Incorporating Camera Metadata for Attended Region Detection and Consumer Photo Classification

Zhong Li
Dept of Computer Science
UNC-Charlotte
Charlotte, NC 28223, USA
zhli@uncc.edu

Hangzai Luo
Software Engineering Institute
East China Normal University
Shanghai, China
hzluo@sei.ecnu.edu.cn

Jianping Fan
Dept of Computer Science
UNC-Charlotte
Charlotte, NC 28223, USA
jfan@uncc.edu

ABSTRACT
Photos taken by human beings significantly differ from the pictures that are taken by a surveillance camera or a vision sensor on a robot, e.g., human beings may intentionally capture photos to express his/her feeling or record a memorial scene. Such a creative photo capture process is accomplished by adjusting two factors: (1) the parameters setting of a camera; and (2) the position between the camera and the interesting objects or scenes. To enable automatic understanding and interpretation of the semantics of photos, it is very important to take all these factors into account. Unfortunately, most existing algorithms for image understanding focus on only the image contents while completely ignoring these two important factors. In this paper, we have developed a new algorithm to calculate what the interestingness of the photographer is and what the core content of a photo is. The gained information (i.e., attended regions and attention of the photographer) is further used to support more effective photo classification and retrieval. Our experiments on 70,000+ photos taken by 200+ different models of cameras have obtained very positive results.

Keywords
Camera metadata, attended regions, image classification.

1. INTRODUCTION
The consumer photos, which are created by human beings, are driven by the photographer’s intention. Such the consumer photos are used to express the photographers’ emotion and feeling and record the memorial scene of objects. Such the human-taken photos significantly differ from the pictures that are taken by a fixed surveillance camera or a vision sensor of a mobile robot: photographers follow the basic rules of esthetic sentiment that shared among the majority of civilized human beings to make the photos attractive, beautiful or even astonishing in all circumstances even in tragedy. For example, people often save certain amount of space for the sky in an outdoor photo to make it looks balanced. The interesting subjects are often placed in the middle of the picture and keep it integrated. If environmental luminance is not sufficient to make a clear picture, photographers may use a flash and they may make sure that the interesting subjects take proper exposure while the surroundings are neglectable.

The human intention is berried under the pixels we can see from pictures and the camera metadata we can read from head files. In order to enable automatic image understanding, it is very important to develop new frameworks by integrating the visual contents of the images and the camera metadata. Unfortunately, most existing techniques for image understanding have completely ignored the camera metadata while focusing on only the visual contents of the images. Recently, some pioneering researches have been done to simply integrate the camera metadata for image classification [6-7].

From the above interpretations, one can observe two important issues [2-3]: (1) not all pixels are born equal and some of them are more important than the others in a picture; and (2) the camera metadata can indicate the photographer’s intention which may help us detect the attended regions of the images easily. The attended region (region of interest) can be defined as a single semantic object or a group of semantic objects in the pictures [4-5]. Such the attended regions can catch the viewers’ interests and represents the semantics of the pictures. Obviously, such the attended regions in a picture can provide an alternative way to interpret its semantics.

Based on these observations and understandings, we ask ourselves the following question: can the unspeakable esthetic sentiment in a photo be learned by a machine through studying large-scale photo collections taken by different people and cameras? If so, in a reverse way, the machine can hopefully understand where the attended region in a photo is and what the subject is in the attended region. Moreover, photos can be categorized and retrieved according to the similarity between their attended regions that contain the core visual contents. Detecting the attended regions in a photo has many other applications, such as more effective image compression (i.e., allocate more bits for the attended regions and less bits for other regions) and compact image content representation.

In this paper, we present an interesting algorithm for attended region detection. By using the attended regions for image content representation and feature extraction, we have built an automatic photo classification system. In section 2, we give a brief overview of the related work and point out our major contributions. In section 3, we describe how
a photo is captured and how metadata can be helpful. The algorithm for attended region detection and photo classification is presented in section 4. The algorithm evaluation is presented in section 5 with discussion, and finally we draw the conclusion in section 6.

2. RELATED WORK

Many content-based image retrieval (CBIR) systems were built based on the assumption that images can be characterized by using global features such as color histogram and textures [1]. Integrating both the global features and the local features for image classification and retrieval has been studied and reported, and the performance was significantly improved by integrating the local features for image classification and retrieval [1]. The local features can be obtained in three ways: image regions, grid-based image partition, and SIFT (scale invariant feature transform) features.

The philosophy behind the local/global argument is whether the human visual perception is a bottom-up or top-down process [2-4]. In a bottom-up approach, the informative regions with different visual properties from the surroundings are considered as attended regions. Because the semantic features (i.e., camera metadata) are not explicitly used for classifier training, most existing algorithms cannot be extended for handling large-scale photo collections with the unconstrained contents. Camera metadata, including exposure time, flash, object distance and focal length extracted from EXIF format (http://www.exif.org/Exif2-2.PDF), as well as the low-level cues were applied in in-door/out-door scene detection by using Bayesian network [6-7]. There is no existing work which can incorporate the camera metadata for attended region detection, photo classification and retrieval.

In this paper, we have developed an automatic approach to incorporate the camera metadata for attended region detection and consumer photo classification. Our major contributions consist of: (1) the attended regions in a photo are detected automatically by considering both the camera metadata and the visual features; (2) the attended regions are used to interpret the semantics of the photos and enhance photo classification.

3. PHOTO CAPTURE AND METADATA

A typical scenario of photo capture is illustrated in Fig. 1. A butterfly, which is the object of interesting to the photographer, appears in the right season at the right time (it’s unlikely to appear in winter on the snow at night). The photographer wears camouflage so that he can approach the object to achieve a better conformation. The parameter of the digital camera is set and the photo is captured intentionally. The camera records: (a) the objects and the scene in pixels that we can see; and (b) the underlying settings and the environment condition (such as GPS value and luminance etc.) in the metadata that we can read from EXIF files. The final photo intensively blur the foreground and the background by setting larger aperture so that the interesting object can vividly stand out.

Through controlling the camera metadata, photographer can intensively capture the objects or the scenes into photos, which he/she wants to share with others. Therefore, the camera metadata can somehow reflect the intentions of the photographers and the focus of the photos. Based on these understandings, camera metadata may play an important role in automatic image/photo understanding. Unfortunately, most existing techniques for image understanding focus on using only the information available in the pictures while completely ignoring the camera metadata (which may effectively reflect the intentions of the photographers and the focus of the photos).

In this paper, we have developed a new approach to explore the camera metadata (recorded in more than 200 different modes of digital cameras on the market in the last decade) for automatic image/photo understanding and retrieval. We have obtained that the following metadata are most helpful for photo understanding: the exposure value (EV), object distance and data/time. Although the GPS can be recorded in the EXIF file and it can provide the information where the photo is taken, there is no photo in our collection which records this message yet. Many digital cameras nowadays can select different mode of focal points such as central, face, or dynamically set anywhere using the touch screen or a joystick like device. With this information, one can know where the attended region is in the photo. Unfortunately, this information is not released to the public in the standard format. Another useful information, which is once available in the early of this century in a few camera models, is the object distance. But the parameter of the object distance is no longer released later probably due to the insufficient accuracy for predicting luminance compensation. Such the object distance can be integrated with the focal length for estimating the object’s size in a photo. With the help of the object distance, one can tell the difference between a photo of mountain scene and a duplicate of the mountain scene from a postcard. Based on the assumption that the objects in photos have temporal and spatial structures, data/time can help to predict the scenes/objects that can only be seen at certain time period.

Figure 1: Three components for photo capture environment: photographer and his/her attention, camera and its metadata, and scene and objects.

Figure 2: Major steps for attended region detection.
night. The photo database we have collected are the images which are captured for testing the quality of the cameras, most of the date/time recorded were not properly set. Thus we cannot incorporate the date/time for attended region detection and photo categorization.

The exposure value (EV) that is available in most of the camera models on the market today can be useful for attended region detection, such the exposure value may be directly provided in the metadata or can be calculated from the exposure time($t$) and the aperture ($f$-number, $N$):

$$EV = \log_2 \frac{N^2}{t}$$  \hspace{1cm} (1)

This value can be set with exposure prior or aperture prior with the other one automatically adjusted by the user or both set automatically in certain mode. The EV value is a function of luminance:

$$EV = \log_2 \frac{LS}{K}$$  \hspace{1cm} (2)

where $L$ is the luminance, $S$ is the ISO speed and $K$ is the reflected-light meter calibration (both $S$ and $K$ can be treated as constant). Given the EV, we can have more accurate measurement of the luminance. The luminance is closely related with the scene and it’s value is set according to the scene for a traditional camera.

Based on these observations, our approach for attended region detection and photo understanding is straight forward:

(a) Given the segmentation of an image, we first integrate the visual features of the image and the camera metadata to train the classifier for attended region detection; (b) The image (photo) classifiers are learned for partitioning the image collections into multiple categories according to the similarity between their attended regions.

4. ATTENDED REGION DETECTION

As shown in Fig. 2, our algorithm for attended region detection consists of the following key steps: (a) images are first segmented into multiple homogeneous regions; (b) a number of images regions are labeled as the training samples; (c) both the visual features of the image regions and the camera metadata are integrated to train a SVM classifier for attended region detection.

We have developed a GMM-based image modeling algorithm to enhance the image segmentation results. To estimate the optimal GMM model structure for image segmentation, the confidence scores of various models are calculated. Rather than obtaining the best configuration, we focus on finding the “average of the good ones” via a hierarchical approach. As shown in Fig. 3, our GMM-based image segmentation algorithm can provide more reliable results. A hierarchical tree can be formed by recursively partitioning the whole image into multiple homogeneous regions according to the confidence map of the contours.

The region-based visual features include 26-dimensional features: (a) 3-dimensional colors; (b) 3-dimensional textures; (c) 2-dimensional alignment and compactness of region shape; (d) 2-dimensional contrasts with the neighboring regions; (e) 16-bins EV histogram (camera metadata). SVM classifier is trained to detect the attended regions by using the region-based visual features and the camera metadata. As shown in Fig. 4 and Fig. 5, one can observe that our algorithm can obtain the underlying attended regions accurately.

From these experimental results, one can observe that such the attended regions can characterize the underlying core contents of the images effectively. Thus detecting the attended regions for image content representation may provide more effective approach for image classification. We have extracted the following visual features for image content representation: (1) 36-bin RGB color histogram to characterize the color distributions of the attended regions; (2) 8-dimensional Tamura texture features to characterize the
Figure 6: (a) Concept ontology for image concept organization and classifier training; (b) EV distribution among different categories.

Figure 7: The comparison results on attended region detection.

5. ALGORITHM EVALUATION

The benchmark metric for algorithm evaluation includes precision $\phi$ and recall $\varphi$. They are defined as:

$$
\phi = \frac{\zeta}{\zeta + \psi}, \quad \varphi = \frac{\zeta}{\zeta + \xi}
$$

where $\zeta$ is the set of true positive regions that are related to the attended regions and are detected correctly, $\psi$ is the set of true negative regions that are irrelevant to the attended regions and are detected incorrectly, and $\xi$ is the set of false positive regions that are related to the attended regions but are mis-detected.

We have compared two approaches for attended region detection: (a) contrast-oriented approach without camera metadata; (b) our classification-oriented approach with camera metadata. The comparison for their average performance on attended region detection are given in Fig. 7. From these experimental results, one can observe that our algorithm can detect the attended regions in the photos more accurately.

Figure 8: The comparison results on the precision $\phi$ between our structured max-margin learning algorithm and flat approach for some image concepts and objects.

We have compared the performance differences between two approaches for image classification: (a) only global and local visual features are used; (b) both camera metadata and the global and local visual features are integrated. As shown in Fig. 8, one can observe that our algorithm scheme can improve the classification accuracy significantly by incorporating the camera metadata for classifier training.

6. CONCLUSION

In this paper, we have developed a new algorithm to calculate what the attention of the photographer is and what the core content of a photo is. The gained information (i.e., attended regions and attention of photographer) is further used to build a more effective photo classification and retrieval system. Our experiments on 70,000+ photos taken by 200+ different models of cameras with variety of interests have obtained very positive results. Our future work will focus on evaluating our algorithms on large-scale image collections.

7. REFERENCES