Predicting the Memorability of Natural-scene Images

Jiaxin Lu [1], Mai Xu [1], Zulin Wang [1,2]

[1]*The School of Electronic and Information Engineering, Beihang University, Beijing, China*
[2]*Collaborative Innovation Center of Geospatial Technology, Wuhan, China*
Corresponding author: MaiXu@buaa.edu.cn

***Abstract*—Recent work has shown that image memorability, in general, can be reliably predicted using some state-of-the-art features. However, all existing methods are not effective in predicting memorability of natural-scene images, far from human. In this paper, we propose a novel method to improve the effectiveness of memorability prediction for natural-scene images. Specifically, we argue that some of HSV colors have either positive or negative impact on memorability of natural-scene images in our Natural-Scene Image Memorability (NSIM) dataset. Then, we develop an HSV-based feature for memorability prediction. Finally, the HSV-based feature is combined with other efficient state-of-the-art features in our approach to predict memorability on natural-scene images. Experimental results validate the effectiveness of our method.**

***Index Terms*—Image analysis, memorability, HSV**

I. Introduction

When exposed to the overflow of visual information, people tend to have the splendid ability to recall thousands of pictures after only a glance [1]. In fact, not all pictures are remembered equally in human brain. Some might be stored in one's mind for a long period, while others fade away in a short time. The reasons why people have the intuition to remember pictures are varied. Some memorable pictures might contain familiar things or events relevant to one's life experience, whereas some are not but still stick in people's mind. Those complicated brain behaviors arouse increasing attention on the field of image memorability. The research on image memorability prediction may have a prospect of application in commercial advertisement, education, medicine and etc.

Recent studies have made some remarkable achievements on how the intrinsic and extrinsic image properties influence the memorability of different scenes. Specifically, the works of [2], [3], [4] have shown that image memorability is an intrinsic property of an image that can be estimated by the state-of-the-art features. In [5], [6], Khosla *et al.* combined local and global features together to predict the memorability scores, and mentioned that it has potential to modify memorability of image region. In addition, Mancas *et al.* [7] argued that the features of image saliency can be added in [4] to improve the accuracy of memorability prediction, for natural images with salient objects. However, the non-object scenario has not been considered in [7]. Besides, [8], [9] focused on the memorability of face photographs, while [10] shed light on the various factors and properties that make an object memorable. Most recently, [11] has been proposed to apply fine-tuned Hybrid-CNN to learn adaptive features for predicting the memorability of generic images. Furthermore, Bylinskii *et al.* [12] aimed to dig out how the context of images changes memorability, highlighting that observers' eye movements provide additional information to predict the memorability of images.

From [10], it is evident that the high memorable object and its category can serve as a strong priority to predict the memorability scores. But what if there is no object in the scenery images? As found in this paper, when predicting the memorability of non-object images, named *natural-scene images*[1] here, the state-of-the-art memorability prediction is far from human. The latest work of [10] has also pointed out that memorability prediction of natural-scene images is an important research direction.

Fig. 1 demonstrates that the low-level features of color are highly correlated with memorability of natural-scene images. Motivated by this observation, we propose in this paper a novel approach to predict memorability of natural-scene images by exploring the HSV-based feature. To our best knowledge, this paper is the first work to concentrate on memorability prediction of natural-scene images, seen as a subset of all generic images. More importantly, our approach develops efficient low-level features for bottom-up prediction of image memorability. The main contributions of this paper are summarized as follows:

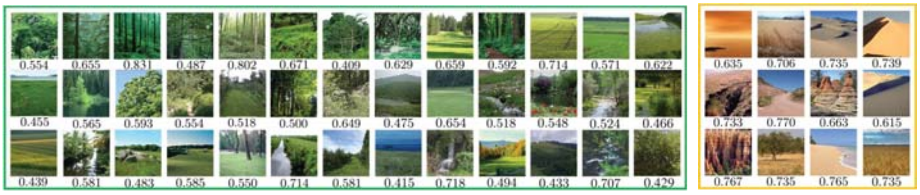
- We establish a dataset, which includes the ground-truth memorability of 258 non-object images, for memorability prediction of natural-scene images.
- We find that some of HSV colors have high correlation with human memorability, through analysis on our established dataset.
- We propose a memorability prediction approach for natural-scene images, which is based on the HSV-based feature.

II. Dataset Analysis

Although we have intuition that the memorability of natural-scene images is correlated with their colors. To shed light

[1] In this paper, we refer natural-scene images as the natural calibrated images, which are without any salient object such as people, animals, man-made objects.

978-1-5090-5316-2/16/$31.00 © 2016 IEEE



(a) Green-oriented (b) Orange-oriented

Fig. 1. The memorability scores of some green-oriented and orange-oriented images. The number below the images indicates ground-truth memorability of (a) and (b) from our dataset to be discussed in Section II. All green-oriented and orange-oriented images, in which green or orange occupies above 50 percent of total pixels, are selected from our dataset and shown in this figure. It can be seen clearly that orange component of HSV-based feature has high correlation with memorability, as their memorability scores are around 0.7. In contrast, the green has weak correlation as memorability scores fluctuate strongly from 0.4 to 0.9.

on the intrinsic relationship between memorability of natural-scene images and low-level features, we introduce the Natural-Scene Image Memorability (NSIM) dataset (see Fig. 1 for some images). Our NSIM dataset contains 258 natural-scene images with their ground-truth of memorability scores, which are selected from [4]. These data allow the analysis on memorability of natural-scene images.

A. NSIM Dataset

To our best knowledge, there exists no dataset on memorability of natural-scene images. Therefore, we established a natural-scene image memorability dataset, namely NSIM dataset, for analyzing humans' memorability on natural scenes. For our NSIM dataset, 258 images with only natural-scenes were selected from the existing memorability dataset [4] by the following way. First, subjective selection was carried out on [4] to find natural-scene images. Specifically, three volunteers participated in selecting natural-scene images. They were asked to pick out natural-scene images according to the following criteria: (1) Each image is comprised by outdoor natural scenes; (2) Each image should not include any salient object, which refers to human, animals and man-made object.

Then, each of three volunteers scanned all 2222 images of [4] in a random order, and annotated whether the image belongs to natural-scene or not. Afterwards, the voting mechanism was applied to determine natural-scene images, which were selected as natural-scene images by at least two volunteers. Finally, 258 images in total were chosen from [4], labelled as natural-scene images. Note that we also have the ground-truth memorability scores of these 258 images by 50 subjects, based on [4].

B. Analysis on human consistency

We first analyze the human consistency on the memorability of natural-scene images. Fig. 2 shows how image scores measured on the participants of Group 1 match image scores given by the participants in Group 2. We can observe from this figure that the memorability of images predicted by Group 2 is indeed similar to that predicted by Group 1. Therefore, these two groups of subjects are highly consistent on memorability of images from our NSIM dataset. To further quantify human consistency, we find that the *Spearman's rank correlation*[2] (ρ) is 0.697 between two sets of scores for Groups 1 and 2. It is worth pointing out that the Spearman's rank correlation between random prediction scores and human prediction scores of Group 1 is 0.003. This again indicates the high consistency on memorability of natural-scene images from our dataset. Thus, it is promising to predict memorability of natural-scene images by exploring some specific features and methods. In the following, we find that the memorability of natural-scenes is correlated with its colors.

C. Analysis on correlation between memorability and color

Since natural-scene images are lack of salient objects, the low-level features are dominant in determining the memorability of such images. Fig. 1 has shown that there exists some correlation between memorability and color for natural-scene images. We further clarify this correlation in Fig. 3, which plots the memorability scores of images with natural-scene being at different colors (i.e., red, orange, green and blue). Note that the results of Fig. 3 are based on all images of our NSIM dataset. As seen from this figure, red color has high correlation, while green has rather low correlation evaluated by Spearman's rank correlation.

Furthermore, we quantify in Fig. 4 the correlation between 12 colors and memorability by analyzing all images of our NSIM dataset. Here, we follow the way of [13] to choose 12 colors, which divides the whole color space of HSV into 12 non-overlapping subspaces. It is evident that different colors have different impact on image memorability. For example, it is more easy to remember red scenes, whereas the black scenes are more likely to be forgotten. On the other hand, the colors of yellow and green have little impact on the memorability of natural-scene images. Therefore, it is necessary to take into account some specific colors in predicting memorability of natural-scene images. In the next section, we propose a method to explore HSV colors in the memorability prediction of natural-scene images.

III. The proposed method

A. Overview of HSV color

HSV [14] is the most common cylindrical-coordinate representation of colors. HSV color space can be transformed from RGB color space by the nonlinear operation. The HSV model re-arranges the geometry of RGB in attempt to be more effective than cube representation of RGB in color characterization. As a result, HSV model is superior to RGB model in segmenting the whole color space.

HSV model contains three variants of hue, saturation and value. The hue (H) of a color refers to the pure color it resembles. The saturation (S), so called tints, describes how white the color is. The value (V), means the brightness of a

[2] Spearman's rank correlation: reflects the correlation of two ranked vectors in the same dimension.

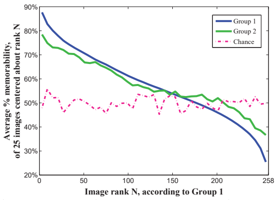


Fig. 2. Measures of human consistency on natural-scene images memorability. The memorability scores are derived from two groups of participants. Images are ranked by memorability scores of participants in Group 1 and plotted against the average memorability scores given in two groups respectively. For clarity, we convolute the resulting plots with a length-6 box filter along the horizonal axis. The chance line is given by allocating random prediction scores as a reference.

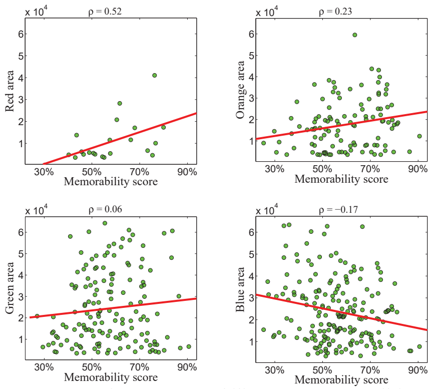


Fig. 3. Color area versus image memorability score among all images with featured color occupying over 5% of total pixels. The red lines are fitting lines, which show the correlation between memorability and color features. For instance, in four pictures, red color has a higher positive correlation with memorability, while green color has a lower positive correlation with memorability.

color. Next, we propose to cluster HSV colors of all natural-scene images from our NSIM dataset, in order to provide low-level features for memorability prediction.

B. HSV-based feature

As analyzed in Section II-C, color features appear to be useful in determining how easy a natural-scene image is to be recalled. Here, we develop HSV-based feature, which makes use of colors as low-level features to predict memorability of natural-scenes.

Assume that $\mathbf{p}_{i,j}$ is the HSV vector of the (i,j)-th pixel in an image. For the t-th training image, we need to identify its primary colors upon HSV vectors of all pixels $\{\mathbf{p}_{i,j}\}_{i=1,j=1}^{I,J}$. Note that the dimension of training image is $I \times J$. Here, we apply the K-means algorithm [15] to cluster $\{\mathbf{p}_{i,j}\}_{i=1,j=1}^{I,J}$ of all training images. Then, we have the centroid set $\mathbf{C}_t = \{\mathbf{C}_{s,t}\}_{s=1}^{S_t}$, in which S_t is the number of primary colors in the t-th training image.

Next, we learn to segment the HSV color space, based on the estimated primary colors of all training images. To be more specific, we define the set of primary colors as $\mathbf{C} = \{\mathbf{C}_t\}_{t=1}^{T}$, for all T training images. Then, the K-means algorithm is again used to cluster primary color set of $\mathbf{C}$, such that the centroid set of HSV space can be obtained. Let $\{\mathbf{o}_l\}_{l=1}^{L} = \{(h_l, s_l, v_l)\}_{l=1}^{L}$ be centroids of segments in HSV color space, where L is the number of segments. We can calculate within-class scatter [16] of each HSV segment, denoted as μ_l, to model the density of primary colors in each segment.

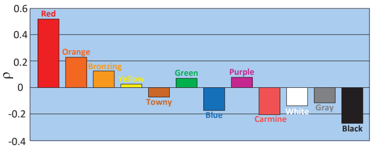


Fig. 4. Measures of Spearman's rank correlation between color and memorability. We spilt the whole color space of HSV into 12 non-overlapping subspace according to [13], and all 258 images are classified to one or more group of these colors. Note that an image can be classified as one color only if more than 5% pixels of the image fall into this color. Then, the Spearman's rank correlations for different colors are arranged in the memorability of the corresponding groups of images, and plotted in the figure.

The thresholds for segmenting HSV color space can be computed by taking into account both centroid and density of each HSV segment. Assume that $\mathbf{o}_l$ and $\mathbf{o}_{l+1}$are centroids of the l-th and $(l+1)$-th segments, which are neighboring in HSV space. The hue of HSV segmentation thresholds can be estimated by

$$\mathbf{b}_l = \frac{\mathbf{o}_l \cdot \mu_{l+1} + \mathbf{o}_{l+1} \cdot \mu_l}{\mu_l + \mu_{l+1}} \tag{1}$$

Given $\mathbf{b}_l$, we are able to divide the HSV color space into L segments.

According to thresholds $\{\mathbf{b}_l\}_l^{L-1}$, we can extract L-dimensional HSV-based feature for an input image as follows. Let $\mathbf{X}$ be the input image, with $\mathbf{X}_{i,j}$ being its pixels. Its HSV-based feature can be represented by $\mathbf{u} = (\mathbf{u}_1, \cdots, \mathbf{u}_l, \cdots, \mathbf{u}_L)$. Then, $\mathbf{u}_l$ can be calculated by counting for the proportion of pixels, the HSV of which fall into the range of $[\mathbf{b}_l, \mathbf{b}_{l+1}]$. This way, we can obtain the HSV-based feature $\mathbf{u}$ for an input image.

C. Memorability prediction with HSV-based feature

For memorability prediction, we need to integrate our HSV-based feature with other effective features of [4]. Assume that $f(\cdot)$ is a prediction function, mapping image features to memorability scores. The n-th feature is denoted as $\mathbf{a}_n$, where $\mathbf{u}$ is seen as $\mathbf{a}_1$ and other features are $\{\mathbf{a}_n\}_{n=2}^{N}$. Here, we learn to map from features of $\mathbf{a}_n$ to m, by training a Support Vector Regression (SVR) classifier on the training data. Finally, memorability for the input image can be predicted by

$$m = f(\mathbf{a}_1, \cdot \cdot \cdot, \mathbf{a}_n, \cdot \cdot \cdot, \mathbf{a}_N), \tag{2}$$

where m is the memorability score of the image.

We empirically find that the features of GIST, HOG2x2, pixels, scene categories and spatial histograms from [4] are all effective in predicting the memorability of natural-scene images. Thus, they are also selected for features $\{\mathbf{a}_n\}_{n=2}^{N}$ in our method. Their effectiveness is to be discussed in Section IV, compared with that of our HSV-based feature. For the kernel of SVR, we use the linear kernel function in HSV-based feature, as it is simple yet effective (verified by experiments). Fig. 5 summarizes the framework of our method.

IV. Experimental results

Setting. For the HSV-based feature, the primary color number S_t is chosen to be $4-7$, according to image content. In addition, we set the HSV segment number L to be 12 in our experiment, to make the results appropriate. For memorability prediction, we follow the settings of [4] in the SVR classifier. Here, our database was equally divided into non-overlapping training and test sets by random, the same as [4]. Also 50 subjects were split into two non-overlapping groups, i.e.,

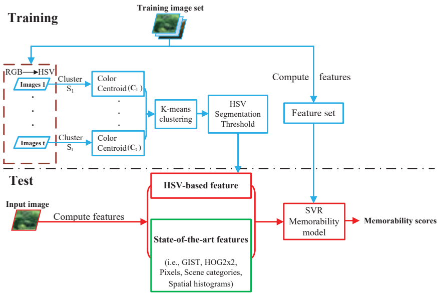


Fig. 5. Framework for our method.

TABLE I
THE COMPARISON BETWEEN OUR METHOD AND CONVENTIONAL METHODS

	Our method	[4]	[11]
ρ	**0.466±0.068**	0.424±0.070	0.404±0.069

Groups 1 and 2. Note the memorability scores of one subject from Group 1 were used as the ground truth of the training set, while the scores of another subject from Group 2 were utilized for the test set. Then, this process was repeated for all 50 subjects, with our database randomly divided for 25 times. The averaged results are reported in the following.

Evaluation. We compute the Spearman's rank correlation between ground truth and test results to evaluate the accuracy of natural-scene image memorability prediction. Here, we set Isola *et al.* [4] and Khosla *et al.* [11] model as baseline prediction model, and we compare accuracy of our method with them. Note that all methods share the same training and test sets for fair comparisons.

We report in Table I the Spearman's rank correlation coefficients ρ, which are obtained by 25 times validation. This table demonstrates that our method is superior to other two methods, in terms of both mean and standard deviation (among 25 times test). Note that the averaged results of [11] performs even worse than [4]. This is probably because the CNN of [11] is not effective in natural-scene images, in which salient object can be hardly found.

In addition, Fig. 6 shows the Spearman's rank correlations of each single feature and feature integration. We can see from this figure that our HSV-based feature ranks the second among all features for memorability prediction, which is slightly worse than GIST. Thus, the effectiveness of the single HSV-based feature can be validated. Besides, one may observe from this figure that adding HSV-based feature can increase ρ by 7.3%[3]. Furthermore, it is obvious that our method outperforms other conventional methods. This verifies the effectiveness of our HSV-based feature in feature integration.

V. Conclusion

In this paper, we have proposed a memorability prediction approach for natural-scene images. Interestingly, we found that human beings are capable of remembering natural-scene images, far better than the state-of-the-art memorability prediction approaches. Thus, it is necessary to develop the effective low level features for memorability prediction of natural-scene images, as there is no salient object in this kind of images. To this end, we investigated that memorability of natural-scene images is highly correlated with some colors. Thus, we developed the HSV-based feature in light of segmented HSV color space. Then, our approach used the developed HSV-based feature and other existing features for predicting image memorability. Finally, the experimental results demonstrated that our approach is superior to other state-of-the-art approaches in memorability prediction of natural-scene images, evaluated in terms of Spearman's rank correlation.

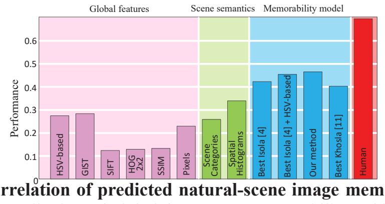


Fig. 6. **Rank correlation of predicted natural-scene image memorability.** The pink region shows the contribution of global features to memorability, while the green region demonstrates the contribution of scene semantic attribute to memorability. Additionally, the blue region shows the results of feature integration. The difference between our method and Best Isola+HSV-based is that SSIM and SIFT are discarded in our method.

VI. Acknowledgement

This work was supported by the NSFC projects under Grants 61573037 and 61471022, and the Fok Ying Tung Education Foundation under Grant 151061.

References

[1] L. Standing, "Learning 10000 pictures," *The Quarterly journal of experimental psychology*, vol. 25, no. 2, pp. 207–222, 1973.

[2] P. Isola, D. Parikh, A. Torralba, and A. Oliva, "Understanding the intrinsic memorability of images," in *Advances in Neural Information Processing Systems*, 2011, pp. 2429–2437.

[3] P. Isola, J. Xiao, D. Parikh, A. Torralba, and A. Oliva, "What makes a photograph memorable?" *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 36, no. 7, pp. 1469–1482, 2014.

[4] P. Isola, J. Xiao, A. Torralba, and A. Oliva, "What makes an image memorable?" in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 145–152.

[5] A. Khosla, J. Xiao, P. Isola, A. Torralba, and A. Oliva, "Image memorability and visual inception," in *SIGGRAPH Asia 2012 Technical Briefs*. ACM, 2012, p. 35.

[6] A. Khosla, J. Xiao, A. Torralba, and A. Oliva, "Memorability of image regions," in *Advances in Neural Information Processing Systems*, 2012, pp. 305–313.

[7] M. Mancas and O. Le Meur, "Memorability of natural scenes: The role of attention," in *Image Processing (ICIP), 2013 20th IEEE International Conference on*. IEEE, 2013, pp. 196–200.

[8] W. A. Bainbridge, P. Isola, I. Blank, and A. Oliva, "Establishing a database for studying human face photograph memory," in *Proc. 34th Annu. Meeting Cognit. Sci. Soc*, 2012, pp. 1302–1307.

[9] A. Khosla, W. Bainbridge, A. Torralba, and A. Oliva, "Modifying the memorability of face photographs," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 3200–3207.

[10] R. Dubey, J. Peterson, A. Khosla, M.-H. Yang, and B. Ghanem, "What makes an object memorable?" in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1089–1097.

[11] A. Khosla, A. S. Raju, A. Torralba, and A. Oliva, "Understanding and predicting image memorability at a large scale," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 2390–2398.

[12] Z. Bylinskii, P. Isola, C. Bainbridge, A. Torralba, and A. Oliva, "Intrinsic and extrinsic effects on image memorability," *Vision research*, vol. 116, pp. 165–178, 2015.

[13] R. C. Gonzalez and R. E. Woods, "Digital image processing," 2002.

[14] A. Hanbury, "Constructing cylindrical coordinate colour spaces," *Pattern Recognition Letters*, vol. 29, no. 4, pp. 494–500, 2008.

[15] J. A. Hartigan, "Clustering algorithms," 1975.

[16] N. Otsu, "A threshold selection method from gray-level histograms," *Automatica*, vol. 11, no. 285-296, pp. 23–27, 1975.

[3] Here, the percentage is calculated by the percentage of increment of [4] plus HSV-based feature (0.451) over [4] (0.424).