A MPEG-7 Based Content-aware Album System for Consumer Photographs

Chen-Hsiu Huang, Chih-Hao Shen, Chun-Hsiang Huang and Ja-Ling Wu

Communication and Multimedia Laboratory,

National Taiwan University,

E-mail: {chenhsiu,shen,bh,wjl}@cmlab.csie.ntu.edu.tw

Abstract

In this paper, an MPEG-7 based album system for managing consumer-generated photographs is introduced. Face detection and recognition technologies are adopted to implement the query-by-face functionality of the content-aware album. A convenient face database construction and training mechanism is also provided. By utilizing both user attention model for visual objects and face information, the most representative region within a photograph, the so-called photo focus, can be easily extracted. Furthermore, with the aid of photo focus and objects importance calculated based on user attention models, two intelligent thumbnail creation approaches, superior to the commonly used sub-sampling method, are also proposed. At last, the face information is incorporated with MPEG-7 visual descriptors to facilitate the similarity calculation between images, allowing users to perform similarity search operations in our album system. The major functionalities introduced have contributed to form a content-aware album system for consumer-generated photographs.

Keywords: MPEG-7, a content-aware album system, query by face, smart thumbnail, photo similarity

I. Introduction

With the rapid development progress of information technology, devices used to create or digitalize photographs, such as digital cameras, have become increasingly affordable for end users. As a result, it is easier for consumers to record their daily life digitally, and the number of digital photographs dramatically increases, too. Users often feel frustrated because that, in the digital world, those contents we cannot manage or handle properly are of no values.

In order to deal with this problem, many digital album systems have been developed to help consumers manage their home photographs. Functions such as helping users to categorize their photographs by events or occasions are provided. Moreover, sorting or browsing photographs according to the date/time information extracted from the digital camera EXIF metadata [3] are also included. Some systems even ask users to write down textual description about photographs thus searching for specific photographs could be possible.

Everything sounds great, but the current solution is not satisfying yet. The reason is that traditional album systems still regard home photographs as meaningless bit streams. In our opinion, an ideal album system should be able to identify the difference between photographs and realize some semantic information about the content; that is, it should be a content-aware album system. Figure 1 shows the user interface of our current implementation.

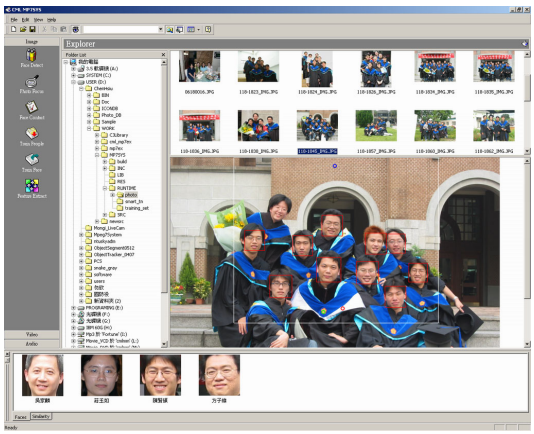

Figure 1. Snapshot of the MPEG-7 based content-aware album system

In our album system, we think that the most meaningful information in consumer home photographs is the presence of human face and thus the face detection and recognition technique [7] is adopted to realize the functionality of "query by face". Second, while viewing an image, there are always objects or scenes

attracting our attention. We call these eye-catching regions as "photo focus". After finding out the regions that catch our attention most by applying the user attention model [9], the album system can understand the digital contents better. Furthermore, base on the photo focus, a smart thumbnail-creation mechanism superior to the traditional one using image sub-sampling is proposed. At last, by combining several visual descriptors in the MPEG-7 standard [4] [5] and the face information, we can calculate the similarity between photographs. With photo similarities, relevant photographs can be automatically grouped together when users browse photographs in the album.

The paper is organized as follows. Section II describes the steps of using face detection and recognition technology to query photos by faces. In section III, the details of adopting user attention models for still images to find out the photo focus is explained, and then a new mechanism called "smart thumbnailing" used to create more informative thumbnail is also introduced. Section IV discusses the photo similarity calculation by using the combination of MPEG-7 visual descriptors and face information. Finally, section V gives some concluding remarks and suggestions for future work.

II. Query Photos By Face

Obviously, the most important semantic feature of home photographs is the human faces. People are always eager to see who is in the photographs. The techniques about locating and identifying faces in images have been developed for years, and it could be used as a more intuitive way for users to search and filter their photographs. In our system, we use the Open Source Computer Vision Library (OpenCV) [7] from Intel as our face detection and recognition component in order to extract the face information from photographs. The OpenCV library is well-known as an efficient and feasible face detection/recognition module.

First of all, all the faces appeared in consumer photographs shall be found out. With the help of the OpenCV library, most human faces can be easily found. Then, a face database will be updated so that new face images that represent a specific person are added. After adaptively training new pictures in the face database, people showed in photographs can be recognized. For each photograph, figures that can be recognized are saved as metadata along with the photograph. With the presence of face description, we can query photographs containing specific persons by instructions such as "find out the photographs that contain George and Mary".

The flowchart illustrating the steps of "query by face" operation is shown in Figure 2. We also developed a convenient interface to help end users build and update their face database easier. The screenshots are listed in Figure 3 and 4.

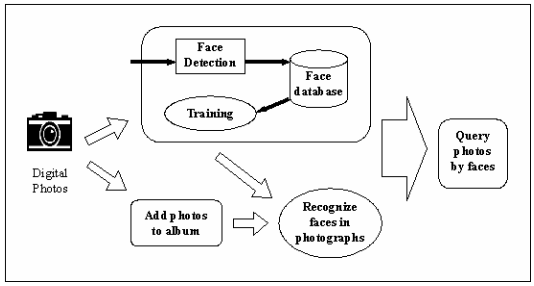

Figure 2. The steps for querying photographs by face

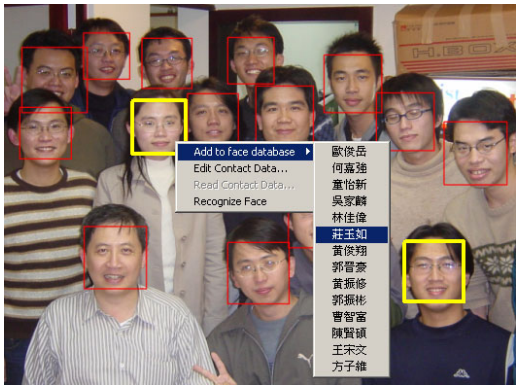

Figure 3. A simple face database updating interface is provided

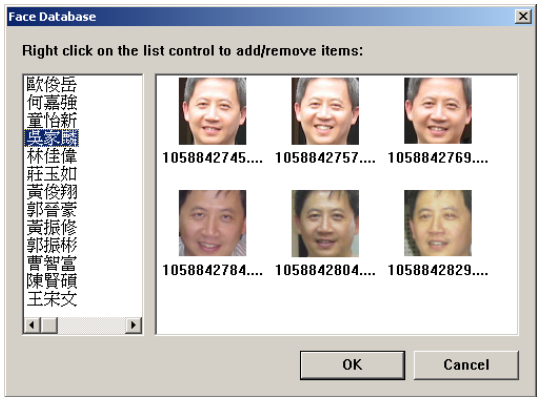

Figure 4. Face pictures belong to a certain user is added to the face database of our album system

III. Smart Thumbnails

Thumbnail representations of images are commonly used in image retrieval and browsing systems. Currently, the most basic way of

automatic thumbnail generation is through sub-sampling, i.e. re-sampling the original images to a smaller target size. This method, which we denote as direct sub-sampling, is simple and adopted widely in current album applications. However, information contained in thumbnails generated in this way is often difficult to be identified and recognized because much information is lost during the sub-sampling process, especially after resizing high-resolution photographs.

Another important reason to find better thumbnails is about image adaptation. Rapid development of information technological structures has contributed to the prevalence of mobile computing devices, such as PDAs and SmartPhones. Users must convert their home photographs to a smaller size first so that they can share these pictures with their friends and families by using mobile devices. Due to the limited display size and computing power, resizing the original photographs to meet the target display resolution is inevitable. Under this circumstance, direct sub-sampling is no longer appropriate for creating thumbnail image because it is often not representative. In order to conquer this problem, we need to identify what is the most informative region in a photograph, and then preserve more information in this region when creating thumbnails.

A. Photo Focus

The term "attention" refers to the ability of one human to focus and concentrate upon some visual or auditory 'object', by careful observing or listening. Visual attention is the ability of biological visual system to detect interesting parts of the visual input [6]. There are mainly two kinds of models in the literature of visual attention and saliency map calculation: top-down and bottom-up methods [8]. Roughly speaking, bottom-up attention models what people are attracted to see, and top-down attention models what people are willing to see.

Itti and Koch [6] have proposed a bottom-up way to compute the saliency map, based on low level features such as color, intensity and orientation. The saliency map is a matrix corresponding to the input image and describes the degree of saliency of each position in the input image. In our implementation, we borrow the low level features of color and intensity from [6] and since the human faces are the most significant feature in consumer home photographs, the face information is also adopted as a high level feature.

In the process of visual attention analysis, we denote the detected face rectangle as attention objects, and use the color and intensity features as saliency points. Besides, since the obtained face detection results still has miss detection, we adopt another low level feature, skin color [2], in order to compensate this problem. By combining the mentioned attention objects and saliency points, a rectangle called focus rectangle is formed as the photo focus.

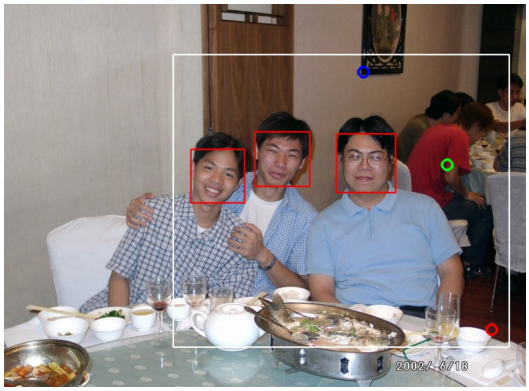

Figure 5. Attention objects and saliency points are combined to form the photo focus

Figure 5 show the example of photo focus computed from those visual attention features. The red circle represents saliency point based on intensity saliency map. The green circle indicates the saliency point calculated using the colour feature, and the blue one is for skin colour regions. The presence of photo focus makes the album system understand more about contents processed and can be further used as the basis to extract more semantic information from photographs. For example, our smart thumbnail creation, which will be discussed soon, is based on the photo focus. Besides, the photo focus region can also be used for Region-Of-Interest (ROI) coding scheme and Spatial/SNR scalability in JPEG 2000 [1] standard.

B. Smart Thumbnails

Two different approaches are proposed to create smart thumbnail base on the photo focus mentioned above. Rather than resizing the original picture to the target size, thumbnails created by clipping the most important region from a photograph would be more informative. Figure 6 illustrates the effect and difference between the traditional thumbnail and our smart thumbnail approaches. The two approaches describe as follows:

Focus-based smart thumbnail: The focus based approach is mainly based on the calculated photo focus region, and is regarded as a static approach. The focus region is cropped first and then shrinking to the target size. In our experiments, the focus-based approach gets pretty well results in most cases. But if there're

too many people on the photograph or the saliency points appeared in regions near image borders, the focus region may be almost as large as the whole photograph size, thus making no difference in comparison to the directly sub-sampled thumbnails. In order to deal with this situation, another method is proposed.

Adaptive selection based smart thumbnail: This method can be regards as the dynamic selection of photo focus. Through our observation, users tend to lay their focus on the center of photograph when they take pictures. Besides, people also use to be together in the same surface plane while take photographs. In other words, most of the detected face regions shall be of the same size.

Upon the two facts observed, we know that for human faces detected around the central of photograph are semantically more important than those appeared near the photo edge. Furthermore, people standing in front of others are often more important (with larger face region). Thus for each attention objects, an importance weighting function is calculated by:

$$W_i = (FR_w \times FR_h)^2 / d_c \qquad (1)$$

Where W_i denotes the calculated importance, FR_w and FR_h is the width and height of detected face region, d_c is the distance between photograph central and face rectangle central. Saliency points are also considered in importance calculation.

We also define the photo focus ratio, Fo_r, which is the ratio between focus region and the whole photograph. Users are allowed to dynamically adjust the photo focus ratio Fo_r. During each iteration, attention objects are sorted according to their importance value to see if the current selection of focus region meets the photo focus ratio Fo_r. If not, the attention object with the lowest importance value will be dropped to decrease the focus region, trying to meet the user-specified photo focus ratio. Finally, when the iteration process stops, people standing near the border of picture will usually be dropped and those located in front or around the central of picture will be preserved.

Figure 6 (c) (d) shows the resulting thumbnail of adaptive selection approach. In this method, users are allowed to choose whether more information or more details are desired as their preference by dynamic selecting photo focus ratio Fo_r.

(a) direct sub-sampling

(b) focus based smart thumbnail

(c) adaptive selection for $Fo_r = 15\%$

(d) adaptive selection for $Fo_r = 10\%$

Figure 6. Different thumbnails created are listed. The photograph with an original size of 2048x1536 is resized to a 320x320 resolution to be properly display on PDA screen.

IV. Photo Similarity

The amount of multimedia contents is explosively growing day by day. How to perform effective and efficient multimedia indexing, searching and retrieval is a long-standing and active research filed. Many tools and systems had been developed for different applications. However, users do want to access heterogeneous content sources in a transparent way, and the need of an interoperable content description scheme has been considered. In 1997, the MPEG organization created a new working item, Multimedia Content Description Interface, to address such a need, and the consolidated result is the MPEG-7 standard [4].

MPEG-7 standardized a set of visual and audio descriptors as representations of features of multimedia data. Two descriptors, Color Layout and Dominant Color, from MPEG-7 visual part [4], are adopted and combined with the number of faces detected in photographs as the basis of photo similarity calculation. Color Layout descriptor effectively represents the spatial distribution of color of visual signals in a very compact form. Dominant Color describes the most suitable colors for representing local (object or image region) features where only a small number of colors are allowed to characterize the color information.

The distance of Color Layout and Dominant Color description, denoted as $dist_{CLD}$ and $dist_{DCD}$ are describe in MPEG-7 visual part. As for the distance of face number description between photographs, it is defined as follows:

$$dist_{FND} = \frac{|FN_i - FN_j|}{\max(FN_i, FN_j)} \quad (2)$$

$dist_{FND}$ denotes the distance of face number descriptor and FN_i, FN_j represents the face numbers detected from photograph i and j. At last, the photo similarity between photo i and j is defined as the average of three description distance:

$$Sim_{ij} = (dist_{CLD} + dist_{DCD} + dist_{FND})/3 \quad (3)$$

The Color Layout and Dominant Color descriptor are simple but powerful. By adopting the face number descriptor, we can easily find out similar photographs with specified number of persons and separate those photographs with people from scenery images without people. Figure 7 shows the photo similarity search result in our album system.

V. Conclusions and Future Work

Three major functionalities: query by face, smart thumbnails, and photo similarity introduced in this paper constitute our content-aware album system for consumer photographs. Future works can be roughly classified into two directions: system aspect and component aspect. In the system aspect, since MPEG-7 has defined the standard multimedia description interface, the album document file format should conform to the MPEG-7 description syntax (currently the album documents are saved in our own XML format for the programming convenience). In the component aspect, more low level features or descriptors in MPEG-7 standard will be used and combined for further semantic meaning extraction. Moreover, the face detection and recognition library could be improved to meet the needs of album applications for figures identification. Besides, features such as texture and edge information in photographs will also be included for creating saliency maps so that a more accurate user attention model of viewing photographs could be obtained.

VI. References

[1] Charilaos Christopoulos, Athanassios Skodras, and Touradj Ebrahimi, "The JPEG2000 still image coding system: An Overview," *IEEE Trans. Consumer Electronics*, vol.46 no.4, pp. 1103-1127, Nov. 2000

[2] Christophe Garcia and Georgios Tziritas, "Face Detection Using Quantized Skin Color Regions Merging and Wavelet Packet Analysis," *IEEE Trans. Multimedia*, vol. 1, no. 3, Sep. 1999.

[3] EXIF.org - EXIF and related resources, http://www.exif.org/specifications.html

[4] Information Technology – Multimedia Content Description Interface, ISO/IEC International Standard 15938-(1-8), 2002.

[5] ISO/IEC JTC1/SC29/WG11/N4980, MPEG-7 Overview, July 2002.

[6] L. Itti, C. Koch, and E. Niebur, "A Model of Saliency-based Visual Attention for Rapid Scene Analysis," *IEEE Trans. Pattern Analysis and Machine Intelligence*, 20(11), pp. 1254-1259, 1998.

[7] Open Source Computer Vision Library, http://www.intel.com/research/mrl/research/opencv/

[8] Ruggero Milanese, Harry Wechsler, Sylvia Gil, Jean-Mare Bost, Thierry Pun,

"Integration of Bottom-up and Top-down Cues for Visual Attention Using Non-linear Relaxation," *Proc of CVPR* 1994, p781-785.

[9] Yu-Fei Ma, Lie Lu, Hong-Jiang Zhang and Mingjing Li, "A User Attention Model for Video Summarization," *ACM Multimedia*, Dec. 2002

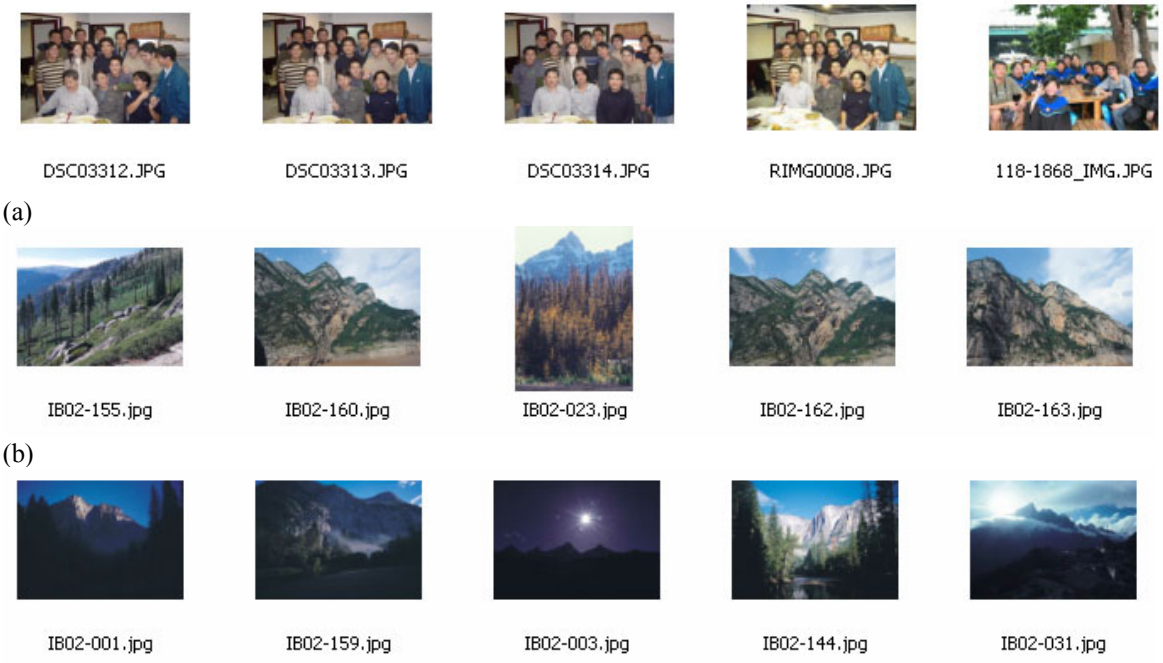

Figure 7. Similar (a) pictures containing people and (b) (c) scenery pictures can be easily and correctly found by the proposed image grouping functionality