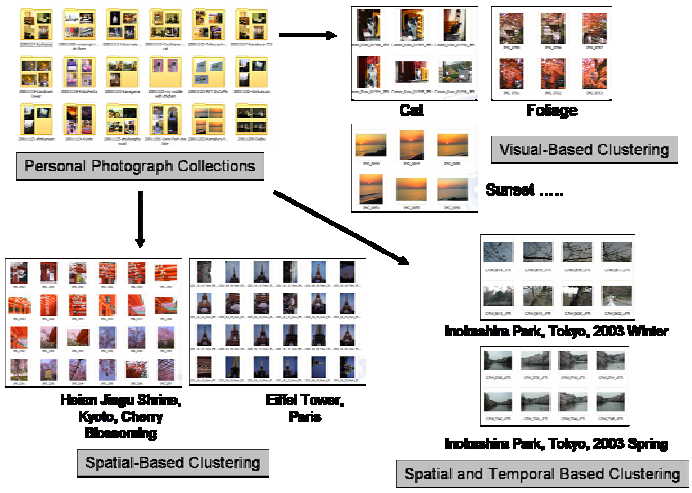

Figure 15 – Spatial and Temporal Based Indexing

4.2 Browsing

In Ref. [8], a browsing user interface with temporal 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 for multimedia contents.

4.3 Retrieving

With a sophisticated DDDC annotation, a number of useful attributes for the image content can be provided. 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 given a well organized location information database. 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 for our proposed DDDC metadata 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. A customized mobile image sharing and delivery platform can be designed based on the embedded spatial and temporal metadata in image database. An application scenario would be a mobile image reservation/sharing service system as well. There is a Chinese saying; many grains of sand piled up will make a pagoda. Currently, most people record digital contents such as image, video or audio just for personal pleasure. In the near future, when large anonymous users could provide their multimedia contents any time, any where, and on any event, the power of content aggregation can not be neglected. The novel utilization models of DDDC embedded multimedia data management system are anticipated to have a significant impact on the future content based retrieval scheme.

5 CONCLUSION AND FUTURE WORKS

We described a semantic description tool of multimedia content constructed and extended from the StructuredAnnotation Basic Tool of MPEG-7 Multimedia Description Schemes (MDS) in this paper. 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). Various multimedia content management applications at semantic level can be envisioned based on the description tool proposed here. In addition to building basic storage, access and distribution services, the more advanced data mining algorithms for efficient higher-level retrieval, recommendation, and utilization of the proposed DDDC still require further investigation. Building a prototype DDDC system utilizing the description tools described in this paper would be the next step of this study.

6 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.