Relative Spatial Features for Image Memorability

Jongpil Kim*
jpkim@cs.rutgers.edu

Sejong Yoon*
sjyoon@cs.rutgers.edu

Vladimir Pavlovic
vladimir@cs.rutgers.edu

Department of Computer Science
Rutgers, The State University of New Jersey
Piscataway, NJ 08854-8136

ABSTRACT

Recent studies in image memorability showed that the memorability of an image is a measurable quantity and is closely correlated with semantic attributes. However, the intrinsic characteristics of memorability are not yet fully understood. It has been reported that in contrast to a popular belief unusualness or aesthetic beauty of the image may not be positively correlated with the image memorability. This counter-intuitive characteristic of memorability hinders a better understanding of image memorability and its applicability. In this paper, we investigate two new spatial features that are closely correlated with the image memorability yet intuitively explainable. We propose the Weighted Object Area (WOA) that jointly considers the location and size of objects and the Relative Area Rank (RAR) that captures the relative unusualness of the size of objects. We empirically demonstrate their useful correlation with the image memorability. Results show that both WOA and RAR can improve the memorability prediction. In addition, we provide evidence that the RAR can effectively capture object-centric unusualness of size.

Categories and Subject Descriptors

I.2.10 [**Artifical Intelligence**]: Vision and Scene Understanding—*modeling and recovery of physical attributes; representations, data structures, and transforms; perceptual reasoning*

Keywords

image memorability; feature extraction; relative correlation

1. INTRODUCTION

How memorable is an image? Some images are clearly more memorable than others. This is especially true for the images with known content, e.g. photos of a family or friends, attended events or visited places. However, studies show that some images are intrinsically more memorable than others without such subjectively recognizable content [2, 11]. According to recent studies, the latter case of memorability is measurable [9], closely correlated with semantic attributes [8] and even predictable [9, 10] to some extent. One key beneficiary of the image memorability prediction may be the advertisement industry, which can use the memorability score to quantitatively measure the effectiveness of prototype designs of advertisements for various consumer items. Moreover, predicting memorability may also affect a system's ability to understand scenes or improve object recognition.

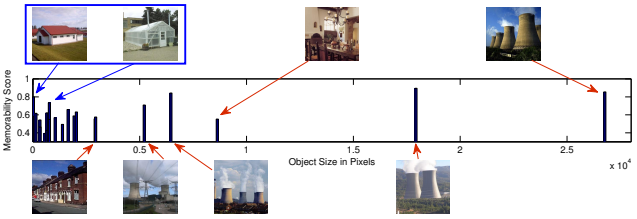

Figure 1: Image memorability score [9] against all 25 images containing the object *chimney*. Images with unusually sized, center-concentrated chimneys (1st, 2nd, 4th and 5th from the right) are more memorable than images with normally sized chimneys (6th from right). 3rd chimney is an indoor scene. Two outliers inside the blue box show that their high memorabilities are not directly related to the chimneys.

However, the natural characteristics of image memorability are not fully understood. Spatial and content-based attributes such as "people with visible faces" and "enclosed spatial structure" are reported to positively affect the memorability scores, while unusualness and aesthetic beauty of an image negatively correlate with the memorability scores [8]. The fact that the unusualness negatively correlates with the memorability score is especially counter-intuitive as we can easily find highly memorable images depicting unusual scenes, as depicted in Fig. 1. Qualitative analysis of these images seems to indicate that what correlates with image memorability is *not* the unusualness as a holistic attribute of an image *but* the unusualness of each object relative to its expected size and location within the image's semantic context. This hypothesis can be roughly supported by the fact that the image memorability is closely correlated with the spatial object size and semantics [9]. Nevertheless, a relationship between the relative unusualness of an object and the memorability has not been properly established.

The goal of this paper is to investigate relative size and location-based features both closely correlate with the image memorability and intuitively explainable for a better under-

*J. Kim and S. Yoon equally contributed to this work.

MM'13, October 21–25, 2013, Barcelona, Spain.

ACM 978-1-4503-2404-5/13/10 ...$15.00.
http://dx.doi.org/10.1145/2502081.2502198.

standing of the image memorability. We propose two spatial features in this paper: the Weighted Object Areas (WOA) and the Relative Area Rank (RAR). Exploiting the same experimental setting as in [9], we show that these basic spatial features are closely correlated with the image memorability. In addition, we provide evidence that RAR can effectively capture the object-centric unusualness defined as the object size realtive to the expected size of the object's class. We discuss possible extentions of the spatial features to the image memorability prediction in the concluding remark.

2. RELATIVE SPATIAL FEATURES

In this section, we introduce two relative spatial features for the image memorability measure, the Weighted Object Area and the Relative Area Rank. Both features are designed to encode spatial characteristics (size and position) of objects in an image as related to image memorability.

2.1 Weighted Object Areas

Size and location of an object in an image are two important factors to determine the object's importance [1]. Objects closer to the center and larger in size tend to be more important as they have higher probabilities of being mentioned by annotaters [1]. While the connection between the importance of objects in an image and the image memorability is unclear, object size and spatial location do have a strong correlation with the memorability [9]. Thus it is natural to presume that the two spatial characteristics of objects need to be treated jointly, with the location information being considered as relative to the center. However, previous studies on memorability used the location information of the object only by considering multiscale area coverage, i.e. we additionally consider the object's subareas covered in the quadrant. While this spatial pyramid encodes the location information, it only captures the existence with limited directional information. Moreover, it does not consider the relative displacement of the object from the center of the image.

Therefore, we introduce a new feature which jointly considers the size and location in an intuitive manner. The basic idea is simple: We give more weight to pixels around the center while reducing the weight of pixels closer to image boundaries. The bivariate Gaussian function over the image pixel locations natually conforms to the criteria. Formally, the weight for a pixel located at a pixel coordinate $\mathbf{x}$, $\mathbf{x} \in \mathbb{R}^2$ is defined as follows: $w(\mathbf{x}) \sim \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\mathcal{N}(\cdot)$ is the bivariate Gaussian probability density function with the vector $\boldsymbol{\mu} \in \mathbb{R}^2$ and the matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{2\times 2}$ as its mean and covariance, respectively. To give the largest weight at the center point, we set $\boldsymbol{\mu}$ to be the location of the image center. $\boldsymbol{\Sigma}$ is the covariance of the weight distribution and we do grid search and cross validation to find it. Then, for a given image Q, the weighted object area (WOA) histogram is calculated as follows:

$$\mathrm{WOA}(Q,k) = \sum_{\mathbf{x}\in \mathrm{OBJ}(Q,k)} w(\mathbf{x}), \qquad (1)$$

where k is an object index, and $\mathrm{OBJ}(Q,k)$ is the set of all pixels of object k in image Q. Clearly, when an object is larger or closer to the center of an image, the WOA for the object is higher. Moreover, the same sized objects may have different WOA values since the object closer to the center has a higher weight than objects closer to image boundaries.

2.2 Relative Area Rank

Earlier studies on image memorability demonstrated that natural scenes are less likely to be remembered [8, 9]. However, little has been known about the underlying reason of this unusual finding. One possible explanation of this phenomenon is that natural scene image components, such as skies, mountains, ground, trees, etc., in many cases do not vary in *relative image size*. Unlike other objects of varying sizes including the most memorable object, a person, the natural scene image components are expected to cover a certain portion of an image. Thus, putting it another way, the lack of unusualness of those natural scene components negatively affected the image memorability.

To verify our intuition, we need a way to capture the unusualness of each object size within an image compared to the object class' expected coverage. Note that considering the absolute size of each object and comparing the value with its own class' size distribution would make little sense since different images have different context and an object's coverage in an image may vary depending on the context. Therefore, we need a measure to capture this *relative coverage* of objects.

In image annotation literature, Hwang and Grauman [7] introduced the relative rank to capture the prominence of an object name by its order of annotation in the tag list. The relative rank is defined as the percentile of the rank for a tag in the given image, relative to all the ranks in the training images with the same tag. The higher the rank value, the more the tag climbs to the top of the list relative to where it typically occurs in any other list.

In this paper, we employ the relative rank to account for relative change of the object area from its expected area distribution. Similar to [7], we define the relative area rank of object k as the percentile of the rank for object k to all the ranks in the training images for that object label. Formally, the relative area rank is defined as follows. Let $S(Q,k)$ be the size of object k. Then, the relative area rank (RAR) of object k is

$$\mathrm{RAR}(Q,k) = \frac{\sum_{i\in T} I\left(S(Q_i,k) \le S(Q,k)\right)}{|T|}, \qquad (2)$$

where T is the training data, $I(\cdot)$ is the indicator function and Q_i is the i-th image in the training data T. If the object k is not present in the image, we set the $\mathrm{RAR}(Q,k)$ to be 0. We next present empirical evidence to show that RAR is an effective feature to predict the memorability score. In this paper, we scale the size of all images to a same size in order to avoid the normalization issue due to image size difference.

3. EXPERIMENTAL SETTINGS AND RESULTS

In this section, we explain our experiment setting and demonstrate the utility of the proposed WOA and RAR features, as a proxy to predicting the image memorability.

3.1 Data and Experimental settings

We examine the correlation of the two proposed measures with the memorability score of the memorability dataset provided by [8] and [9]. The dataset has 2222 images randomly sampled from the SUN dataset [15]. As a pre-processing, all the images are cropped and scaled to 256×256 pixels. The memorability scores for the images are calculated us-

Figure 2: Upper panel: Top 2 images for which WOA and RAR reduce the prediction errors compared to `Object`, respectively. Bottom panel: Bottom 2 images for which the WOA and RAR increase the prediction errors compared to `Object`, respectively.

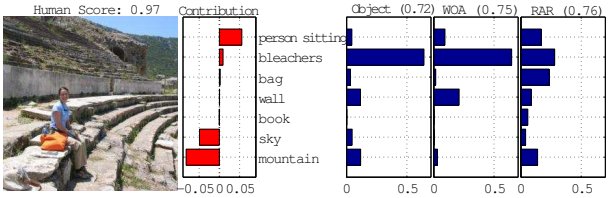

Figure 3: An example image where RAR gives the best prediction results. Three histograms on the right correspond to `Object` (left), WOA (center) and RAR (right) with numbers in parenthesis showing predicted memorability. Horizontal axis is the normalized size. Red plot on the left shows the averaged object-wise contribution to the memorability. Note that objects with large positive contribution gained more normalized area, leading to the better prediction.

ing Amazon Mechanical Turk with a "visual memory game". The game shows participants a sequence of images and asks them to press a space bar whenever they see a repeated image. After collecting the responses from the participants, the memorability score is calculated by the percentage of correct detections. To evaluate the proposed method, we follow the same process as [9]; we used the averaged Spearman's rank correlation (ρ) as the evaluation measure over 25 random split trials. This measure computes the correlation between the ranks of predicted and ground truth memorability scores.

3.2 Results

We compared the proposed features to previously used features, `Object`, `Scene` and `Attributes` which were used in [8, 9]. `Object` is a histogram of the labeled object areas over the spatial pyramid representation (i.e. labeled multiscale object areas). In `Object`, each bin corresponds to one object. `Scene` is a scene category label assigned to the image. Because all images used in the experiments are in a subset of the SUN database, every image has a scene category label. `Attributes` is a collection of visual attributes, not only human-understandable but also highly informative for the memorability [5, 8]. `Attributes` data is built by Amazon Mechanical Turk workers. We trained a support vector regression (ϵ-SVR [3]) to predict the memorability as in [9]. Histogram intersection kernels are used for `Object`, `Scene` and the proposed features, and a Radial Basis Function (RBF) kernel for `Attributes`. We also evaluated combinations of the features using a kernel sum.

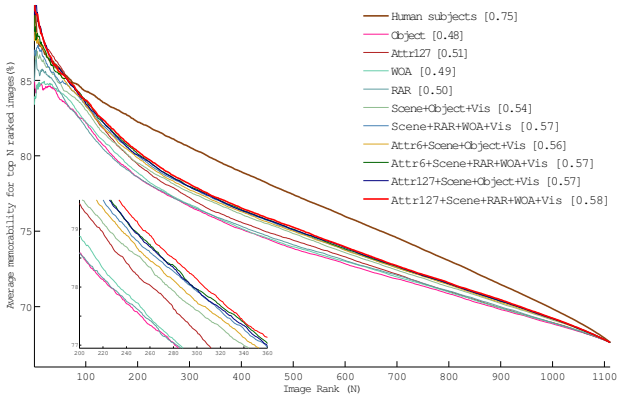

Figure 4: Comparision of prediction results averaged over 25 random splits. Images are sorted by predicted memorability scores and plotted against the cumulative ground truth memorability scores. Best viewed in color.

Table 1 shows the prediction results of the memorability for `Object`, `Scene`, `Attributes`, WOA and RAR. `Human Subjects` denotes the prediction results by people. The proposed features outperform `Object` and `Scene` used in [9]. Surprisingly, RAR is almost the same as the combination of `Object and Scene` by itself. Using all 127 attributes from `Attributes` shows the best performance. However, WOA/RAR are better than the top 6 best attributes proposed in [8]. Fig. 2 shows example images when WOA/RAR improved/degraded the prediction performance compared to `Object`.

We conducted further experiments using combinations of the considered features including additional global visual features. The visual features are Pixel histograms, GIST [13], SIFT [12], HOG2 $\times$ 2 [4, 6, 15] and SSIM [14]. We used the RBF kernel for GIST and the histogram intersection kernels for the other features as in [9]. Table 2 shows the results of predicting memorability using this large set of visual features, coupled with attributes. `VIS` denotes the combination of all the visual features using the kernel sum. We sought to substitute the `Object` with WOA in the best known combination of `Object`+`Scene`+`VIS` and `Attributes`. The result shows that WOA can constantly boost the available approach, from 0.56 to 0.57 and 0.57 to 0.58 correlation, for 6 and all 127 attributes, respectively. RAR showed little improvement in terms of the rank correlation but as shown in Bottom 20/100 rows of Table 2, RAR shows the larger improvement for the prediction of unmemorable images than Top 20/100 rows, i.e. it can classify unmemorable images. This confirms our intuition claimed in Section 2.2. Note that in the absense of attributes that requires the expensive high level human annotation, WOA and RAR have improved the prediction as from 0.50 to 0.52 without `VIS` and from 0.54 to 0.57 with `VIS`. Fig. 4 shows the comparison between `Object` and the proposed features pictorially.

Fig. 3 depicts another evidence in support of RAR. This image is the 3rd most memorable image in the dataset. However, using only the multiscaled area does not take the person in the image into account enough due to overwhelming size of *bleacher*. However, with RAR, the person's size is boosted due to the other object sizes being decreased, leading to improved overall memorability prediction.

Table 1: Results of memorability predictions on various features. Four rows with percentages show the average memorability scores (Recall that the memorability score is calculated by the percentage of correct detections). Top 20/100 row shows the average memorability score over the images of top 20/100 images with the highest predicted memorability scores. Bottom 20/100 row shows the average score over the bottom 20/100 images with the lowest predicted memorability scores. Thus, for Top 20 and Top 100, higher the better while for Bottom 20 and Bottom 100, lower the better. See [9] for details.

	[9]			[8]		Proposed		[9]
	Object	Scene	Object + Scene	Attributes (6)	Attributes (127)	WOA	RAR	Human Subjects
Top 20	84.4%	81.6%	84.5%	85.2%	87.6%	84.9%	85.5%	86.9%
Top 100	82.2%	77.3%	82.3%	82.5%	83.4%	82.2%	81.9%	84.3%
Bottom 100	56.2%	57.5%	55.3%	55.5%	55.3%	55.6%	55.6%	46.9%
Bottom 20	52.7%	56.2%	51.7%	55.6%	52.3%	51.2%	52.0%	39.6%
ρ	0.48	0.36	0.50	0.48	0.51	0.49	0.50	0.75

Table 2: Results of memorability predictions on combination of features. Attr127 used all 127 attributes from [8] while Attr6 used only 6 features suggested therein. Note that using features of [8], we can only obtain $\rho = 0.53$ even with the global visual features. Numbers at the bottom of each top and bottom box section show the improvements of average memorability score using our features over traditional feature `Object`. Note that Bottom shows larger improvements than Top for RAR.

	Without Attributes				Top Attributes (6)				All Attributes (127)			
	Object + Scene	WOA+RAR + Scene	Object + Scene + VIS	WOA+RAR + Scene + VIS	Object + Scene + VIS	WOA + Scene + VIS	RAR + Scene + VIS	WOA+RAR + Scene + VIS	Object + Scene + VIS	WOA + Scene + VIS	RAR + Scene + VIS	WOA+RAR + Scene + VIS
Top 20	84.5%	85.2%	86.1%	**86.5**%	87.0%	86.9%	**87.4**%	87.1%	87.5%	87.2%	**87.8**%	87.4%
Top 100	82.3%	82.7%	83.0%	**83.3**%	83.3%	83.5%	83.4%	**83.7**%	83.5%	83.6%	83.4%	**83.7**%
		+1.1%		+0.7%		+0.1%	+0.5%	+0.5%		-0.2%	+0.2%	+0.1%
Bottom 100	55.3%	54.6%	53.7%	**52.8**%	53.4%	53.2%	53.0%	**52.8**%	53.0%	53.0%	52.9%	**52.6**%
Bottom 20	51.7%	50.3%	49.4%	**48.5**%	49.4%	49.5%	**48.8**%	48.9%	49.6%	49.3%	**48.5**%	48.9%
		+2.1%		+1.8%		+0.1%	+1.0%	+1.1%		+0.3%	+1.2%	+1.1%
ρ	0.50	0.52	0.54	**0.57**	0.56	**0.57**	0.56	**0.57**	0.57	**0.58**	0.57	**0.58**

4. CONCLUSION

We introduced two relative spatial features to gain a better understanding of the image memorability. Weighted Object Area (WOA) jointly utilizes the object size and location. This contrasts to the traditional spatial pyramid based features that treat the two aspects independently. Thus WOA effectively replaces such features, e.g. `Object` in [8, 9]. Relative Area Rank (RAR) captures the relative changes of an object size compared to the other objects of same class in the training set, leading to effective encoding of the object-based unusualness. The proposed features can be easily calculated and do not require additional human labor. Experimental results show that, by replacing the classic simple area-based features, the proposed features can boost available image memorability prediction in combination with features from other domains, such as the image attributes. Moreover, the proposed features can improve the prediction without the attributes that normally require costly high level human annotations.

5. ACKNOWLEDGMENTS

This work was supported in part by the National Science Foundation under Grant No. IIS 0916812.

6. REFERENCES

[1] A. C. Berg, T. L. Berg, H. Daumé III, J. Dodge, A. Goyal, X. Han, A. Mensch, M. Mitchell, A. Sood, K. Stratos, and K. Yamaguchi. Understanding and Predicting Importance in Images. In *CVPR, IEEE Intl. Conf. on*, 2012.

[2] T. F. Brady, T. Konkle, G. A. Alvarez, and A. Oliva. Visual long-term memory has a massive storage capacity for object details. *PNAS*, 105(38):14325–14329, September 2008.

[3] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Trans. on Intelligent Systems and Technology*, 2:27:1–27:27, 2011.

[4] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR, IEEE Intl. Conf. on*, volume 1, pages 886 –893 vol. 1, june 2005.

[5] S. Dhar, V. Ordonez, and T. L. Berg. High Level Describable Attributes for Predicting Aesthetics and Interestingness. In *CVPR, IEEE Intl. Conf. on*, 2010.

[6] P.F. Felzenszwalb, R.B. Girshick, D. McAllester, and D. Ramanan. Object Detection with Discriminatively Trained Part-Based Models. *IEEE Trans. Pattern Anal. Mach. Intell.*, 32(9):1627 –1645, sept. 2010.

[7] S. J. Hwang and K. Grauman. Reading between the Lines: Object Localization Using Implicit Cues from Image Tags. *IEEE Trans. Pattern Anal. Mach. Intell.*, 34(6):1145–1158, June 2012.

[8] P. Isola, D. Parikh, A. Torralba, and A. Oliva. Understanding the Intrinsic Memorability of Images. In *NIPS 24*, 2011.

[9] P. Isola, J. Xiao, A. Torralba, and A. Oliva. What makes an image memorable? In *CVPR, IEEE Intl. Conf. on*, 2011.

[10] A. Khosla, J. Xiao, A. Torralba, and A. Oliva. Memorability of Image Regions. In *NIPS 25*, 2012.

[11] T. Konkle, T. F. Brady, G. A. Alvarez, , and A. Oliva. Scene Memory Is More Detailed Than You Think: The Role of Categories in Visual Long-Term Memory. *Psychological Science*, 21(11):1551–1556, 2010.

[12] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR, IEEE Intl. Conf. on*, volume 2, pages 2169 – 2178, 2006.

[13] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *Int. J. Comput. Vision*, 42(3):145–175, May 2001.

[14] E. Shechtman and M. Irani. Matching local self-similarities across images and videos. In *CVPR, IEEE Intl. Conf. on*, pages 1 –8, june 2007.

[15] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba. SUN Database: Large-scale Scene Recognition from Abbey to Zoo. In *CVPR, IEEE Intl. Conf. on*, 2010.