An Autonomous Aesthetics-driven Photographing Instructor System with Personality Prediction

Chin-Shyurng Fahn and Meng-Luen Wu

Department of Computer Science and Information Engineering,
National Taiwan University of Science and Technology
Taipei 10607, Taiwan, R. O. C.
{csfahn,D10015015}@mail.ntust.edu.tw

Abstract

In this paper, an autonomous aesthetics-driven photographing instructor system is proposed, which gives instructions to help camera users to take good images. There are two kinds of instructions: image composition and personality feature enhancement. As for composition, a salient region is used to match the aesthetical template. To keep the personality of the autonomous photographing instructor system, the correlation between user types and image features is extracted through data mining on social networks, which is called "personality prediction." The proposed system is run in realtime and workable on all mobile devices with cameras.

Keywords: Computer vision, computational aesthetics, autonomous photographing instructor, personality prediction, social network.

1. Introduction

In this paper, we propose an autonomous aesthetics-driven photographing instructorsystem. This system guides users to move their cameras zoom in and out, as well as changing exposure values and so on, which is shown in Fig. 1.

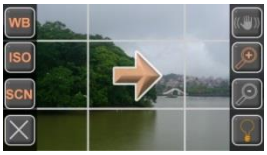


Fig. 1: Prototype of our developed photographing instructor system.

The instructions are based on existing aesthetics rules and theories. Rule of thirds, both horizontal and perspective compositions are popular image composition rules abided by professional photographers. Figure 2 graphically showsan instance of the golden ratio that is defined below:

$$\frac{a+b}{a} = \frac{a}{b} = \varphi, \text{ with } \varphi = \frac{1-\sqrt{5}}{2} \cong -0.618 \quad (1)$$

The value 0.618 is approximate to $2/3$, which is the basis of rule of thirds, and this ratio is adopted by most artists for deciding the proportions in their work.

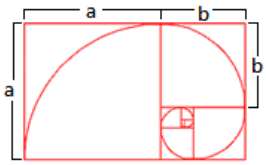


Fig. 2: An example of the golden ratio.

Low level features, such as brightness, color contrast, color harmony, and simplicity, are obtained from the color distribution of images. Figure 3 shows two images with different low level feature values.

(a) (b)

Fig. 3: Images with different color contrasts: (a) high color contrast; (b) low color contrast.

In our proposed method, low level features are chosen todevelop theautonomous photographing instructor system.

2. Related Work

The research fields similar to our work are computational aesthetics, image composition, and robot photographers. The following states these topics.

In [1], Ke et al. described the design of image features for assessing the quality of images. In [2], Obrador et al. indicated that the image composition is the key criterion for evaluating the aesthetics value of an image. In [3], Yeh et al. applied image features to perform photo ranking, where the images with higher qualities have preferable rankings.

A saliency map illustrates the eye-catching regions, which is a factor for evaluating the image composition. The salient region in an image is a set of areas which are eye-catching. There were some salient region detectors published in recent years. In [4], Achanta et al. presented a frequency tuned (FT) method. In [5], Chen et al. proposed both histogram contrast (HC) and region contrast (RC) methods. The RC method has the best accuracy, but ittakes $O(n^2)$ to compute a pixel of an image of sizen^2.To the contrary,the FT method only requires O(1) for a pixel. Although FT has the least time complexity, it's not capable enough for marking up the saliency of the same object according to our experiments.

In [6], Raughdeep et al. developed a system to capture images automatically, which employs the saliency map only. Sony and Microsoft also have their photographing robot products or concepts. Sony party-shot uses face detection todo image composition. Microsoft EDDIE uses Kinect and human skeleton to accomplish image composition. However, neither of them can work when no people are in the field of views (FOV), which cannot captureimages under complex natural environments.

3. Our Proposed Method

In this section, we will describe the system implementation. The system gives two types of instructions to users: image composition and feature enhancement. The design of instructions iselaborated in the following respectively.

3.1 Salient region detection

In the state-of-the-art algorithms, accurate foreground detection spends much computational time. What's more, in many cases, there exists no foreground, especially in landscape scenes. As a result, we resort to the salient region instead of foreground detection. In some of images, there are multiple salient regions, and photographers only choose one of them to execute image composition. Therefore, cropping the salient region is needed for image composition.

First, we adopt the HC method to obtain the saliency map of the original image [6]. The saliency map is set up by:

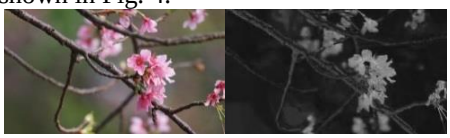
$$S(p_i) = S(c_l) = \sum_{j=1}^{n} prob_j D(c_l, c_j) \quad (2)$$

where c_l is the color of pixel p_i, n is the number of colors in image I, D is the distance of two colors in the CIELab color space, and $prob_j$ is the probability that c_j appears in image I.

Because the number of colors is huge in natural images, color reduction must be applied to group similar colors into one. The color reduction formula is expressed as follows:

$$c_i = \left\lfloor \frac{c_i}{c_{max}/K} \right\rfloor \times (c_{max}/K) \quad (3)$$

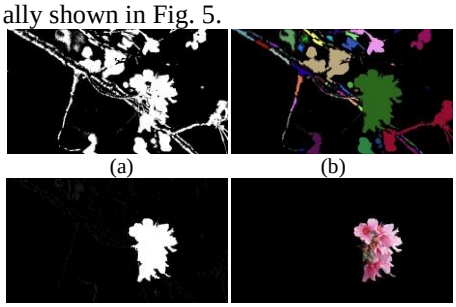
where c_i is the color of pixel p_i, c_{max} is the maximum color intensity of the color space, and K is the desired color count. The example of the saliency map is shown in Fig. 4.



(a) (b)

Fig. 4: Salient region detection: (a) the original image; (b) the saliency map obtained from the HC method.

Subsequently, we perform binarization of the saliency map that is used for finding the contours. The salient contour is computed by the method proposed in [7]. Next, the largest contour is chosen as the foreground mask. We choose the largest contour for two reasons. First, the proportion of the foreground in an image should be as large as possible. Second, the smaller contours are often noises. The process of finding the foreground is visually shown in Fig. 5.



(a) (b)

(c) (d)

Fig. 5: Finding the foreground for image composition: (a) the binary saliency map; (b) the contours of the saliency map; (c) the largest contour; (d) the foreground in the largest contour.

3.2 Feature extraction and personality

In this paper, low-level features are utilized for deciding whether an image matches camera user's taste. There were quite a few image features proposed in previous literatures. Here, we present the features that are able to help users.

Hue, Saturation, and Brightness

In computational aesthetics, the hue, saturation, and brightness channels represent the features of image. We define three features: f_{hue}, $f_{saturation}$, and $f_{brightness}$ as the average pixel values of the channels respectively. For example, the brightness feature is depicted below:

$$f_{brightness} = \frac{\sum_{x=1}^{width} \sum_{y=1}^{height} I(x,y)}{width \times height} \quad (4)$$

where $I(x,y)$ is the intensity of a pixel at (x,y).

Simplicity

The simplicity feature is computed from the color distribution of an image. In [2], the formula of the simplicity feature yields:

$$f_{simplicity} = \left(\frac{|\{ l \mid k(c_l) \geq \gamma\, k_{max} \}|}{4096} \right) \times 10 \quad (5)$$

where $k(c_l)$ is the color count for color c_l, k_{max} is the maximum color count, and γ is set to 0.001. In this formula, the color count in the image is reduced to 4096; that is, the color counts of R, G, and B are all reduced to 16.

Color contrast

Color contrast is a measurement for image and display qualities. The color contrast feature is defined as:

$$f_{contrast} = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} (1 - d(i,j)) \frac{D(i,j)}{A_i A_j} \quad (6)$$

where $d(i,j)$ is the spatial distance between the centroids of two segmentations i and j; $D(i,j)$ is the color distance between the two segmentations in the CIELab color space; A_i and A_j are the areas of the segmentations individually. And n is the number of segmentations.

Harmony

Harmony colors are known to be aesthetically pleasing in terms of human vis-

ual perception. The optimization function is provided by Cohen-Or et al. for the measurement of harmony feature as [8]:

$$F(m,\alpha) = \sum_{p\in I} \Delta\left(H(p),\ h_{T_m(\alpha)}(p)\right)\cdot S(p) \quad \text{(7-a)}$$

where H and S are the hue and saturation channels for an image, respectively, and p is denoted each pixel in the input age$I$; $\Delta(i,j)$ is the arc-length distance on the hue wheel measured in us; $h_{T_m(\alpha)}(p)$ is the sector border hue of a harmonic template T_m together with an associated orientation $\alpha \in [0,2\pi)$, and m is the index for harmonic templates. The best harmonic template and the orientation are determined to minimize the optimization function so as to create the most pleasant visual result. According to this, we define the harmony feature below.

$$f_{harmony} = Minimize\ F(m,\alpha) \quad \text{(7-b)}$$

Blur

A blurry image is almost worse than a sharp image of the same scene. Measurement for image sharpness is as follows:

$$f_{\text{sharpness}} = \frac{|\{(u,v)||F(u,v)|>\xi\}|}{width\times height} \propto \frac{1}{\sigma} \quad \text{(8-a)}$$

$$f_{\text{blur}} \propto \frac{1}{f_{\text{sharpness}}} \quad \text{(8-b)}$$

where $I_{blur} = G_\sigma * I$ is the blurred image derived through convolving the original image I with a Gaussian filter G_σ, and $F(u,v) = FFT(I_{blur}(x,y))$ is the blurred image transformed into the frequency domain via the fast Fourier transform. Here, ξ is set to 5.

Color components

In a human visual system, there are three channels for color perception as Fig. 6 shows. The main colors that human perceive are red, green, yellow, and blue. The feature for a color component is defined as:

$$f_{colorcomponent} = \sum_{x=1}^{width}\sum_{y=1}^{height}\frac{D(c_l,c(x,y))}{width\times height} \quad (9)$$

where $D(i,j)$ is the color distance between two colors in the CIELab color space, $c(x,y)$ is the color at (x,y), and c_l is the color component to measure. Here, c_l is assigned to red, green, yellow, or blue.

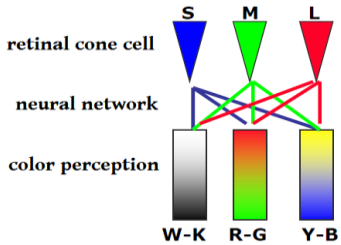


Fig. 6: Phase theory for human color perception.

Feature relation mining

Thanks to the evolution of social networks, which enables users to upload and link images to their own personal pages, we can collect and download their posted images by web crawlers. We adopt feature extractors to acquire the information of each crawled image. We choose the "cover photo" in the personal page, where the image is selected and uploaded by the user. Such strategy is chosen because the "cover photo" always represents the style of a user.

4. Photographing Instructions

In this section, we will describe how to implement the factors in Section 3 into our autonomous aesthetics-driven photographing instructor system.

4.1 Image composition

In [1], some professional photographing composition templates are introduced, as shown in Fig. 7.

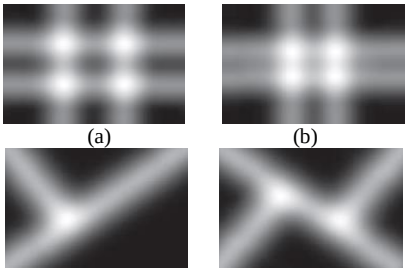


Fig. 7: Image composition templates in form of: (a) the rule of thirds; (b) the golden ratio; (c) the golden triangle; (d) the golden triangle combination.

The intersections of lines are called "power points." Saliency regions close to one of the power points have preferable composition. The centroid of the foreground region put at the brighter points of the template has higher composition scores.

The centroid of the largest contour is employed to guide users to move the camera to match the composition template. However, it's impossible to request users to do point to point template matching. Thus, an approximate matching method is adopted. We set a score for the feature of template matching as:

$$f_{template\ matching} = \min\left\{\sqrt{(C_{lcx} - P_{ix})^2 + (C_{lcy} - P_{iy})^2}\right\} \quad (10\text{-a})$$

$$S_{template\ matching} \propto \frac{1}{f_{template\ matching}} \quad (10\text{-b})$$

where P_i is the position of the ith power point, and C_{lc} is the centroid of the largest contour. A threshold is set that $f_{templatematching}$ should be less than the half of the distance of two power points. Here, we set a threshold τ to be the minimum distance of two power points.

The ratio of a foreground tothebackground determinates a good composition.If the ratio is far from 0.618, an indication of zoom in or zoom out will be signaled.In addition, if the estimated moving distance is too long, then it's possible that the largest contour will change while moving, and the trial of matching should be abandoned. The estimated moving distance is formulated as follows:

$$dist = \min\left\{\sqrt{\left(C_{lcx} - P_{ix}\right)^2 + \left(C_{lcy} - P_{iy}\right)^2 + \left(\frac{Area_{lc}/(width \times height) - 0.618}{K}\right)^2}\right\} \quad (11)$$

where i is the index for all power points. $\boldsymbol{P_i}$ is the position of a power point, and $\boldsymbol{C_{lc}}$ is thecentroid of the largest contour. Besides, $\boldsymbol{Area_{lc}}$ is the area of the largest contour; $\boldsymbol{K}$ is the distance for changing one percent area ratio of the largest contour within the image, which is obtained by camera intrinsic parameter calibration.

The rules for deciding whether the user is guidedby means of image composition are summarized as follows:

IF $f_{template\ matching} \leq \tau$ **THEN** *Keep guiding*
IF $f_{template\ matching} > \tau$ *and* $dist \leq \delta$ **THEN** *Stop guiding and take an image*
IF $f_{template\ matching} > \tau$ *and* $dist > \delta$ **THEN** *Keep Guiding*

where δ settles the precision of template matching, which should be less than τ.

4.2 Personality feature enhancement

As illustrated in Fig. 8, digital cameras have the ability to enhance image features. In our system, the features of views are extracted to help users take better images.

The instructions can be given according to user attributes, such as age, education, gender, and so on. The correlation between a feature and a user group can serve as the weight of thefeature, which is called "personality prediction." The weights are used for calculating the score of a view, which is defined as:

$$\boldsymbol{S_{view}} = \sum_{i=1}^{n} \frac{\boldsymbol{S_i(corr_i + B)}}{\boldsymbol{n}} \quad (12)$$

where n is the number of features; S_i is the normalized score of the ith feature that has a range of 0 to 1; $corr_i$ is the correlation between the user type and its corresponding feature, which is normalized

from 0 to (1-B). And B is a base value ranging from 0.5 to 1.

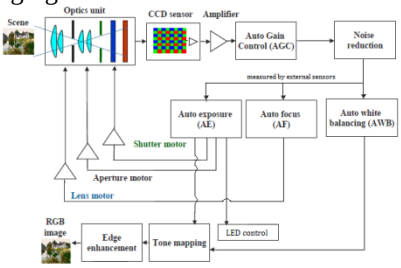


Fig. 8: Digital camera imaging pipeline diagram in simplified form.

The base value is usedfor bias prevention; in some cases, $corr_i$ is close to 0, which will incur a biased score. B is set to a lower value larger than 0.5 if more information is adopted for computing the correlation; otherwise, B is set to a higher value less than 1. If there is no user information retrieved, we set B to 0.5.

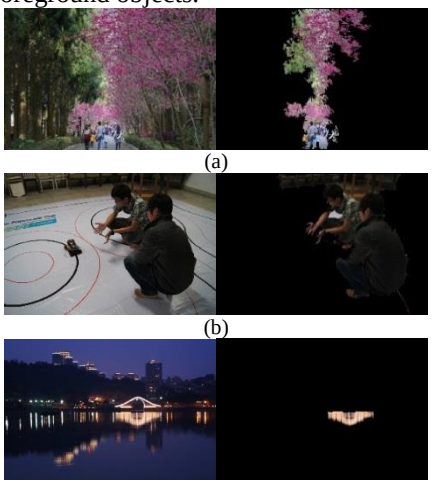
Only when the score is above the threshold will instructions be given. These enhancements are relied on the result of predicted personalities. The following describes the instructions for feature enhancements as listed in Table 1.

Table 1. Instructions to Enhance Features

Conditions	Instructions
Brightness low	1. Increase the EV value 2. Increase the ISO value 3. Turn on the LED light
Brightness high	1. Decrease the EV value 2. Decrease the ISO value 3. Install a glare shield
Simplicity low	1. Zoom in on a subject 2. Move to another scene
Contrast low	1. Change the tone-mapping method
Saturation low	1. Change the tone-mapping method 2. Change the white balance
Harmony low	1. Change the tone-mapping method 2. Change the white balance

5. Experimental Results

The main contributions of our proposed method are camera orientation guided by the result of salient region cropping. As shown in Fig. 9. The proposed method is capable of handling complex environments and FOV without foreground objects.



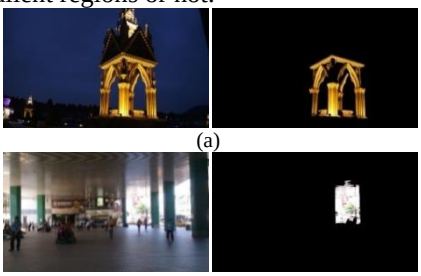
(a)

(b)

(c)

Fig. 9: The salient region detection results from our proposed method: (a) without actual foreground; (b) two dark foregrounds and a bright background; (c) the special case that the water reflection is larger than the actual foreground object.

In Fig. 10, we can see that the result from an under-exposure image is correct, while in an over-exposure image, the salient regions are always the sky or areas illuminated by sun. Therefore, for an image under daylight, the regions for white areas in the largest contours must be examined first and decide whether they are salient regions or not.



(a)

(b)

Fig. 10: Salient region detections in two types of exposure for: (a) under-exposure; (b) over-exposure.

Second, for the "personality prediction," we examine correlations between image features and user types. We collect over

100,000 images and some attributes from the image owners, such as age, education, sex, and occupation. By clustering the images with respective features as well as calculating the correlations between user types and image features, we found some results worth mentioning. There are some clusters that cover people from all types. In these clusters, the images have higher scores in most of the features. However, some image groups are shared by specific user types. The result is as illustrated in Fig. 11.

Fig. 11: Groups of images which are only accepted by specific user types: (a) lower sharpness and color contrast; (b) lower brightness and higher yellow color feature; (c) lower brightness and higher sharpness; (d) higher brightness and higher color contrast.

6. Conclusion and Future Work

In this paper, an autonomous aesthetics-driven photographing instructor system has been developed to help camera users take good images. The system runs in realtime and works under all environments that no detected object is needed. The system is also capable of dealing with multiple salient regions in an image. We also present how to handle personality of aesthetics with the aid of crowd mining.

The developed system can be further installed in a robot to attainautonomous photographing. Search path optimization canbe applied in such a system; thus, itimproves the performance in capturing a series of images.

Acknowledgement

The authors thank the National Science Council of Taiwan for supporting this work in part under Grant NSC101-2221-E-011-140-MY2.

References

[1] Y. Ke, X. Tang, and F. Jing, "The design of high-level features for photo quality assessment," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*,vol. 1, pp. 419-426, New York, NY, 2006.

[2] P. Obrador, L. Schmidt-Hackenberg, and N. Oliver, "The role of image composition in image aesthetics," in *Proceedings of International Conference on Image Processing*, pp. 3185-3188, Hong Kong, China, 2010.

[3] C.-H Yeh, Y.-C. Ho, B. A. Barsky, and M. Ouhyoung, "Personalized photograph ranking and selection system," in *Proceedings of the International Conference on Multimedia*, pp. 211-220, Firenze, Italy, 2010.

[4] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 1597-1604, Miami, Florida, 2009.

[5] M.-M. Cheng, G.-X. Zhang, N. J. Mitra, X. Huang, and S.-M. Hu, "Global contrast based salient region detection," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 409-416, Colorado Springs, Colorado, 2011.

[6] R. Gaddeand K. Karlapalem, "Aesthetic guideline driven photography by robots," in *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence*, vol. 3, pp. 2060-2065, Barcelona, Spain, 2011.

[7] S. Suzuki, "Topological structural analysis of digitized binary images by border following," *Computer Vision, Graphics, and Image Processing*, vol. 30, no. 1, pp. 32-46, 1985.

[8] D. Cohen-Or, O. Sorkine, R. Gal, T. Leyvand, and Y.-Q. Xu, "Color harmonization," *ACM Transactions on Graphics*, vol. 25, no. 3, pp. 624-630, 2006.