Visual Attention-driven Spatial Pooling for Image Memorability

Bora Celikkale Aykut Erdem Erkut Erdem
Hacettepe University, Ankara, TURKEY
{ibcelikkale,aykut,erkut}@cs.hacettepe.edu.tr

Abstract

In daily life, humans demonstrate astounding ability to remember images they see on magazines, commercials, TV, the web and so on, but automatic prediction of intrinsic memorability of images using computer vision and machine learning techniques was not investigated until a few years ago. However, despite these recent advances, none of the available approaches makes use of any attentional mechanism, a fundamental aspect of human vision, which selects relevant image regions for higher-level processing. Our goal in this paper is to explore the role of visual attention in understanding memorability of images. In particular, we present an attention-driven spatial pooling strategy for image memorability and show that the regions estimated by bottom-up and object-level saliency maps are more effective in predicting memorability than considering a fixed spatial pyramid structure as in the previous studies.

1. Introduction

We humans have an astonishing ability to rapidly perceive and understand complex visual scenes. When exploring parts of a city that we have never visited before, glancing at the pages of a magazine or a newspaper, watching a film on television or in a movie theatre, or the like, we are constantly bombarded with a vast amount of visual information, yet we are able to process this information and identify certain aspects of the scenes almost effortlessly [25, 27]. Humans also have an exceptional visual memory [29, 3] that we can remember particular characteristics of the scenes with ease even if we look at them only a few seconds [30]. Here, what is being remembered is considered nothing like an identical representation of the scene itself but the gist of it [33, 34]. Although there is no general agreement in the literature about the contents of this "gist", the most common definitions include statistical properties of the scene such as the distributions of basic features like color and orientation, the structural information about the scene layout like the spatial envelope of Torralba and Oliva [24], and the semantic knowledge about the existing objects and their spatial relationships.

It is intuitive that not all images are equally memorable. We can recall some images surprisingly well whereas some are lost in our minds. In [15], Isola *et al.* carried out the first computational study about understanding the memorability of images using computer vision and machine learning techniques. They quantified the memorability of 2222 photographs (See Figure 1) by performing experiments on Amazon's Mechanical Turk service to collect the rate at which the workers detect a repeat presentation of the images. Then, they investigated the contributions of several factors to memorability such as simple object statistics, scene semantics and global image features, both of which can be related to the aforementioned definitions of the 'gist' of a scene. The authors also showed that the memorability of an image can be estimated reasonably well by a machine. In a follow-up work [14], it was demonstrated that extending the previous framework to incorporate a set of human-understandable visual attributes of scenes such as attractiveness, peacefulness, etc. further improves the predictions. In a more recent study, Khosla *et al.* [20] proposed an algorithm to estimate memorability of local image regions and obtain memorability maps of images. They showed that these local features, when combined with global features, also increase the performance of memorability estimates.

As humans, we use attentional mechanisms to filter the flow of sensory information and select only a small portion of the visual stimuli in complex visual scenes for further processing to perform higher level cognitive tasks in an efficient way. Despite the recent advances in understanding image memorability from a computational viewpoint, the available models do not make use of any such attentional mechanism. In this study, we wanted to explore the role of visual attention in understanding intrinsic memorability of images. Specifically, we proposed a visual attention-driven spatial pooling strategy and analyzed its contribution to predicting image memorability in detail. Our approach made use of two complementary feature pooling schemes which are both related to visual attention. First, we investigated selecting features only from the most salient regions of the images determined according to a recently proposed bottom-

978-0-7695-4990-3/13 $26.00 © 2013 IEEE
DOI 10.1109/CVPRW.2013.142

Figure 1. Sample images from the MIT Image memorability dataset [15]. The images are sorted from more memorable (top left) to less memorable (bottom right).

up visual saliency model [8]. Our second scheme, on the other hand, considers a top-down definition of visual attention and employs an object-centric spatial pooling scheme. Pooling strategies similar to ours have been recently suggested for image and scene classification [26, 9]. Our experimental results demonstrated that memorability predictions can be improved by integrating attentional mechanisms. These results are also in line with a body of research in cognitive sciences which argues that attention plays an important role in understanding natural scenes and enhancing visual memory [34, 12, 11, 4, 13]. Here we should note that the authors of [20] utilized a visual saliency model in their model but they used saliency values as complementary features not as a part of feature pooling.

The system diagram of the proposed pooling approach is given in Figure 2. First, dense visual features such as SIFT and HOG are extracted from the input image. These features are then encoded into higher dimensions through vector quantization using a bag of features approach. In the meantime, bottom-up and object-level saliency maps are estimated from the image and then thresholded to obtain both the salient regions and those which possibly containing foreground objects. Next, to form histogram-based visual descriptors the encoded vectors are pooled together over the extracted attention-driven spatial layouts. Finally, these descriptors are concatenated together to generate the final image-level representation for memorability prediction.

2. Related Work

In this section, we briefly review previous work on image memorability [15, 14, 20] and give some details about the bottom-up visual saliency model [8] and the generic objectness measure [1] that we made use of in our work.

2.1. Image Memorability

In a recent study [15], Isola *et al.* devised a "Visual Memory Game" experiment and utilized Amazon's Mechanical Turk service to quantify the memorability of 2222 natural images of scenes and objects from the SUN dataset [35]. In the game, a total of 665 participants (the workers) were shown a sequence of images, each of which was displayed for 1 second with a 1.4 second gap in between image presentations, and asked to provide feedback any time whenever he/she thinks an identical image is displayed. In the end, the memorability score of an image was measured as the number of subjects who correctly reported a repeated presentation of the image. The authors showed that the Spearman's rank correlation between two halves of the subjects (averaged over 25 random split-half trials) was found to be 0.75, that is the memorability of an image is consistent across subjects and across a wide range of contexts. This suggests that image memorability is in fact an intrinsic property of images which is shared across different people.

In their pioneering work, Isola *et al.* also showed that intrinsic memorability of images can be predicted by using computer vision and machine learning techniques. Their framework is based on support vector regression (SVR) trained on simple image features, object and scene semantics, and, popular image features such as SIFT and HOG. Although measuring memorability of images is considered a difficult problem, the proposed model predicted image memorability significantly better than chance. In a follow-up work [14], the authors investigated the role of visual attributes to quantify memorability of images and the study revealed that predicting and exploiting attributes greatly increases the quality of the predictions. To understand which attribute is a better indicator of memorability, they investigated a greedy feature selection approach to select the relevant set of attributes. More recently, Khosla *et al.* [20] presented a probabilistic model to quantify memorability of image regions, which can predict image memorability as well as which regions are more likely to be remembered. They used ranking SVM (SVM-Rank) framework and employed saliency maps and responses of a large number of pre-trained generic object detectors from Object Bank [22] as additional features.

2.2. Visual Saliency

In recent years, there has been an increasing interest in computational models of visual saliency estimation and

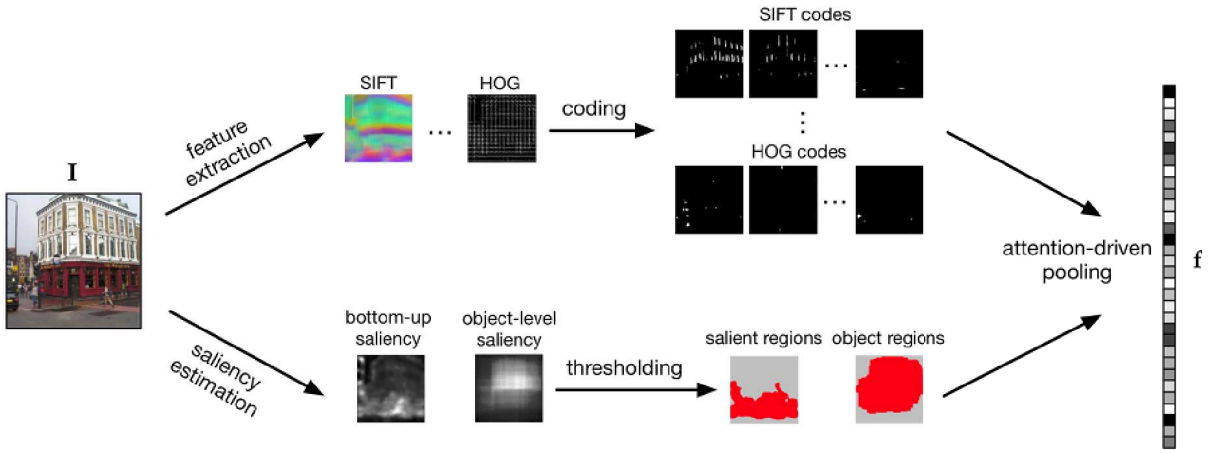


Figure 2. The proposed visual attention-driven spatial pooling pipeline for image memorability (See text for a detailed description).

their use for several computer vision tasks. Starting from the seminal work by Itti, Koch, and Niebur [16], most of the existing models consider a bottom-up strategy in which center-surround differences of various features at multiple scales are computed for each feature channel and then the final saliency map is formed by linearly combining feature maps after a normalization step. For a recent survey, please refer to [2]. In this study, we employed a recently proposed saliency model [8], which gives state-of-the-art results and differs from other models in that visual features are non-linearly integrated using region covariances without any need for intermediate steps. In our experiments, we made use of the implementation of the authors[1] which examines only the second-order statistics of simple visual features such as color, edge and spatial information.

2.3. Objectness Measure

In [1], Alexe *et al.* introduced a generic (category-independent) "objectness" measure[2] to quantify how likely an image window contains an object. In more detail, the authors first analyzed several image cues, namely multi-scale saliency, color contrast, edge density (near window borders) and superpixel straddling, each of which were shown to be an indicator of objectness, but to a certain degree. Then they proposed a Bayesian learning framework to combine these four cues to distinguish object windows from background. It was demonstrated that the approach is very general and can detect objects of novel classes not seen during training. As compared to the visual saliency model reviewed in the previous section which solely depends on bottom-up visual cues, the generic objectness measure can be used to estimate object-level saliency of images and provide top-down high-level information as will be described in Section 3.

[1]The source code is available at `http://web.cs.hacettepe.edu.tr/~erkut/projects/CovSal/`

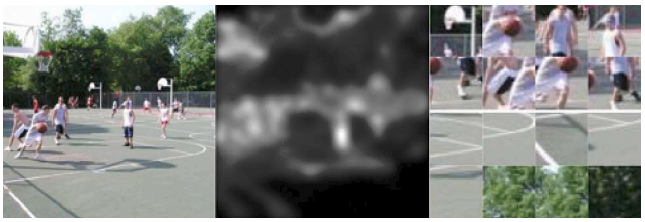
[2]The code is publicly available at `http://groups.inf.ed.ac.uk/calvin/objectness/`

3. Visual Attention-driven Spatial Pooling

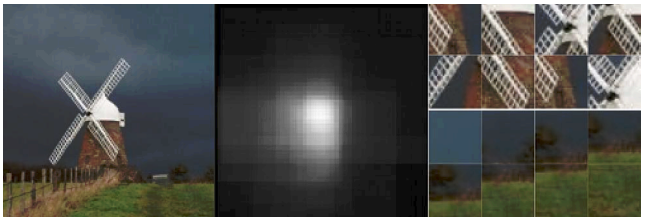
The memorability work by Isola *et al.* [15] employs spatial pyramid matching (SPM) based pooling [21]. Recently, several papers have described ways to learn optimal spatial layouts for feature pooling, *e.g.* [10, 17, 18, 26], instead of considering a fixed pyramidal structure as in SPM. In this study, we pursue a similar direction and propose an alternative visual attention-driven spatial pooling scheme for image memorability, which will be shown to be superior to SPM approach. Our approach is in part motivated by the pooling scheme proposed for scene classification [9], which is based on Itti-Koch-Niebur saliency maps [16]. As in [9], we obtain an image-specific spatial layout for feature pooling but by using a more recent bottom-up saliency model [8]. Moreover, we derive another layout structure from a complementary object-level saliency map which captures information about foreground objects in the images. It is worth mentioning that a similar object-driven pooling idea was recently presented for image classification [26].

Consider Figure 3 where we present examples for bottom-up and object-level saliency estimation. From the saliency maps given in the second column, we randomly sample a number of image patches (rightmost four columns). Those sampled within the top 10% salient locations are given in the top two rows whereas the bottom two rows show sample patches from the bottom 20% salient locations. As can be seen, the saliency values are strongly correlated with the interestingness of the regions [6, 7]. While the most salient patches captures interesting objects, the least salient ones mostly correspond to background or those which have little importance when we consider the image content.

As illustrated in Figure 2, the proposed spatial pooling pipeline has four main stages, as we formally define below:



(a) Bottom-up saliency



(b) Object-level saliency

Figure 3. Interesting and uninteresting patches extracted from two natural images based on visual attention. From the images, 8 image patches are sampled randomly from the top 10% salient locations (top 2 rows) and 8 others from the bottom 20% salient locations (bottom 2 rows) according to (a) a bottom-up visual saliency map and (b) an object-level saliency map, respectively.

(1) Feature Extraction. For an image $\mathbf{I}$, we obtain a global description of $\mathbf{I}$ by extracting D-dimensional local features such as SIFT [23], HOG [5], SSIM [28] at N different locations, denoted with $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N]^T \in \mathbb{R}^{N \times D}$.

(2) Coding. Assuming that we have a learned codebook of K visual words, denoted with $\mathbf{B} = [\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_K] \in \mathbb{R}^{D \times K}$, each local feature $\mathbf{x}_i \in \mathbf{X}$ is encoded into a code vector $\mathbf{c}_i = \left[c_1^i, c_2^i, \ldots, c_K^i,\right]^T$ by applying vector quantization. Alternative coding schemes include sparse coding [37] and locality-constrained linear coding (LLC) [32]. After the coding step, $\mathbf{I}$ is represented by a set of codes $\mathbf{C} = [\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_N] \in \mathbb{R}^{N \times K}$.

(3) Bottom-up and object-level saliency maps. To obtain the attention-driven spatial layouts for the proposed feature pooling scheme, we make use of bottom-up and object-level saliency maps. The bottom-up visual saliency map of image $\mathbf{I}$ is computed by a recently proposed model [8], which was shown to provide state-of-the-art performance in predicting eye fixations. For the object-level saliency map, we randomly sample a large number of windows from $\mathbf{I}$ and measure the objectness of these image windows by using the generic objectness measure proposed in [1]. To obtain the generic objectness map of $\mathbf{I}$, we then compute an objectness score for each pixel by averaging over all the scores of the windows which contain that pixel.

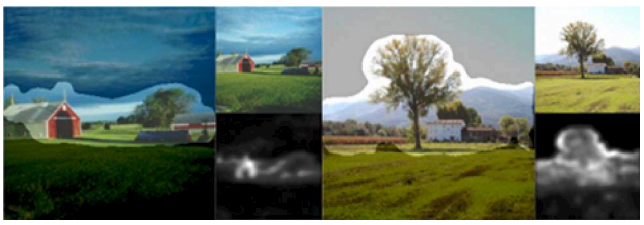
(4) Pooling. In the pooling step, instead of considering a fixed – image-independent – set of spatial regions, as employed in [15], here we propose to use image-specific spatial regions for feature pooling. Specifically, we follow an approach similar to the one in [9], in which regions of interest are located by respectively segmenting the bottom-up and object-level saliency maps into salient/non-salient and object/non-object regions by thresholding. In our experiments, we varied the threshold value to find the optimum thresholds to determine salient and object regions in the images for spatial pooling of features. We found out that the mean works well for the bottom-up saliency maps whereas the best performance for the object-level saliency maps is achieved when the threshold is set to 0.25 times the maximum objectness value. Figure 4 shows some examples of these attention-driven regions. For each region of interest $\mathcal{R}$, we then perform average-pooling, i.e. compute a histogram (or take the average of) the codes over the region $\mathcal{R}$:

$$f(\mathcal{R}) = \frac{1}{|\mathcal{R}|} \sum\nolimits_{i \in \mathcal{R}} \mathbf{c}_i \qquad (1)$$

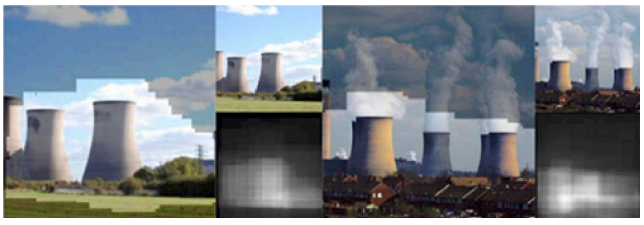
where $|\mathcal{R}|$ denote the number of dense features in $\mathcal{R}$. Moreover, the final feature vector $f(\mathcal{R})$ is renormalized to have L_1-norm of 1.

4. Experimental Results

In this section, we demonstrate the effectiveness of the proposed feature pooling strategy with a series of experiments. We first give brief details about the experimental setup and the memorability image dataset used in the experiments. Then, we describe the global visual features that were used in predicting the memorabilities. Finally, we dis-



(a) Bottom-up saliency-driven feature pooling



(b) Object-level saliency-driven feature pooling

Figure 4. Visual attention-driven feature pooling scheme. For a given image, (a) a bottom-up saliency map and (b) an object-level saliency map are estimated and then the feature vectors are pooled together in the salient regions of the images (depicted as bright areas in the images).

cuss the results of our experiments.

4.1. Experimental Setup

Dataset. For memorability predictions we used the MIT Image memorability dataset[3] introduced by Isola *et al.* [15]. This dataset contains 2222 natural images from the SUN dataset [35], which were cropped and resized to 256×256 pixels. Each image has an associated memorability score which was obtained via the Visual Memory Game discussed in Section 2.1 and was defined as the percentage of correct detections by participants of the game. Moreover, object annotations and scene category label are also available for all these images.

Evaluation. For the quantitative analysis we used Spearman's rank correlation measure (ρ) and a precision-recall measure. The performance was evaluated over 25 different splits of the dataset containing 1111 training and 1111 testing images (the same splits used in [15]). These train and test splits were scored by different halves of the participants, showing a human consistency of $\rho = 0.75$. Thus, the effectiveness of a computational image memorability model can be assessed by measuring how close the model's Spearman rank correlation to this score. In addition, for a precision-recall analysis, test images can be ranked according to their predicted memorability and then the cumulative average of measured empirical memorability scores can be examined for different sets of images. For instance, a good image memorability model should have cumulative averages close to 100% for the top most memorable images predicted by a model.

4.2. Global Image Features

In the experiments, we considered a combination of three groups of global image features which were also used by Isola *et al.* in [15]. These are color histograms, GIST [24] and a set of dense visual features including SIFT [23], HOG [5] and SSIM [28]. We briefly discuss these features in the subsequent sections. Note that we perform attention-driven feature pooling only for the dense visual features.

Color histograms. Color is an important attribute of visual perception. Here, we considered simple color information of images as a complementary cue to the other visual features. We used the 3-dimensional pixel histograms with 21 bins per channel in RGB color space to obtain a 63-dimensional color descriptor for each image.

GIST. We used the GIST scene descriptor [24] which produces a holistic low-dimensional view of the input image by decomposing it by a multiscale oriented filterbank and then taking filter responses over a grid or image regions. We considered 8 orientations, 4 scales and a 4×4 grid, which resulted in a $4 \times 8 \times 16 = 512$ dimensional vector.

Dense visual features. We considered three different dense visual features, SIFT [23], HOG [5] and SSIM [28], which were pooled using the proposed attention-driven scheme. The SIFT descriptor gives the local image structural information whereas the HOG descriptor provides rich local orientation information that can be related to the receptive fields found in early human vision areas. Lastly, the SSIM descriptor captures the local layout of geometric patterns. We densely sampled these features at the pixel level and then pooled each feature together over two different saliency maps, one from bottom-up visual saliency and the other from generic objectness. The size of the used codebook is 200 and thus we obtain a 400-dimensional feature vector for each dense feature.

All global features. We combined the global image features listed above and devised a single high-dimensional feature vector by concatenating all the feature vectors respectively. This process produced $63 + 512 + 3 \times 400 = 1765$ dimensional image-level representation for memorability prediction. It is important to note that the final dimension is nearly half of that of Isola *et al.* [15] which depends on a SPM based pooling on a $2 \times 2 + 1 \times 1$ spatial tiling.

4.3. Results

We first examined the use of eight different attention-driven spatial layouts in predicting image memorability. In particular, we tested several combinations of salient/non-salient and object/non-object regions as suggested by a bottom-up saliency model [8] and a generic objectness measure [1] to determine the image-dependent layout structure for pooling. In each case, the final image-level representation $\mathbf{f}$ was obtained by concatenating the related descriptors. Consequently, we separately trained eight different SVRs to map from the features pooled over these maps to memorability scores.

Table 1 summarizes our results. The first four pooling layouts are solely based on salient, non-salient, object and non-object regions, respectively, and they all have the same performance ($\rho = 0.46$), which is also equal to the score of Isola *et al.* [15]. This is interesting because it illustrates that excluding the corresponding regions from the prediction estimation does not hurt the performance. The worst performance is achieved ($\rho = 0.41$) when the features are pooled over the combination of non-salient and non-object regions. Here, what's more interesting about our results is that pooling the features over salient and object regions together achieves a rank correlation of $\rho = 0.47$, provid-

[3]The dataset is publicly available at `http://web.mit.edu/phillipi/Public/WhatMakesAnImageMemorable/`

Predicted as highly memorable (89%) Predicted as typically memorable (67%) Predicted as least memorable (48%)

Figure 5. Memorability predictions by the proposed attention-driven feature pooling strategy. Out of all test images, the 8 images in (a) are found to be the most memorable, the ones in (b) are predicted as typically memorable and the other 8 images in (c) are guessed as the least memorable. The numbers denote the average prediction scores of the given image sets. The images predicted as highly memorable contains highly distinctive visually salient elements as compared to other groups of images.

Table 1. Comparison of predictions via different combinations of attention-based feature pooling schemes (pooling over S: salient, $\neg S$: non-salient, O: object, $\neg O$: non-object regions, with '+' denoting concatenation) versus empirically measured memory scores. In example, the first row indicates average empirical memorability over the images with the *top 20* highest predicted memorabilities, and ρ is the Spearman rank correlation between model predictions and empirical results.

	S	$\neg S$	O	$\neg O$	$S+\neg S$	$O+\neg O$	$\neg S+\neg O$	$S+O$
Top 20	83%	83%	83%	83%	84%	84%	83%	84%
Top 100	80%	80%	80%	80%	81%	80%	79%	81%
Bottom 100	56%	56%	56%	56%	56%	56%	57%	56%
Bottom 20	53%	53%	53%	53%	55%	55%	57%	55%
ρ	0.46	0.46	0.46	0.46	0.46	0.45	0.41	0.47

ing the best memorability prediction performance across the set. That is, the top down information provided by object-level saliency combined with the bottom up information predicted by visual saliency gives better results than those of using top down or bottom up information alone. This result strongly supports our claim that the image regions which retain in human memory is highly correlated with the areas that attract our attention.

Figure 5 shows sample images from the memorability predictions based on salient and object regions. In addition to these qualitative results, we also compare our results with Isola *et al.* [15] and Khosla *et al.* [20]. In Figure 6, we present the precision-recall performances of our model together with Isola *et al.*'s global features model, predictions based on annotated objects and scenes, and human predictions [15]. For the topmost 300 images our model gives slightly better predictions than Isola *et al.*'s global features model. Table 2 summarizes the performances of our model and other computational models in terms of Spearman's rank correlation measure (ρ) and the precision-recall measure. As it can be seen, we achieved a better performance as compared to Isola *et al.* [15] even if we used the same global features. Here, it is important to note that the size of our image level descriptor is nearly half of the one used by Isola *et al.* [15]. This demonstrates another benefit of visual attention-based feature pooling for image memorability. It should be noted that Khosla *et al.* [20]'s global and full models provided predictions better than ours but they employed semantically more complex features.

Figure 7 shows sample images on which the memorability predictions based on our approach are incorrect as compared to the empirical results. To argue about why our model fails to capture the intrinsic memorabilities, in Figure 8, we provide the bottom-up and object-level saliency maps of two of the images from Figure 7 together with their memorability maps obtained from object annotations. In the memorability maps, the red regions illustrate the objects that contribute positively to the predicted memorability and the blue regions show the objects that contribute negatively to the predicted memorability. For the "iceberg"

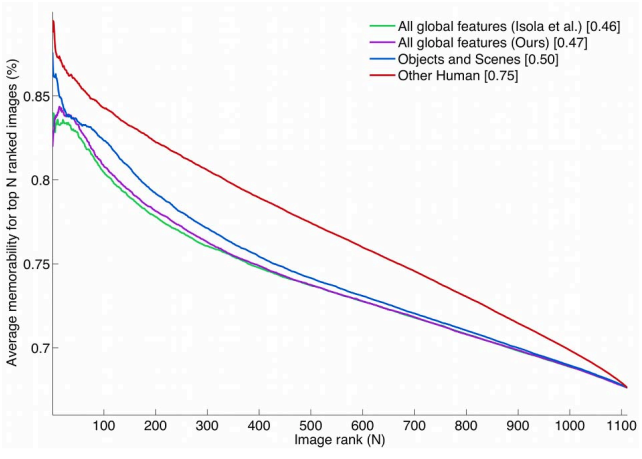


Figure 6. Comparison of regression results averaged across the 25 splits. Test images are ranked according to their predicted memorability and plotted against the cumulative average of measured memorability scores.

image whose memorability rank was overshot by the proposed prediction scheme (Figure 8(a)), our pooling did not correctly identify the memorable image regions. For the "street view" image whose memorability rank was undershot by the proposed scheme (Figure 8(b)) our pooling had given more prominence to the object regions that affects the memorability predictions negatively.

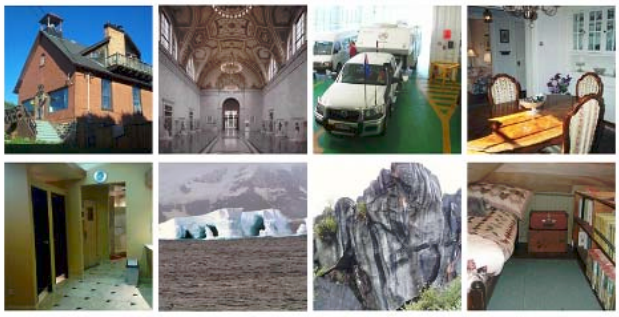
5. Conclusion

We have presented a novel feature pooling strategy for image memorability based on visual attention. The new strategy is derived from the observation that main memorable areas of an image are the ones that attract the most attention [34, 12, 11, 4, 13]. The proposed scheme is practically advantageous and effective than the traditional SPM based pooling as it forms a lower dimensional image-level representation while enhancing the memory prediction performance further. Instead of considering a fixed pyramidal structure as in [15, 20], our regression model learns memorability scores of images by taking the concatenation of pooled features over the saliency maps as input. In the suggested scheme, we employed two saliency maps, one by a bottom-up saliency model [8] and the other by a generic objectness model [1]. These maps respectively model bottom-up and top-down attentional influences that affect memorability estimations. Experiments on the MIT image memorability dataset demonstrated that the proposed pooling scheme improves the prediction quality of the Isola *et al.*'s model [15] by using the same set of global image features but with lower dimensional descriptors. We expect the proposed scheme would be quite effective also for other global features such as the semantic ObjectBank features used in [20].

In the context of this work, we investigated how attention driven spatial pooling strategies, which are defined based on bottom-up and object-level saliency, can help to improve predicting image memorabilities. For future work, it would be interesting to investigate the inverse problem, *i.e.* how visual memorability affects visual saliency. A recent trend in visual saliency estimation is to pose saliency estimation as a supervised learning problem [31, 19, 38, 36]. Most of these models with the exception of [31, 36] try to predict where human look in the images under free-viewing conditions. Motivated with these works, one can try to devise a task-dependent model with the task being defined as to memorize image content.

Table 2. Test images are ranked by their predicted memorabilities suggested by different models (denoted by column headings) and as in Table 1, the average predicted memorabilities are reported for different sets of images, together with the Spearman rank correlation ρ between model predictions and empirical results.

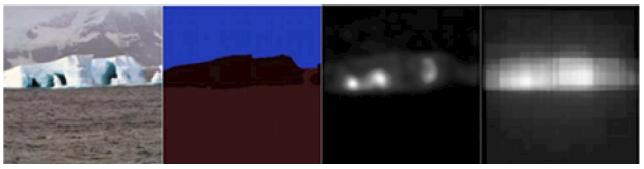
	Isola *et al.* [15]	Khosla *et al.* [20]			Our global
	global	local	global	full	model
Top 20	83%	84%	83%	85%	84%
Top 100	80%	80%	80%	81%	81%
Bottom 100	57%	56%	57%	55%	56%
Bottom 20	55%	53%	54%	52%	55%
ρ	0.46	0.48	0.45	0.50	0.47



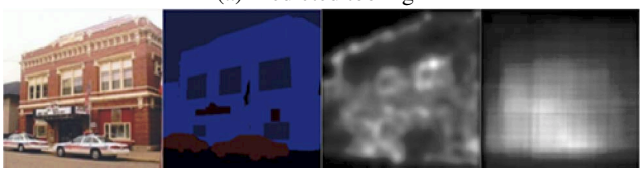
Predicted too high (+832/1111)

Predicted too low (-866/1111)

Figure 7. Sample images on which our proposed scheme failed to capture the memorability. The memorability ranks are predicted too high for the images in (a) and too low for the ones in (b), as compared to their empirical memorability ranks. The numbers in the parentheses show the mean rank error between the predicted and the empirical ranks across each group.



(a) Predicted too high



(b) Predicted too low

Figure 8. Memorability maps versus bottom-up saliency and object-level saliency maps of two of the images from Figure 7 (See text for details)

Acknowledgments

This research was supported in part by The Scientic and Technological Research Council of Turkey (TUBITAK), Career Development Award 112E146.

References

[1] B. Alexe, T. Deselares, and V. Ferrari. What is an object? In *CVPR*, 2010.

[2] A. Borji. State-of-the-art in visual attention modeling. *IEEE-TPAMI*, 35(1):185–207, 2013.

[3] 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, 2008.

[4] M. A. Cohen, G. A. Alvarez, and K. Nakayama. Natural-scene perception requires attention. *Psychological Science*, 22:1165–1172, 2011.

[5] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.

[6] W. Einhäuser, M. Spain, and P. Perona. Objects predict fixations better than early saliency. *Journal of Vision*, 8(14):1–26, 2008.

[7] L. Elazary and L. Itti. Interesting objects are visually salient. *Journal of Vision*, 8(3):1–15, 2008.

[8] E. Erdem and A. Erdem. Visual saliency estimation by nonlinearly integrating features using region covariances. *Journal of Vision*, 13(4):1–20, 2013.

[9] M. Fornoni and B. Caputo. Indoor scene recognition using task and saliency-driven feature pooling. In *BMVC*, pages 98.1–98.12, 2012.

[10] T. Harada, Y. Ushiku, Y. Yamashita, and Y. Kuniyoshi. Discriminative spatial pyramid. In *CVPR*, 2011.

[11] A. Hollingworth and J. M. Henderson. Accurate visual memory for previously attended objects in natural scenes. *Journal of Experimental Psychology: Human Perception and Performance*, 28(1):113–136, 2002.

[12] A. Hollingworth, C. C. Williams, and J. M. Henderson. To see and remember: visually specific information is retained in memory from previously attended objects in natural scenes. *Psychonomic Bulletin & Review*, 8(4):761–768, 2001.

[13] K. Inoue and Y. Takeda. The role of attention in the contextual enhancement of visual memory for natural scenes. *Visual Cognition*, 20(1):94–107, 2012.

[14] P. Isola, D. Parikh, A. Torralba, and A. Oliva. Understanding the intrinsic memorability of images. In *NIPS*, pages 2429–2437, 2011.

[15] P. Isola, J. Xiao, A. Torralba, and A. Oliva. What makes an image memorable? In *CVPR*, 2011.

[16] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE-TPAMI*, 20(11):1254–1259, 1998.

[17] Y. Jia, C. Huang, and T. Darrell. Beyond spatial pyramids: Receptive field learning for pooled image features. In *CVPR*, 2012.

[18] Y. Jiang, J. Yuan, and G. Yu. Randomized spatial partition for scene recognition. In *ECCV*, 2012.

[19] T. Judd, K. Ehinger, F. Durand, and A. Torralba. Learning to predict where humans look. In *ICCV*, pages 2106–2113, 2009.

[20] A. Khosla, J. Xiao, A. Torralba, and A. Oliva. Memorability of image regions. In *NIPS*, pages 305–313, 2012.

[21] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006.

[22] L.-J. Li, H. Su, E. Xing, and L. Fei-Fei. Object bank: A high-level image representation for scene classification & semantic feature sparsification. In *NIPS*, pages 1378–1386, 2010.

[23] D. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(4):91–110, 2004.

[24] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *IJCV*, 42(3):145–175, 2001.

[25] M. C. Potter. Short-term conceptual memory for pictures. *Journal of Experimental Psychology: Human Learning and Memory*, 2(5):509–522, 1976.

[26] O. Russakovsky, Y. Lin, K. Yu, and L. Fei-Fei. Object-centric spatial pooling for image classification. In *ECCV*, 2012.

[27] P. G. Schyns and A. Oliva. From blobs to boundary edges: Evidence for time- and spatial-scale-dependent scene recognition. *Psychological Science*, 5(2):195–200, 1994.

[28] E. Shechtman and M. Irani. Matching local self-similarities across images and videos. In *CVPR*, 2007.

[29] R. N. Shepard. Recognition memory for words, sentences, and pictures. *Journal of Verbal Learning and Verbal Behavior*, 6:156–163, 1967.

[30] L. Standing. Learning 10,000 pictures. *Quarterly Journal of Experimental Psychology*, 25:207–222, 1973.

[31] A. Torralba, A. Oliva, M. Castelhano, and J. Henderson. Contextual guidance of eye movements and attention in real-world scenes: The role of global features on object search. *Psychological Review*, 113(4):766–786, 2006.

[32] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong. Locality-constrained linear coding for image classification. In *CVPR*, 2010.

[33] J. M. Wolfe. Visual memory: What do you know about what you saw? *Current Biology*, 8:R303–R304, 1998.

[34] J. M. Wolfe, T. S. Horowitz, and K. O. Michod. Is visual attention required for robust picture memory? *Vision Research*, 47:955–964, 2007.

[35] J. Xiao, J. Hayes, K. A. Ehinger, A. Oliva, and A. Torralba. SUN database: Large-scale scene recognition from abbey to zoo. In *CVPR*, 2010.

[36] J. Yang and M.-H. Yang. Top-down visual saliency via joint CRF and dictionary learning. In *CVPR*, 2012.

[37] J. Yang, K. Yu, Y. Gong, and T. Huang. Linear spatial pyramid matching using sparse coding for image classification. In *CVPR*, 2009.

[38] Q. Zhao and C. Koch. Learning visual saliency by combining feature maps in a nonlinear manner using AdaBoost. *Journal of Vision*, 12(6):1–15, 2012.