I-009

Attention-based Clustering for Grouping Photo Collections by Quality of Composition

デシルヴァ ガムヘワゲ† 相澤 清晴†‡
Gamhewage de Silva† Kiyoharu Aizawa†‡

1. Introduction

Recent years have seen a rapid increase of the number of digital photos shared on the Internet. These photos come from different types of cameras, and photographers with a wide range of skill levels and tastes. As a result, the photographic quality of such photos is extremely diverse. A user will find it easier to retrieve photos if search results are grouped by photographic quality. There has been some recent research on automated estimation of photo quality, to support such applications. Most of these approaches are based on low level image features such as distribution of color and brightness, and the presence of sharp edges and blurry regions [1].

Photographic composition (also referred to as *composition*) is one of the most important features of a good photo. A photo is considered to be well-composed if the subject of the photo stands out from the rest of the content [2]. Therefore, automatic categorization of photos by composition can be very helpful in selecting good photos.

The intention of a photographer when composing a photo is to ensure that the subject of the photo gets the most amount of attention. Recent work by Itti et al. [3]proposed *saliency maps* as a reliable way to computationally model visual attention on digital images. Saliency maps have also been successfully used for automatically cropping images and creating thumbnails [4].

In this paper, we investigate the possibility of using saliency maps to automatically estimate the quality of composition. Our intention is to group a collection of photos based on the quality of composition, making it easier for users to find photos with desired composition and quality from a large collection. We first derive features from saliency maps to represent the characteristics of well-composed images. Thereafter, we extract these features from a collection of reviewed photographs from a stock photography archive and images uploaded to Flickr. We apply both supervised and unsupervised learning algorithms on this data collection, and present the results.

2. Feature Selection

A saliency map of a given image is a two dimensional representation of the amount of visual attention received by different regions of the image. Figure 1 shows a digital photo with its saliency map. The brightness of a region in the saliency map is proportional to the amount of visual attention received by that region. Matching the photo with the saliency map shows that salient points are generally located close to corners, sharp edges and contrasting regions.

The first step in classification of images by using saliency maps is to extract features from them to represent the quality of photographic composition. We use the following simple guidelines about good composition, to derive feature vectors:

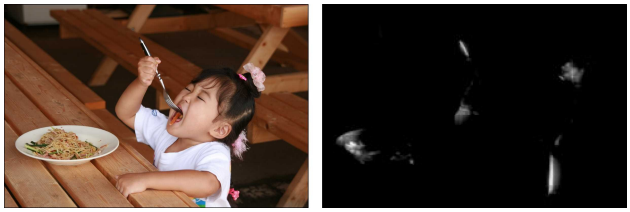

Fig. 1. A photo and its saliency map

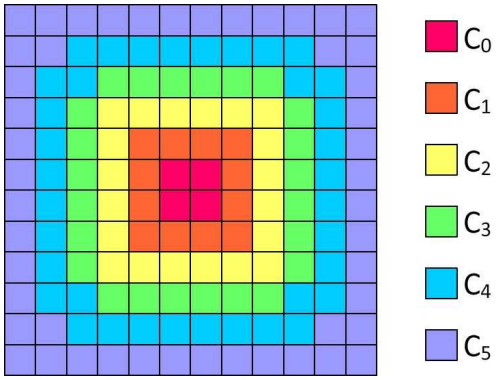

Fig. 2. Deriving feature vectors from composition map.

(1) A well-composed photo normally contains a subject that attracts attention of the viewers, and a background that receives little attention.
(2) A photo is not considered as well-composed if there are several, distributed objects that attract attention.
(3) In some well-composed photos, objects are arranged in a way that the viewer's eye follows a given (generally linear) path
(4) The rule of the thirds: important compositional elements of a photo are generally positioned around the intersections of the lines that divide it in to thirds (both horizontally and vertically).
(5) The corners of a photo should not contain objects that distract a viewer's attention from the subject.

We extracted the following features from the saliency map of an image, to correspond to the guidelines (1) and (2):

- Average value of saliency
- Minimum value of saliency
- Maximum value of saliency
- Ratio of points with non-zero saliency to the total number of points on the saliency map

The following two features were extracted to represent guideline (3):

- The length of the path connecting the centers of the five most-salient regions of the image, in the descending order of saliency.
- The length of the shortest path connecting the above centers, as determined using the *greedy algorithm*.

† 東京大学大学院情報学環学府
‡東京大学大学院情報理工学研究科

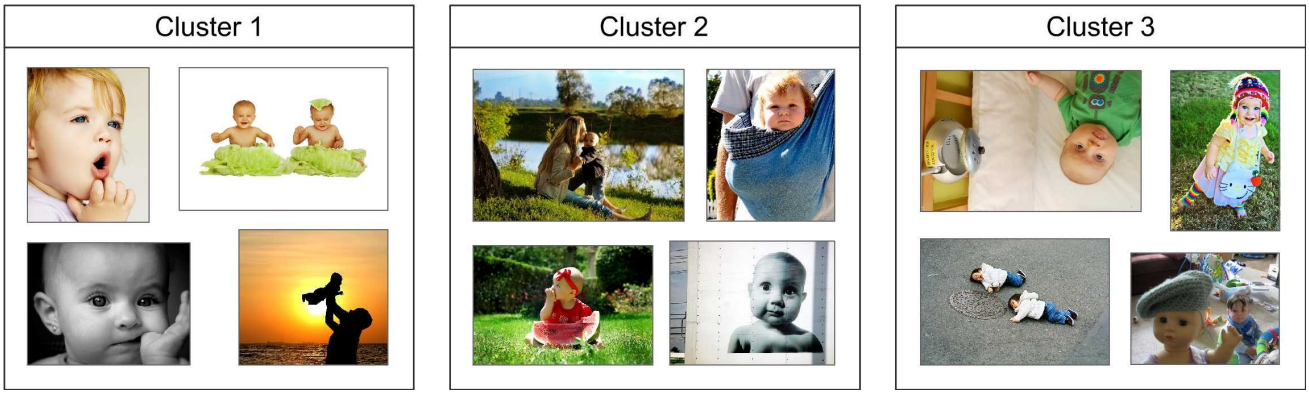

Fig. 3. Example images from different clusters.

Since the salient points are located mostly on the perimeter of a subject, it is not straightforward to match the rule of the thirds to the positions of the salient points. We examined the distribution of salient points in 200 well-composed images and designed a 12×12 “composition map” as shown in Figure 2. Each saliency map was re-scaled to the size of the composition map, and the feature values C0 to C6 were estimated as the fraction of non-zero salient points in the region with the corresponding color.

The coordinates of the centroid of the salient regions, normalized to the size 12×12, is also used as a feature. This point acts as an approximation of the location of the subject, for images with one subject and a blurry background.

3. Experiments and Results

3.1 Data collection

We selected a small data collection to evaluate the feasibility of photo classification using the above features. 180 images corresponding to the keyword “baby” were extracted from two online photo galleries. Half of them came from *Stock Photo Exchange* (*http://sxc.hu)*, a stock photographers' site where images are reviewed for quality before being accepted. The other half was split between the most recent and the most interesting images from *Flickr*. Each photo was labeled as *well-composed or other*, by two photographers including one of the authors.

The following subsections describe the experiments that we carried out using this data set.

3.2 Supervised Learning

We evaluated the performance of different classifiers based on supervised learning for classifying images according to the class labels in Secton 3.1. Due to the small size of the data set, 10-fold cross validation was used to test the classifiers. *MultiBoost Adaboost* classification gave the maximum accuracy of 70%. While the accuracy was fairly low, we believe that the results can be improved by fine tuning the features.

3.3 Unsupervised Learning

We also investigated the natural groupings within the complete data set by clustering them using simple expectation maximization. This resulted in three distinct clusters. Figure 3 shows representative images from each cluster. Cluster 1 was dominated by well-composed photos and consisted of photos with subjects that stand out from the background. The majority of photos in Cluster 3 were not composed well, and the others had a high depth of field. Cluster 2 contained photos with a large background area and well-exposed subjects.

The results of unsupervised learning seem more convincing than those of supervised learning. Given that the perception of photo quality is highly subjective, we believe that it is a better approach to cluster photos according to composition than making a binary decision on whether a photo is well-compoed or not.

4． Conclusion and Future Work

We reported the results of an initial study on the use of computational visual attention for classification of digital photos based on photographic composition. Supervised learning based on a small data set yielded an accuracy of 70% in recognition of well-composed photos. Unsupervised learning using expectation maximization yielded three clusters of photos with a meaningful grouping.

These results show that it is feasible to achieve automated image classification according to photographic composition. However, the data set used for the experiments is quite small. We intend to use a much larger image collection and revise the features to improve the accuracy. It is also possible to classify photos based on other criteria (such as color balance) as well, to allow additional methods of grouping image search results.

References

[1] R. Datta, D. Joshi, J. Li, J. S. Wang, “Studying Aesthetics in Photographic Images Using a Computational Approach”, in *Proc. ECCV*, III:288-301 (2006).
[2] C. de Silva, “Take Great Photos with Your Digicam”, ISBN 978-0-557-22130-1, p.29 (2009).
[3] L. Itti, C. Koch, “Computational Modelling of Visual Attention”, Nature reviews, Neuroscience, Vol. 2, No. 3. (March 2001), pp. 194-203.
[4] M. Nishiyama, T. Okabe, Y. Sato, I. Sato, “Sensation-based Photo Cropping”, *in Proc. ACM Multimedia 2009*, pp. 669-672.

† Interfaculty Initiative in Information Studies, the University of Tokyo

‡ Dept. of Information and Communication Engineering, the University of Tokyo