SALIENCY-ENHANCED IMAGE AESTHETICS CLASS PREDICTION

Lai-Kuan Wong *and* *Kok-Lim Low*

National University of Singapore
{lkwong, lowkl}@comp.nus.edu.sg

ABSTRACT

We present a saliency-enhanced method for the classification of professional photos and snapshots. First, we extract the salient regions from an image by utilizing a visual saliency model. We assume that the salient regions contain the photo subject. Then, in addition to a set of discriminative global image features, we extract a set of salient features that characterize the subject and depict the subject-background relationship. Our high-level perceptual approach produces a promising 5-fold cross-validation (5-CV) classification accuracy of 78.8%, significantly higher than existing approaches that concentrate mainly on global features.

Index Terms— Aesthetics, saliency, classification

1. INTRODUCTION

Computational image aesthetics evaluation can be very useful in various photographic applications, such as digital photo-editing, content-based image retrieval, content-based document design, and even during photo-taking. Existing work [2, 6, 9] based on the computation of aesthetics features and photographic rules have shown promising results. However, despite the different approaches and the different set of features used to train the evaluation models, they yield about the same classification accuracy of about 70% to 72%. One underlying limitation may be that these methods, except Datta et al.'s [2], focus mainly on global image features. Datta et al.'s method computes a set of region composition features based on regions extracted using low-level color analysis, and no higher level perceptual information is used to determine the regions.

In this work, we explore the use of higher-level perceptual information, based on *visual attention*, for the classification of *professional photos* and *non-professional snapshots*. Studies have shown that there exists strong correlation between visual attention and visual aesthetics. According to Lind [5], aesthetic objects are interesting and thus, can hold and attract attention. Similarly, Coe [1] discovered that aesthetics is a means to create attention to an object or a person. These studies suggest that visual attention may be a key to aid the evaluation of photographic aesthetics. The main contribution of our work is, in addition to a set of global image features, our method identifies a set of salient regions in the photo and computes a set of saliency features within them. These salient regions potentially correspond to the attention-catching areas in the photo. We also extract a set of features based on the relationship between the salient regions and the background. Together, these provide higher-level visual information, such as the quality of the photo subject and the subject-background relationship, which we have found to be significantly discriminative features.

2. RELATED WORK

Classification of photographs based on aesthetics measures was first attempted by Tong et al. [6], in which they took a black-box approach to classify photographs into professional or snapshots. A large set of 846 low-level features were combined exhaustively with a standard set of learning algorithms for classification. Although this approach successfully classifies photographs with an accuracy significantly better than chance, it offers little insight into why certain features are selected, or how to design better features. Yan et al. [9] tried to address the above limitations by using a principled approach. They studied the perceptual criteria that people use to judge a photo and presented a top-down approach to construct high-level semantic features for assessing the quality of the photos. With a small set of highly discriminative high-level semantic features, they achieved a classification accuracy of 72.3% using a Naïve Bayes classifier, an accuracy comparable to that of Tong et al.'s approach.

In a similar work, Datta et al. [2] computed a set of 56 features based on rules of thumb in photography, common intuition and observed trends in ratings. Combining filter-based and wrapper-based methods, they shortlisted a set of 15 features and used them to classify photos into 'high' and 'low' classes. Using a set of photos from Photo.net, with aesthetics scores ranging from one to seven, and excluding photos with average scores between 4.2 and 5.8, they obtained a classification accuracy of 70.12% using an SVM classifier.

Comparing these approaches, Yan's and Datta's approaches are more objective and efficient for aesthetics class prediction. Yan's has a smaller set of more discriminative features compared to Datta's larger but weaker set of features. However, all methods obtained about the same classification accuracies even though different sets of features and classifiers have been used. A possible bottleneck of these approaches is that none of these methods consider features specific to the photo subject, which potentially provide insight into a better set of discriminative features.

3. OUR APPROACH

The photo subject of an image may be one of the most distinguishing factors between *professional photos* and *snapshots* [9]. We address the limitation in existing work by employing a saliency approach to analyze and extract features associated with the subject. We make the assumption that the salient regions of an image contain the subject and use the salient regions to represent the subject. The salient regions are identified using the visual attention model of Itti et al. [4], which is built upon a biologically plausible architecture.

We collected a set of peer-rated images from Photo.net, which is an online photo-sharing community. We chose Photo.net because it has a better consensus over its ratings [3]. Following

Yan's approach [9] to discern *professional photos* from *snapshots*, we extracted the top and bottom 10% of the photos and assigned them as *high-quality professional photos* and *low-quality snapshots*. Each image I is converted to the HSV and LUV color spaces and the resulting two-dimensional matrices with size of $X\times Y$ are denoted, as I_H, I_S, I_V, and I_L, I_U, I_V respectively. First, we compute the saliency map SM on I and determine a set of salient locations L. Using the salient locations in L as seeds, we perform seeded segmentation to create a salient mask K that indicates the salient regions. Both the saliency map SM and salient mask K are of size $X\times Y$. The salient region with an area of M, is defined for each HSV channel as

$$S_{ch} = \{ I_{ch}(x, y) \mid K(x, y) > 0\} \tag{1}$$

where $M = |S_{ch}|$ and $ch = \{\text{H, S, V}\}$. Similarly, the background region with an area of N is defined as

$$B_{ch} = \{ I_{ch}(x, y) \mid K(x, y) = 0\} \tag{2}$$

where $N = |B_{ch}|$.

Each image I and its corresponding saliency map SM, salient region S_{ch}, and background region B_{ch}, are then used to extract a set of global image features and a set of features that characterize the subject and its relationship with the background. Altogether, we compute a total of 44 candidate features, $F = \{f_1, f_2, \ldots , f_{44}\}$. Finally, using a set of images, each with a set of features F, we build a two-class classification model that classifies an image I into either class 1 or 0, where

$$class(I) = \begin{cases} 1 & \textit{professional photo} \\ 0 & \textit{snapshot} \end{cases} \tag{3}$$

4. SALIENT REGIONS EXTRACTION

Extracting the main subject from a photo is non-trivial. Many methods exist but none is close to perfect or may only work well for certain type of photographs. However, some do give good hints on the salient regions of interest that attract attention and thus can be utilized to identify photo subjects.

For each image, we compute the saliency map and the salient locations using Itti's visual saliency model [4], which is built upon a biologically plausible architecture that exploits multi-scaled intensity, color and orientation image features and learnt the salient locations using a Winner-Take-All (WTA) neural network framework. The salient locations are learnt in a sequential manner and we observed that the sequence eventually returns to the first salient location after a certain number of locations. We trace and extract only the unique salient locations. The sequence of these locations mimics the navigation pattern of the viewers, and could potentially provide clue about the image attractiveness. For our purpose, we use only the first three salient locations as seeds for segmentation as previous work [4] has shown that in over 90% of the cases, the main subject is discovered within the first three salient locations. We then perform segmentation on image, I using the CTM segmentation engine [10], a state-of-the-art segmentation technique for natural images and extract all segmented regions that contain these three seeds to create a salient mask. This salient mask is used to create the salient region S_{ch} and its complimentary background region B_{ch} for each HSV channel. Figure 1 shows a photo with the salient locations and its corresponding saliency map, the segmentation result and the salient mask. The white pixels of the salient mask indicate the salient regions whereas the black pixels denote the background region.

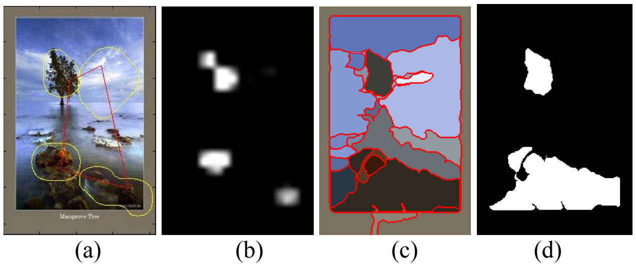

Figure 1. Salient image regions extraction. (a) Original image with salient locations, (b) saliency map, (c) segmented image, (d) salient mask based on the first three salient locations.

5. VISUAL FEATURES EXTRACTION

We consider three types of features—global image features, features of salient regions, and features that depict the subject-background relationship.

5.1. Global Image Features

There are three categories of global image features, based on basic techniques, photographic rules, and camera settings. For basic techniques, we chose sharpness, contrast and exposure. A professional photo would be well-exposed, have a sharp main subject, and high contrast. The blur estimation method used by Yan et al. [9] is discriminative but rather complex with the need to combine a number of techniques: Fourier transform, Haar wavelet transform, and Naïve Bayes. We have extended their Fourier transform method to derive a simpler integrated method for detecting sharp images, but due to the space constraint, we leave out the details here. Altogether, we computed three sharpness features, f_1, f_2 and f_3. We have also enhanced Yan's method to compute image contrast, f_4, by taking the middle 98% mass of the luminance I_L histogram, instead of using the combined RGB histogram as computed by [9]. The luminance records how light intensity is perceived by the human eye and therefore the I_L histogram is a more intuitive representation of image contrast. For measure of exposure, we compute the average pixel intensity, f_5 [2].

For the photographic rules, we measure the texture details, the low depth of field (DOF), and the rule of thirds. We adopted Datta's method to compute features for texture details, (f_6 to f_{17}), low DOF (f_{18} to f_{19}), and rule of thirds (f_{20} to f_{21}). The low-DOF and rule-of-thirds features are computed only for the saturation I_S and intensity I_V channels.

Camera settings information can be obtained from the EXIF data. However, EXIF data is not implicit to an image and not readily available for all images. Therefore, for practicality, we do not consider features related to camera settings.

5.2. Features of Salient Regions

We compute the measures of exposure, sharpness, and texture details for the salient regions utilizing the same respective techniques used to compute the global features. The exposure feature for the salient regions is

$$f_{22} = \frac{1}{M}\sum_{m=1}^{M} S_V(m). \tag{4}$$

The saturation for the region, f_{23} is similarly computed. The sharpness features, f_{24} to f_{26}, are computed by applying our enhanced sharpness detection technique on the salient regions of the image. For texture details, we only compute the sum of the average wavelet coefficients over all levels to produce f_{27} to f_{29} for each HSV channel of the salient regions. Another photographic rule, *fill the frame*, suggests that the subject should occupied a large portion of the image. We represent the size of the salient regions by the dimension of the salient regions, and we have $f_{30} = |S_{ch}| = M$.

In addition to the features of the salient regions, we analyze the position, distribution, and the total number of salient locations. A professional photo has a strong focus and the subject can be easily identified. This feature is characterized by a small number and dense distribution of salient locations. We let $f_{31} = |L|$, where L is the number of unique salient locations of the image. To represent the distribution of the salient locations, we compute the standard deviation of all the salient locations, f_{32}.

A saliency map provides additional useful information about the salient regions. In addition to just the locations, the intensity and size of the salient regions represent the degrees of saliency of the corresponding salient regions. We compute the saliency map mean

$$f_{33} = \frac{1}{XY}\sum_{x=1}^{X}\sum_{y=1}^{Y} SM(x,y) \tag{5}$$

and the standard deviation

$$f_{34} = \sqrt{\frac{1}{XY-1}\sum_{x=1}^{X}\sum_{y=1}^{Y}(SM(x,y)-f_{33})^2} \tag{6}$$

to capture the saliency strength information. Comparing the two images in Figure 2, with aesthetics score of 6.56 and 3.83 respectively, image (a) has a total of 7 salient locations compared to 10 in image (b). In addition, image (a) yields smaller scores of 5.6 and 33 for f_{33} and f_{34} compared to scores of 7.8 and 38 obtained for image (b).

5.3. Features Depicting Subject-Background Relationship

From the survey conducted by Yan et al. [9], they concluded that simplicity is the most distinguishing factor of professional photos. They used two *global* image features—the edge spatial distribution and hue count—to measure the simplicity factor. As simplicity of a photo is mostly characterized by a simple background as well as clear contrast between the subject and the background, it would be more intuitive to compute simplicity measure based on the differences between the subject and its background in a number of aspects, such as the exposure, saturation, hue, blurriness, texture details, and edge spatial distribution. For example, difference in hue between the subject and its background represents the color contrast, which is an important photographic rule.

For these set of features, we apply the same methods used for the global image features to compute exposure, saturation, hue, blurriness, and texture details, for both the salient and background regions. We compute the differences of the respective features of the subject and its background using squared difference. For example, the subject-background difference for exposure is

$$f_{35} = \left(\frac{1}{M}\sum_{m=1}^{M} S_{\mathrm{V}}(m) - \frac{1}{N}\sum_{n=1}^{N} B_{\mathrm{V}}(n)\right)^2 . \tag{7}$$

Similarly, f_{36} and f_{37} are computed for subject-background differences in saturation and hue respectively. The sharpness of the subject and background are computed by applying our enhanced sharpness detection technique to both the salient and background regions separately. The subject-background sharpness difference features, f_{38} to f_{40}, are then computed using the squared difference.

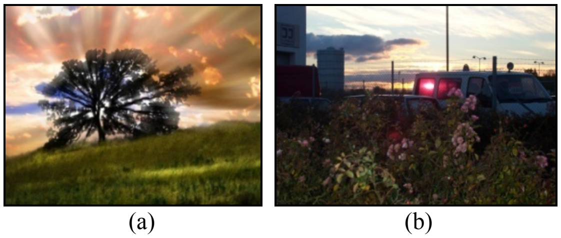

Figure 2. (a) Professional photo with a prominent subject; aesthetics score = 6.46. (b) Snapshot without any prominent subject; aesthetics score = 3.83.

The texture difference between the subject and background, f_{41} to f_{43}, are computed by taking the squared difference between the sum of the average wavelet coefficients over all three levels for each HSV channel of the salient regions and of the background region.

For edge simplicity, we first compute the edge spatial distribution [9] for the both the subject and background separately. We applied Canny edge detector to detect edges and compute the edge distribution by dividing the magnitude of the edge by the size of the salient regions and that of the background region. Then, the edge simplicity feature, f_{44}, is computed by calculating the squared difference between the edge spatial distributions of the subject and its background.

6. CLASSIFICATION

To allow direct comparison with the results of Datta et al. [2], we downloaded the same set of photos used by them from Photo.net. However, since some users have removed their photos from Photo.net, we managed to collect only a subset of 3161 photos out of the 3581 photos used by Datta et al. Each photo has an aesthetics score in the range of one to seven. The mean aesthetics score for this set of photos is 5.1. For training our classifier, we used only the top 10% and bottom 10% photos that have ratings above 6.2 and below 4.0 respectively.

Our feature set F consists of our global image features f_1 to f_{21}, and the salient features f_{22} to f_{44}. To illustrate the effectiveness of our salient region features, we have also created a feature set G by augmenting a selected set of Datta's most discriminative global image features with our set of salient features.

We perform attribute selection and classification on both feature sets using one-dimensional support vector machine (SVM) [7] tool provided by Weka Explorer [8]. We select SVM to build our model because it is a powerful binary classifier and is most appropriate for two-classification task. Instead of using SVM with RBF kernel [2], we chose to perform SVM classification without any kernel because we believe that professional and snapshots classes are linearly separable if our features are discriminative enough.

7. EXPERIMENTAL RESULTS

After performing one-dimensional SVM on feature set G, we obtain a set G_S that contains the top 15 features. Out of these 15 features, there are eight global image features and seven salient features. The top two features are our sharpness features for the salient region. This result shows that salient features do play an important role to differentiate professional photos from snapshots.

For classification, we obtained a 78.2% 5-fold cross-validation (5-CV) accuracy for feature set G_S, with a high class precision of 82.9% and low class precision of 75.6%. For feature set F_S, the accuracy achieved is 78.8%, marginally higher than accuracy of feature set G_S. The precision for professional photos is increased to 83.7% but remains about the same at 75.2% for snapshots. This indicates that our enhanced sharpness and contrast global features are able to increase the discrimination of professional photos.

For comparison with Datta's approach, we run the SVM classifier using standard RBF kernel ($\gamma = 3.7$, cost = 1) on the top 15 features specified in their paper [2]. Since our dataset is the subset of the original dataset used by Datta et al., for a fair comparison, we also perform SVM attribute selection and classification without RBF kernel on their full feature set. Figure 3 shows the comparison of our results with Datta et al.'s [2] and Yan et al.'s [9]. Both of our feature sets, G_S and F_S outperformed existing work across the top and bottom n% datasets. Specifically, for the top and bottom 10% images, our 5-CV accuracy is about 6% to 8% higher than all existing works. Another interesting observation is that our results have much higher stability across the different datasets, maintaining an accuracy of 77% to 78%.

8. DISCUSSIONS AND CONCLUSIONS

Results show that our saliency-enhanced approach is indeed a promising direction for aesthetics class prediction. Our higher-level approach to extract salient region features proves to be more effective than the low-level region-based approach used by Datta et al. [2]. This result is not surprising since aesthetically-pleasing photos tend to direct the focus of attention to the intended subject that normally coincides with the salient locations. Furthermore, in professional photos, the subject normally stands out from the background. Representing the subject with the extracted salient regions and using it to identify the subject-background relationship enable the determination of the level of conspicuity of the subject. All this distinctive information of professional photos can be found in the saliency map as well as in the features extracted based on the salient regions, contributing to significantly higher classification accuracy. The consistency and stability of the performance of both our feature sets across the different image datasets illustrate that our saliency approach is less prone to misclassification due to noise inherent in individual image features.

Despite the promising result of our work, we believe that there are various limitations that can be addressed to further enhance the classification performance. Our salient region extraction relies on a saliency model and image segmentation, both being open problems with results still not truly accord with human perception. There are possibilities in which the extracted salient regions may not represent, or only partially represent the photo subject, causing decreased classification performance. Thus, future improvement in both saliency model and segmentation techniques would likely lead to better classification accuracy. Another limitation of our current work is that the number of salient locations used as seeds to extract regions from a segmented image is fixed to three. There are possibilities that these three seeds may not coincide with the subject, or in cases where the subject has high level of texture details as in low DOF macro images, three seeds are likely not sufficient to fully segment out the entire subject. Thus, finding the optimal number of seeds to be used for salient region segmentation may potentially lead to better performance. Additionally, an area worth investigating is the relationship among multiple salient regions for discovery of discriminative features.

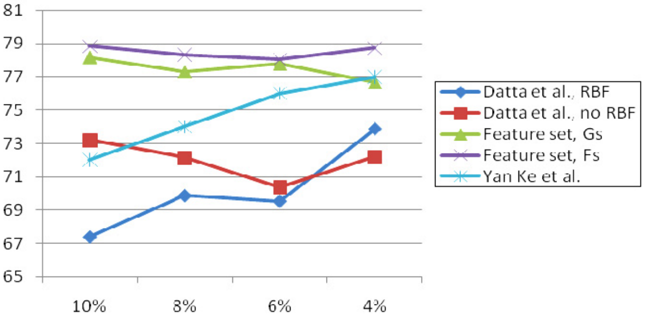

Figure 3. Comparison with existing work on datasets consisting of the top and bottom *n*% of photos.

One area for future work is to extend our saliency-enhanced approach for score prediction and apply it to application areas such as content-based image retrieval. In addition, we will also look into combining our saliency approach with a category-based approach. Different categories of photos have different set of features to determine whether they are good. For example, low DOF is one of the significant features that make good portrait photographs but it is not for landscape photographs. Vice versa, rule of thirds is more significant for landscape photographs than to portraits. Thus, combining both approaches may be a key for better classification performance.

9. REFERENCES

[1] Coe, K. (1992), *Art: The replicable unit - An inquiry into the possible origin of art as a social behavior*, Journal of Social and Evolutionary Systems, 15(2), 217-234.

[2] R. Datta, D. Joshi, J. Li, and J. Z. Wang (2006), *Studying Aesthetics in Photographic Images Using a Computational Approach*, Proc. of European Conference on Computer Vision, 288-301.

[3] R. Datta, J. Li, and J. Z. Wang (2008), *Algorithmic Inferencing of Aesthetics and Emotion in Natural Images: An Exposition*, Proc. of IEEE International Conference on Image Processing, Special Session on Image Aesthetics, Mood and Emotion, 105-108.

[4] Itti, L., Koch, C., Niebur, E. (1998), *A Model of Saliency-Based Visual Attention for Rapid Scene Analysis*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 20(11), 1254-1259.

[5] Lind, R. W. (1980), *Attention and the Aesthetics Object,* Journal of Aesthetics and Art Criticism, 39(2), 131-142.

[6] Tong, H.,, Li, M., Zhang, H., He, J., and Zhang, C. (2002), *Classification of Digital Photos Taken by Photographers or Home Users*, Proc. of Pacific-Rim Conference on Multimedia. 367-376.

[7] Wang, J.Z., Li, J., and Wiederhold, G., (2001), SIMPLIcity: Semantics-Sensitive Intergrated Matching for Picture Libraries, IEEE Transactions on Pattern Analysis and machine Intelligence, 23(9), 947-963.

[8] Witten, I. H., and Frank, E., (2005), *Data Mining: Practical machine learning tools and techniques*, 2nd Edition, Morgan Kaufmann, San Francisco, 2005.

[9] Yan Ke, Xiaoou Tang, Feng Jing (2006), *The Design of High-Level Features for Photo Quality Assessment*, Proc. of Computer Vision and Pattern Recognition, pp. 419-426.

[10] Yang, A., Wright, J., and Yi, M. (2008), *Unsupervised segmentation of natural images via lossy image compression*, Computer Vision and Image Understanding, 110(2), 212-225.